MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2770
Measure Title: Family Experiences with Coordination of Care (FECC) Measure Set
Measure Steward: Seattle Children´s Research Institute
Brief Description of Measure: In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs was to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum. Our Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN) was charged by CMS and AHRQ to develop measures assessing the quality of care coordination for children with medical complexity (CMC).

The Family Experiences with Coordination of Care (FECC) Survey was developed to gather information about the quality of care coordination being received by children with medical complexity (CMC) over the previous 12 months. The FECC Survey is completed by English- and Spanish-speaking caregivers of CMC aged 0-17 years with at least 4 medical visits in the previous year, and it includes all of the information needed to score 20 separate and independent quality measures, a sub-set of 10 of which are included in this submitted measure set. CMC are identified from administrative data using the Pediatric Medical Complexity Algorithm (PMCA)1, which uses up to 3 years' worth of International Classification of Diseases—9th Revision (ICD-9) codes to classify a child's illness with regard to chronicity and complexity. CMC are children identified by the PMCA as having complex, chronic disease.

This submission relates to a set of 10 of the FECC quality measures. The short descriptions of each included quality measure follows; full details are provided in the Detailed Measure Specifications (see S.2b):

FECC-1: Has care coordinator
FECC-3: Care coordinator helped to obtain community services
FECC-5: Care coordinator asked about concerns and health changes
FECC-7: Care coordinator assisted with specialist service referrals
FECC-8: Care coordinator was knowledgeable, supportive and advocated for child's needs
FECC-9: Appropriate written visit summary content
FECC-14: Health care provider communicated with school staff about child's condition
FECC-15: Caregiver has access to medical interpreter when needed
FECC-16: Child has shared care plan
FECC-17: Child has emergency care plan

Each of the quality measures is scored on a 0-100 scale, with higher scores indicating better care. For dichotomous measures, a score of 100 indicates the child received the recommended care; a score of 0 indicates that they did not. Please see Detailed Measure Specifications (see S.2b) for additional measure-specific scoring information.
Developer Rationale: Increasing numbers of children in the United States are living with medical complexity.(2) Although these children with medical complexity (CMC) comprise only 13% of the pediatric population, they account for a disproportionately high 26-49% of hospital days(3,4) and 70% of overall health expenditures.(5) Given the cost and complexity of caring for these children, optimizing the quality of their care is likely to yield significant health and economic benefits.

Comprehensive, well-coordinated care in a medical home improves patient and family experiences of care6-8 and patient medical outcomes.(6,7,9,10) Care coordination interventions among CMC have also been associated with decreased unmet specialty care

need11 and improved utilization of health care services, decreasing hospitalizations and cost.8,9,12-14 Improving care coordination for CMC is likely to improve many aspects of care received by these children and their families.

Little is known about the quality of care coordination received by CMC. Present assessments of care coordination are generally limited to whether care coordination was received or not, without any attempt to identify potentially beneficial components of care coordination or the manner in which they were delivered. The evidence that is available suggests that 29-41% of parents of children with special health care needs report not getting needed help with care coordination;(15,16) little is known about the quality of the help that is being received.

While limited information on quality of care coordination exists, data do demonstrate disparities in receipt of care coordination. Latino and black children have been found to be more likely to have unmet care coordination needs compared to non-Hispanic white children.(16) In addition, children from families with limited English proficiency have reported higher unmet care coordination needs and greater difficulty getting needed referrals compared to English proficient families.(15) These data suggest that there may also be disparities in quality of care coordination received by race/ethnicity and language. The FECC Survey can be collected with data on child and parent race, ethnicity and language, which will allow for tracking of disparities in care coordination quality over time.

references:

2. Bethell CD, Read D, Blumberg SJ, Newacheck PW. What is the prevalence of children with special health care needs? Toward an understanding of variations in findings and methods across three national surveys. Matern Child Health J. 2008;12(1):1-14.
3. Berry JG, Hall M, Hall DE, et al. Inpatient growth and resource use in 28 children´s hospitals: a longitudinal, multi-institutional study. JAMA Pediatr. 2013;167(2):170-177.
4. Simon TD, Berry J, Feudtner C, et al. Children with complex chronic conditions in inpatient hospital settings in the United States. Pediatrics. 2010;126(4):647-655.
5. Ireys HT, Anderson GF, Shaffer TJ, Neff JM. Expenditures for care of children with chronic illnesses enrolled in the Washington State Medicaid program, fiscal year 1993. Pediatrics. 1997;100(2 Pt 1):197-204.
6. Farmer JE, Clark MJ, Sherman A, Marien WE, Selva TJ. Comprehensive primary care for children with special health care needs in rural areas. Pediatrics. 2005;116(3):649-656.
7. Farmer JE, Clark MJ, Drewel EH, Swenson TM, Ge B. Consultative care coordination through the medical home for CSHCN: a randomized controlled trial. Matern Child Health J. 2011;15(7):1110-1118.
8. Palfrey JS, Sofis LA, Davidson EJ, Liu J, Freeman L, Ganz ML. The Pediatric Alliance for Coordinated Care: evaluation of a medical home model. Pediatrics. 2004;113(5 Suppl):1507-1516.
9. Counsell SR, Callahan CM, Clark DO, et al. Geriatric care management for low-income seniors: a randomized controlled trial. JAMA. 2007;298(22):2623-2633.
10. Rocco N, Scher K, Basberg B, Yalamanchi S, Baker-Genaw K. Patient-centered plan-of-care tool for improving clinical outcomes. Qual Manag Health Care. 2011;20(2):89-97.
11. Boudreau AA, Perrin JM, Goodman E, Kurowski D, Cooley WC, Kuhlthau K. Care coordination and unmet specialty care among children with special health care needs. Pediatrics. 2014;133(6):1046-1053.
12. Casey PH, Lyle RE, Bird TM, et al. Effect of hospital-based comprehensive care clinic on health costs for Medicaid-insured medically complex children. Arch Pediatr Adolesc Med. 2011;165(5):392-398.
13. Dorr DA, Wilcox AB, Brunker CP, Burdon RE, Donnelly SM. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. J Am Geriatr Soc. 2008;56(12):2195-2202.
14. Gordon JB, Colby HH, Bartelt T, Jablonski D, Krauthoefer ML, Havens P. A tertiary care-primary care partnership model for medically complex and fragile children and youth with special health care needs. Arch Pediatr Adolesc Med. 2007;161(10):937-944.
15. Zickafoose JS, Davis MM. Medical home disparities are not created equal: differences in the medical home for children from different vulnerable groups. J Health Care Poor Underserved. 2013;24(3):1331-1343.

Numerator Statement: The numerators for each of the 10 FECC quality measures included within the FECC measures set are specified in the Detailed Measure Specifications (see S.2b). A brief description of each numerator is laid out in Table 1 in section De.3, and a more detailed description follows:

FECC-1: Caregivers of CMC should report that their child has a designated care coordinator.

FECC-3: Caregivers of CMC who report having a designated care coordinator and who require community services should also report that their care coordinator helped their child to obtain needed community services in the last year.

FECC-5:Caregivers of CMC who report having a care coordinator and who report that their care coordinator has contacted them in

the last 3 months should also report that their care coordinator asked them about the following:

- Caregiver concerns
- Health changes of the child

FECC-7: Caregivers of CMC who report having a care coordinator for their child should also report that the care coordinator assists them with specialty service referrals by ensuring that the appointment with the specialty service provider occurs

FECC-8: Caregivers of CMC who report having a care coordinator should also report that their care coordinator:

- Was knowledgeable about their child's health
- Supported the caregiver
- Advocated for the needs of the child

FECC-9: Caregivers of CMC who report receiving a written visit summary during the last 12 months from their child's main provider's office should report that it contained the following elements:

- Current problem list
- Current medication list
- Drug allergies
- Specialists involved in the child's care
- Planned follow-up
- What to do for problems related to outpatient visit

FECC-14: Caregivers of CMC who report their child's condition causes difficulty learning, understanding, or paying attention in class should also report that one of their child's health care providers (i.e., primary care physician, specialist physician, care coordinator, nurse practitioner, nurse, social worker, etc.) communicated with school staff at least once a year about the educational impacts of the child's condition.

FECC-15: Caregivers of CMC who self-identify as having a preference for conducting medical visits in a language other than English should have access to a professional medical interpreter (live or telephonic) at all visits for which an interpreter is needed.

FECC-16: Caregivers of CMC should report that their child's primary care provider created a shared care plan for their child.

FECC-17: Caregivers of CMC should report that their child's main provider created an emergency care plan for their child.

Denominator Statement: The eligible population of caregivers for the FECC Survey overall is composed of those who meet the following criteria:

1. Parents or legal guardians of children 0-17 years of age
2. Child classified as having a complex, chronic condition using the Pediatric Medical Complexity Algorithm (PMCA) (see Simon TD, Cawthon ML et al. 2014)
3. Child had at least 4 visits to a healthcare provider over the previous year

While some of the FECC measures only apply to a subset of the overall eligible population for the survey (e.g., measures related to the quality of care coordination services provided are only scored for those caregivers who endorse having a care coordinator), eligibility for these quality measures can only be gleaned from responses to the FECC Survey itself. This is analogous to the situation with many H-CAHPS measures, where, for example, measures about blood draws and laboratory testing are scored only for those who had the relevant service performed during the time frame or hospitalization in question.

Denominator Exclusions: Denominator exclusions:

1. Child had died
2. Caregiver spoke a language other than English or Spanish

Measure Type: Process
Data Source: Administrative claims, Patient Reported Data/Survey
Level of Analysis: Health Plan, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: not applicable

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer reports the following:

- The developer states the FECC includes all of the information needed to score 20 ***separate and independent quality measures***; ***a subset of 10 have been submitted for endorsement consideration***. The Level of Analysis is health plan or population: state.
- These are process measures. For all measures, a higher score = better quality.
- The developer created a conceptual framework for care coordination/fragmentation for children with medical complexity that indicates events that may lead to fragmented care and that demonstrates how care coordination relates to short and long-term outcomes. Based on the conceptual framework, it identified six topics for evidence review and conducted a literature review for each. The measures were drafted based on the results of the review and identified evidence.
- The evidence for each measure was rated using the Oxford Centre for Evidence Based Medicine grading scale. The quality for each measure ranges from randomized controlled trials to consensus or mechanism based reasoning. The developer provided the following evidence for each measure. Additional details are provided in the Evidence attachment tables and narrative.

    - FECC 1. Caregivers of CMC should report that their child has a designated care coordinator: One randomized control study (RCT) (Farmer et al 2011), one cohort study, and five case series, case-control, or historically-controlled studies demonstrate that outcomes improve when caregivers of children with medical complexity (CMC) report that their child has a designated care coordinator.
        - The RCT involved intervention for 100 children with chronic illness on Medicaid (6-month intervention supporting 32 primary care provider [PCP] offices), wherein the care coordinator worked with the family to develop a written health plan for the child to provide access to services and coordination with doctors and home visit/ telephone support.
            - In the between-group analyses, participants in the intervention reported significantly higher satisfaction with mental health services and specialized therapies as measured by a family survey adapted from the Shared Responsibilities Tool Kit—Version 1.0, and significantly lower need for information as measured by the Family Needs Survey (FNS).
            - In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs as measured by the FNS. There was a significant improvement in satisfaction with specialty care and care coordination as measured by the Shared Responsibilities Tool Kit—Version 1.0. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status as measured by the Functional Status II (Revised)—14 item version. There was a significant decrease in personal and family strain, as measured by the Impact on Family Scale (IFS).
        - The balance of the evidence cited involves studies with 43 to 277 children and findings from them include (depending on the study): significantly fewer barriers to services; improved satisfaction with care coordination services; significant increase in Emergency Department (ED) use; significant

decrease in hospitalization and length of stay; significant increase or decrease (depending on study) in outpatient visits; decreased cost of care; significant increase in satisfaction; significant decrease in lost work days; significant decrease in school absence; significant decrease in unmet needs, and/or decrease in family strain.

- FECC 3. Caregivers of CMC who report having a designated care coordinator and who require community services should also report that their care coordinator helped their child to obtain needed community services in the last year: One randomized controlled trial and two uncontrolled intervention studies demonstrate that outcomes improve when care coordinators assist families with obtaining needed community services.
  - In the same Farmer RCT as that for FECC 1, the care coordination intervention included, among other components: a) facilitating communication among families, primary and specialty care providers, and community service agencies, and b) providing information to help the family access needed educational and community resources. The developer notes it is not possible to determine which elements of this bundled intervention resulted in improved outcomes, but that the authors stated that the improvement could be related to obtaining community services (e.g., mental health services and therapies). The authors also found decreased unmet needs, some of which may be met by community services.
    - In the between-group analyses, participants in the intervention reported significantly higher satisfaction with mental health services and specialized therapies, and significantly lower need for information. [Note to Committee: to streamline and avoid repetition the particular assessment tools or surveys for the Farmer RCT are not repeated for each FECC after FECC 1.].
    - In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status. There was a significant decrease in personal and family strain.
  - The two additional studies cited (n=227 or 43) found (depending on the study) a significant decrease in the number of hospitalizations and lengths of stay; an increase in the use of outpatient services; a decrease in tertiary care center payments.

- FECC 5. Caregivers of CMC who report having a care coordinator and who report that their care coordinator has contacted them in the last 3 months should also report that their care coordinator asked them about the following: a. caregiver concerns, b. health changes of the child: One randomized controlled trial demonstrates that having a care coordinator that asks about the child with medical complexity's progress is associated with improved outcomes.
  - The developer cites the same RCT as noted for FECC 1. In that RCT, the care coordination intervention included telephone contact to discuss the child's progress at least once each month as one component of the intervention. The authors suggest that suggest that caregiver concerns were being addressed because of the contact as represented by a significant decrease in personal and family strain. There also was a significant improvement in satisfaction with specialty care and care coordination, a significantly improved overall child health rating, and a trend toward improved child functional status.

- FECC 7. Caregivers of CMC who report having a care coordinator for their child should also report that the care coordinator assists them with specialty service referrals by ensuring that the appointment with the specialty service provider occurs within 3 months of referral initiation: One randomized controlled trial and three uncontrolled intervention studies demonstrate that outcomes improve when care coordinators assist families with making sure specialty service referrals are successfully completed.
  - Again, in the previously cited Farmer RCT, the care coordination intervention included: a) facilitating

communication among families, primary and specialty care providers, and community service agencies, and b) direct advocacy for needed care, as required. The developers report the authors did not track the completion of appointments, but state several findings suggest that families who received the intervention were receiving needed services.

- In the between-group analyses, participants in the intervention reported significantly higher satisfaction with mental health services and specialized therapies.
- In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status.

- The developers cite three additional pre- and post-intervention studies (n=43 to 227) that found (depending on the study) a significant decrease in hospitalizations and lengths of stay, increase in the use of outpatient services, decrease in tertiary care center payments, decrease in missed work days, and that it was easier to obtain services. The developer specifically notes that one study did not track completion of appointments; no information is provided about the second study.

- FECC 8. Caregivers of CMC who report having a care coordinator should also report that their care coordinator: a. is knowledgeable about their child's health, b. supports the caregiver, c. advocates for the needs of their child: One randomized controlled trial and three other studies demonstrate that outcomes improve when care coordinators are knowledgeable, supportive, and good advocates for the child's needs.
  - In the previously cited Farmer RCT, the care coordination intervention included: a) facilitating communication among families, primary and specialty care providers, and community service agencies, b) direct advocacy for needed care, as required; and c) telephone contact to discuss the child's progress at least once each month as one component of the intervention. These activities were intended to ensure that the care coordinator was informed about the child's health and could support the caregiver and advocate for the child's needs.
    - In the between-group analyses, participants in the intervention reported significantly lower needs for information.
    - In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs. There was a significant improvement in satisfaction with specialty care and care coordination, a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status. There was a significant decrease in personal and family strain.
  - The developers cite three additional studies (n=51 to 227) that found. Two of the studies involved a pediatric nurse case manager or a designated pediatric nurse practitioner. The developer posits that these single points of contact likely resulted in the individual having knowledge of the child's health and supporting and advocating for the child and family. Depending on the study, there was a significant decrease in hospitalizations and lengths of stay, increase in the use of outpatient services, decrease in tertiary care center payments, decrease in parents' missed work days, and that it was easier to obtain services. The developers state the third study provided care coordination, information about resources and services, emotional support, and empowerment for families to advocate for their children; there was a statistically significant decrease in specialty care, an increase in satisfaction with care coordination services, a significant decrease in missed work days and missed school days, and a significant decrease in unmet needs and family strain.

- FECC 9. Caregivers/patients of CMC should report receiving a written visit summary following all outpatient visits in the last 12 months (or report access to a patient portal that provides a visit summary) and it should contain the following elements: a. current problem list; b. current medication list; c. drug allergies; d. specialists involved in the child's care; e. planned follow-up; f. what to do for problems

related to the outpatient visit: One study that used a pre/post intervention comparison design and two other expert consensus sources (medical home standards from the NCQA and guidelines from the American Academy of Pediatrics (AAP) support that caregivers of CMC should report receiving a written visit summary following all outpatient visits.

- The developer reports Palfrey et al. (2004) evaluated the medical home model in Massachusetts through six pediatric practices that introduced interventions to operationalize the medical home for children with special health care needs (n=117). One outcome measured was receipt of a written care plan. After the intervention, more families reported that their PCP gave them a written health care plan (30% at before and 47% after, $p<0.01$). In addition, there were fewer hospitalizations and a decrease in parents missing > 20 days of work. There was no change in ED use or school absences. The developers note that since receipt of a written plan was itself an outcome, conclusions are limited to noting correlation of the receipt of a written care plan with other improvements in outcomes. The developers do not indicate if the written care plan contained all elements required by the FECC 9 specifications.
- The developers also cite 2011 NCQA standards and 2005 AAP guidelines on the importance of a written plan (NCQA) and that the "medical home physician should share information among the child, family, and consultants." The developer reports neither document specified the elements that should be included in the written plan or communication.

- FECC 14. Caregivers of CMC should report that one of their child's health care providers (i.e., primary care physician, specialist physician, care coordinator, NP, nurse, social worker, etc.) communicated with school staff at least once a year about the educational impacts of the child's condition: One paper that synthesizes the authors' experience and provides guidance supports that caregivers should report that their child's health care providers communicated with school staff about the educational impacts of the child's condition.
  - The developers report that Savage et al. (2004) conducted a study (n=66) involving the treatment and recovery of children with a traumatic brain injury, and synthesized their experience to provide guidance for transitioning back into school. The authors identify the importance of having a representative of the patient-centered medical home share suggestions for easing transitions between school and medical facilities and request training for school staff working with the student regarding the condition, best practices, and related educational impacts. The developers do not report whether the paper identified improved outcomes resulting from such communication or the periodicity/frequency of the communication.

- FECC 15: Caregivers of CMC or CMC who self-identify as having a preference for conducting medical visits in a language other than English should have access to a professional medical interpreter (live or telephonic) at all visits for which an interpreter is needed: One systematic review, one randomized controlled trial, two non-randomized controlled interventions, and one retrospective cohort study support that provision of professional interpreter services improves patient outcomes. While these studies do not examine outcomes among medically complex children specifically, they all included patients with a heterogeneous mix of conditions. The developer states it would expect that an intervention to improve communication, associated with improved outcomes, would be at least as beneficial in patients with greater complexity as in those without complex conditions, if not more so.
  - The developers report that Karliner (2007) conducted a systematic review to determine if the use of professional interpreters improves medical care for patients with limited English proficiency (LEP). The review included one randomized controlled trial and 27 cohort studies comparing professional interpreter use to another group (no interpreter use, bilingual provider use, or different types of interpreter use), published between 1966 and 2005, and assessing satisfaction, utilization, clinical outcomes, or comprehension. Sample sizes of participants/encounters ranged from 13 to 4,146. The developer does not report whether the evidence in the systematic review was graded.

- Use of professional interpretation, compared to ad hoc (family or friend as interpreter) or no interpretation, was consistently associated with better outcomes, generally approaching or equaling those of patients with language concordant physicians (both Spanish speakers with Spanish-speaking physicians, or English-speakers with English-speaking physicians).
- The review concluded that professional interpreter use was associated with decreased disparities in utilization and adherence to follow-up care, fewer interpretation errors and better patient diagnosis comprehension, better clinical outcomes (fewer obstetrical interventions, better hemoglobin A1C, lipid levels, creatinine levels), and greater patient satisfaction.
- Study populations and types of outcomes were too varied to permit meta-analysis.
- As the developer notes, the study populations are not specific to CMC, but indicates it would expect that an intervention to improve communication, associated with improved outcomes, would be at least as beneficial in patients with greater complexity as in those without complex conditions, if not more so.

▪ The developers also cite four additional studies, including an RCT, that demonstrate improved outcomes with the presence of a medical interpreter, including (depending on the study): significantly greater degrees of patient-reported understanding and satisfaction with communication; greater satisfaction with communication among ED physicians and nurses; fewer hospitalization; decreased length of hospital stay and lower risk of 30-day readmission; and decreased resource utilization.

o FECC 16: Caregivers of CMC should report that the child's main provider created a shared care plan for their child: Seven randomized controlled trials (two in children), three non-randomized controlled trials, six uncontrolled interventions with a pre-post comparison, a non-systematic review including unpublished program evaluations, and a consensus statement from the AAP support that interventions that include a shared care plan are associated with improved outcomes among children and adults with chronic disease or medical complexity. Of note, most identified studies evaluated outcomes associated with shared care plan use in the context of larger care coordination or disease-specific management interventions; however, the shared care plan was generally a central feature of the multi-factorial intervention.

▪ In the previously cited Farmer RCT, the intervention included a care coordinator who worked with the family to develop and implement a written health plan for the child to provide coordination with doctors and home visits/ telephone support.
- In the between-group analyses, participants in the intervention group reported significantly lower needs for information.
- In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs. There was a significant improvement in satisfaction with specialty care and care coordination, significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status, and a significant decrease in personal and family strain.

▪ In a different RCT involving children (Lozano 2004), the authors conducted a multisite cluster (N-678). In the intervention, asthma nurses conducted an assessment, developed individualized shared care plans with the family, and provided self-management support and telephone follow-up., which resulted in significantly fewer asthma symptom days, fewer oral steroid bursts per year, and greater controller adherence (by parent report).

▪ The developers cite additional RCTs in the adult population that demonstrate that an individualized shared care plan results in (depending on the study): better general health, vitality, social functioning and mental health; significantly fewer ED visits; fewer hospital admissions in a pre-defined group at high risk for admission; better illness self-management; knowledge of illness-related resources; lower symptom-related distress; higher self-rated health; improvements in clinical measures of depression; adherence to therapy; and improved quality of life, functional status, and management of co-morbid diseases.

        - The developer also cites six additional studies (some and a non-systematic review (adults, 29 programs) that found an intervention of a shared care plan was associated with (depending on the study): decreased hospitalizations, costs of care, unmet needs, work loss and school absences, and increased ED use, outpatient visits, satisfaction with services, and cost savings.

    - FECC 17: Caregivers of CMC should report that the child's main provider created an emergency care plan for their child: A consensus statement from the American Academy of Pediatrics supports the importance of having an emergency care plan for children with complex medical problems for optimizing outcomes.
        - The developer cites a consensus statement from AAP on the importance of having an emergency care plan for children with complex medical problems in order to optimize outcomes.
        - The developer does not indicate whether the consensus statement was only expert opinion-based or included empirical evidence and/or some review of the quality, quantity, and consistency of the evidence.

- **Per the NQF Algorithm for Evidence**:
    - Only FECC 15 cites underlying evidence involving an independent systematic review. The quality, quality, and consistency of that review are described by the developer, but no grading system was reported by the developer. The eligible ratings for FECC 15 are HIGH, MODERATE, or LOW depending on the assessment of the strength of evidence (box 5a.)
    - The evidence for FECC 17 provided by the developer does not include empirical evidence (box 5->box 10). The eligible ratings for FECC 17 are INSUFFICENT WITH EXCEPTION or INSUFFICIENT.
    - For all other FECC measures, the developer provides empiric evidence and did conduct its own review indicating quality, quantity, and consistency, and graded the evidence. If the Committee constitutes these as systematic reviews qualifying for the path box 3->box 5a, the eligible ratings for these FECC measures also are HIGH, MODERATE, or LOW, but if the Committee does not, the highest eligible rating path is MODERATE (box 7->box 9)

Questions for the Committee

- *For each of the 10 FECC measures, is the evidence directly applicable to the process of care being measured? Please identify by FECC number any concerns you have about evidence for that measure.*
- *For each of the 10 FECC measures, is the process of care proximal and closely related to desired outcomes? Please identify by FECC number any concerns.*
- *For FECC 5, the evidence cited is an RCT that included at least monthly contact, but the measure requires only contact in the last three months. Does the Committee wish to seek additional comment/justification from the developer on the discrepancy between the evidence and the specifications?*
- *For FECC 7, the evidence cited is no RCT and three intervention studies, but the developer notes 2 of 4 papers did not track appointments; no information is provided on appointment timing for the other two studies. Does the Committee wish to seek additional comment/justification from the developer on the discrepancy between the evidence and the specifications, which require that the appointment with the specialty service provider occurs within a specific timeframe (3 months).*
- *For FECC 9, the evidence cited focuses on a written summary, but the developer notes two of the citations do not specify the components of the summary; no information on the presences or lack of elements is provided for the third reference. Does the Committee wish to seek additional comment/justification from the developer on the six specific elements that must be present in the written summary to meet the measure specifications?*
- *For FECC 14, the study cited focuses on the importance of healthcare provider-school communication, but the developer does not provide information on whether the communication resulted in improved outcomes or at what frequency the communication occurred. Does the Committee wish to seek additional comment/justification from the developer on whether outcomes were improved or the requirement in the specification that the communication occur at least annually?*

- *For FECC 17, are you aware of empirical evidence that might support a rating of MODERATE? Does the Committee wish to seek additional comment/justification from the developer about the details of the AAP consensus statement?*
- *Do you believe FECC 17 (eligible for INSUFFICIENT WITH EXCEPTION) or any other measure needs separate voting (because some of above you judge HIGH, some MODERATE, and others LOW) at the in-person meeting? If so, please indicate which one(s) and why. In general, NQF would otherwise ask only for a single vote on Evidence for all measures.*

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer has provided the following information:

- Although CMC comprise only 13% of the pediatric population, they account for a disproportionately high (26-49%) of hospital days and 70% of overall health expenditures.
- Little is known about the quality of care coordination received by CMC. Present assessments of care coordination are generally limited to whether care coordination was received or not, without any attempt to identify potentially beneficial components of care coordination or the manner in which they were delivered. The evidence that is available suggests that 29-41% of parents of children with special health care needs report not getting needed help with care coordination; little is known about the quality of the help that is being received.
- While limited information on quality of care coordination exists, data do demonstrate disparities in receipt of care coordination. Latino and Black children have been found to be more likely to have unmet care coordination needs compared to non-Hispanic white children. In addition, children from families with limited English proficiency have reported higher unmet care coordination needs and greater difficulty getting needed referrals compared to English proficient families.

Questions for the Committee (as appropriate):

- *Is there a gap in care that warrants a national performance measure for each of the 10 FECC measures?*
- *Should any of the 10 FECC measures be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b)

1a. Evidence.

- Regarding the presence of a care coordinator (FECC 1), there is a lack of clarity about what that person does (as process) versus exists (structurally). For example, many insurers assign care coordinators for high utilizing patients. That coordinator would have different responsibilities than a clinically assigned care coordinator. This would then impact the subsequent measures substantially.
- For many of the measures, there is not a lot of empirical evidence to go on (one RCT). This should be considered in relation to the time points for the outcomes. For example, #5 and #7, why is 3 months the appropriate interval?
    - For #9, Medicaid Meaningful Use requires after visit summaries although not as comprehensive as this.
    - #14: not outcome driven. Study on children with new onset brain injury who are likely very different than other children with complexity
    - #15: this is positive but not a lot else on cultural competence
    - #17: not outcomes based
- Yes, for each measure the evidence is directly applicable to the process of care being measured.
    - FECC16: The evidence provided seems to demonstrate that the care coordinator created the shared care plan and not necessarily the main provider yet the question specifically identifies the main provider.
    - As already identified by the NQF review staff, neither FECC14 or FECC17 directly identify the impact on outcomes. It would be satisfactory if the developers provided evidence for FECC14 using studies of children with chronic conditions. The developers should contact the AAP around evidence to support FECC17

    - FECC5: Yes, it would be good to hear from the developer why “in at least three months” was considered adequate.
    - FECC7: Yes , it would be good to hear from the developer why they chose the specific 3 month timeframe for appointment without justification from evidence provided. Though I suspect most guidelines would recommend quarterly visits.
    - FECC9: NO, these six are pretty standard for written summaries.
    - FECC14: No. studies in most chronic disease populations show that interaction with the school improve outcomes. At least annually shows that the school is being apprised of the student’s condition for each new grade they enter. This does not preclude notifications when conditions change.
    - FECC17: AAP consensus statements are strong evidence and are usually themselves based on systematic reviews. Additional details on how the consensus statement was developed would be helpful. I do not believe using this as evidence requires a moderate designation and requires that each measure be evaluated separately
- FECC 5 & FECC 7- Would like insight as to the rationale for the timeframes associated with each measure.
- FECC 9 - Assume that the specific elements that should be present in the summary constitute what is usually found in a written summary?
- FECC 14 - Is there perhaps additional evidence that could be looked at regarding children with chronic (not necessarily complex) conditions and how communication with school has improved outcomes for children (fewer missed school days, etc.).
- FECC 17 - Would like to see the actual AAP statement that supports the measure.
- In general I was concerned that there was no reference in the measures to continuity of care or cultural competence, other than translation.
    - For FECC-5, it is unclear if caregiver concerns include the health of the caregiver in order to provide care; more people enter institutional care due to caregiver burnout rather than deterioration of their condition.
    - For FECC-7, clarification on whether the assistance with appointments includes complex care scheduling (e.g. multiple specialists on one site) would be helpful.
    - For FECC-9, some missing components include a list of hospitalizations, behavioral plan in the chart if applicable, and letters to emergency room providers if applicable (e.g. adrenal insufficiency.) In addition, all allergies (e.g. latex, food, etc.), not just medication allergies, should be included.
    - For FECC-15, I strongly support professional interpretation as often children are asked to do this for families resulting in misinterpretation of medical information and loss of privacy.
    - For FECC-16, clarification is need on who is the "main" provider as in some cases it is the pediatrician but for others it could be a specialist especially if the condition affects all other health conditions. For example, a child with a renal transplant would need to consult with nephrology prior to imaging studies, or if medications are being recommended for a secondary condition.
    - For FECC-17, in addition to an emergency care plan, children with medical complexity also need a plan for emergency preparedness (e.g. natural disaster) as recommended by the American Academy of Pediatrics (AAP – see http://pediatrics.aappublications.org/content/125/4/829.full.) I would like the source citation on the "consensus statement" as I was unable to locate this information.
- Regarding the NQF algorithm for evidence, I am somewhat concerned that there is "no grading system" for FECC-15. I am concerned that FECC-17 has "no empirical evidence."
- Regarding the questions to the committee, again my main concern is lack of information on FECC-17. I do not think that some of the evidence from adult studies is applicable; for example certain conditions such as end stage renal disease affects children differently and can result in growth and cognition being affected. I agree that the committee needs more information on FECC-5 which cites monthly contact yet differs from the measure requiring three months. I also agree that more information is needed on FECC-7 regarding the timing of the appointments if the measure timeframe is three months. I think more information is needed under FECC-9 regarding the components of the written summary. I agree that the committee needs outcome information from the result of provider-home communication. As stated above, I agree that the committee needs more information on FECC-17 regarding empirical evidence and the details of the AAP consensus statement. I would recommend a separate vote on FECC-17 only if this information isn't provided.

- The body of evidence for measures 1, 3, 5, 7, 8, and 9 is overall weak with most of it repeatedly relying on one RCT that involved 100 children for a 6 month intervention trial. Given the inherent long term health issues and outcomes when dealing with children with medical complexity, this evidence is at best tangential and minimal.
- The evidence for measure 14 is barely tangential as it is one trial with 66 children with traumatic brain injury. This is insufficient to justify a broad measure asking for communication on educational impacts of child's health for all children with medical complexity.
- Evidence for measure 15 and 16 is acceptable.
- There is no evidence provided for measure 17, only a consensus statement.

1b. Performance Gap.

- This population (less than 1% of children) is high utilizing. This utilization may be mutable (from inpatient to outpatient). There are likely more data on disparities that could be highlighted.
- While more information could have been provided, there is definitely a gap in care that warrants the development of the FECC questions presented. Evidence also shows racial disparity in the care of medically complex children but none of the questions seem to be disparities sensitive.
- Overall, this should be considered a low occurrence/high risk population - noting the low occurrence in the overall pediatric population but the high number of hospital days and expense. Potentially all of these measures could highlight disparities within this population, especially if measure results are reported in a stratified manner.
- I support the developer's information regarding disproportional high utilization of this subgroup. Disparities for the underserved populations noted result in delayed diagnosis, even for life-threatening conditions, increasing morbidity and mortality. Regarding questions for the committee, I agree that there is a gap in care for these populations, as well as Asian families. The measures I would most recommend to address disparities would be FECC-1 having a care coordinator and FECC-15 access to professional translation but would like to see disparities addressed in all measures.
- No performance data supplied. Measures developers provided a few very general statements suggesting their is gap but overall minimal is known about care coordination, for any patient let alone one with medical complexity.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The data sources are administrative and caregiver survey.
- Each of the ten measures included in this set has a different numerator:
  - FECC-1: Caregivers of CMC should report that their child has a designated care coordinator.
  - FECC-3: Caregivers of CMC who report having a designated care coordinator and who require community services should also report that their care coordinator helped their child to obtain needed community services in the last year.
  - FECC-5:Caregivers of CMC who report having a care coordinator and who report that their care coordinator has contacted them in the last 3 months should also report that their care coordinator asked them about the following: Caregiver concerns; Health changes of the child
  - FECC-7: Caregivers of CMC who report having a care coordinator for their child should also report that the care coordinator assists them with specialty service referrals by ensuring that the appointment with the specialty service provider occurs
  - FECC-8: Caregivers of CMC who report having a care coordinator should also report that their care coordinator: Was knowledgeable about their child's health; Supported the caregiver; Advocated for the needs of the child
  - FECC-9: Caregivers of CMC who report receiving a written visit summary during the last 12 months from

their child's main provider's office should report that it contained the following elements: Current problem list; Current medication list; Drug allergies; Specialists involved in the child's care; Planned follow-up; What to do for problems related to outpatient visit
    - FECC-14: Caregivers of CMC who report their child's condition causes difficulty learning, understanding, or paying attention in class should also report that one of their child's health care providers (i.e., primary care physician, specialist physician, care coordinator, nurse practitioner, nurse, social worker, etc.) communicated with school staff at least once a year about the educational impacts of the child's condition.
    - FECC-15: Caregivers of CMC who self-identify as having a preference for conducting medical visits in a language other than English should have access to a professional medical interpreter (live or telephonic) at all visits for which an interpreter is needed.
    - FECC-16: Caregivers of CMC should report that their child's primary care provider created a shared care plan for their child.
    - FECC-17: Caregivers of CMC should report that their child's main provider created an emergency care plan for their child.
  - The denominator for the measures is: The eligible population of caregivers for the FECC Survey overall is composed of those who meet the following criteria:
    - Parents or legal guardians of children 0-17 years of age
    - Child classified as having a complex, chronic condition using the Pediatric Medical Complexity Algorithm (PMCA) (see Simon TD, Cawthon ML et al. 2014)
    - Child had at least 4 visits to a healthcare provider over the previous year
    - While some of the FECC measures only apply to a subset of the overall eligible population for the survey (e.g., measures related to the quality of care coordination services provided are only scored for those caregivers who endorse having a care coordinator), eligibility for these quality measures can only be gleaned from responses to the FECC Survey itself.
- The denominator details, including ICD-9 codes, are included in the detailed measure specifications (see Excel file – NQF_detailed specs…xls).
- The measure is risk-adjusted using case mix adjustment.
- The developer recommends adjusting for survey mode (telephone-only vs. mixed mode) and for respondent education level.

Questions for the Committee (as appropriate):

- *Are all the 10 FECC survey questions being considered (1, 3, 5, 7, 8, 9, 14, 15, 16, 17) clear, unambiguous, and at an appropriate comprehension level?*
- *Are all appropriate codes included for the denominator?*
- *Is the logic or calculation algorithm clear?*
- *Can each of the 10 FECC measures be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer provides the following:

- The measures were tested at the performance score level.
- Empirical testing of the denominator was conducted by validity testing at the data element level using the Pediatric Medical Complexity Algorithm. Per NQF policy, validity testing at this level does not require additional reliability testing.
- The measures within each domain are not meant to function as a scale, since they do not measure a single underlying construct, but instead measure separate aspects of care coordination quality. Accordingly, the

developers do not present any measurement of the reliability within domains. The 10 FECC measures in this submission may be used independently of one another.

- Caregivers of CMC insured by Medicaid in Washington and Minnesota were surveyed
  - The developer indicates in at least one instance that 3,000 surveys in each state were deployed, but in others indicates 1,500 per state.
  - 600 completed surveys were returned in Washington, and 609 from Minnesota.
  - The developer used a subset of the overall participants for its reliability analysis (n=889) and states this subset was nearly identical to the overall population; data on the subset as compared to the full population are provided in Table T1.
- For the measures that function as a scale, a variation on Cronbach's alpha was used, polychoric ordinal alphas, since the measures are ordinal. The polychoric ordinal alpha statistic is a disattenuated Cronbach's alpha for ordinal scales and the commonly accepted rules for describing internal consistency may be employed: $\alpha \geq 0.9$ = excellent; $0.9 > \alpha \geq 0.8$ = good; $0.8 > \alpha \geq 0.7$ = acceptable; $0.7 > \alpha \geq 0.6$ = questionable; $0.6 > \alpha \geq 0.5$ = poor; $0.5 < \alpha$ = unacceptable. Similar to Cronbach's alpha, the polychoric ordinal alpha statistic increases with the number of items on the scale—i.e., it may be lower with fewer items on the scale
- Score reliability was tested by calculating intra-class correlation coefficients.
- The reliability results are presented in Table T2 and narrative)
  - The developer notes that for FECC 5, 8, and 9, which are the multi-part measures, the polychoric ordinal alpha reliability results ranged from acceptable to good.
- For reliability at the performance score level, the developer calculated the intraclass correlation coefficent by affiliated group practices. The per entity recommended sample size was >30. The affiliated group practice reliability testing at the performance score level was conducted (although not the intended Level of Analysis) because the developer's state/Medicaid plan-level testing involved only two units. The results are presented in Table T3.
  - The developers report ICC reliability testing at the practice level generally ranged from acceptable to excellent for 8 of the 10 measures.
  - FECC 3 and FECC 15 did not achieve good reliability at the performance score level, which the developer attributes to small sample size. ICC's also are not provided for FECC 5 and FECC 8, but FECC 5 and FECC 8, as multi-part measures, have polychoric ordinal alphas reported.
- **Per the NQF Algorithm for Reliability**, empirical testing at the level of computed performance score level may be rated HIGH, MODERATE, LOW, OR INSUFFICENT, depending on the results.

Questions for the Committee (as appropriate)

- *Is the test sample adequate to generalize for widespread implementation? For the ICC's, N ranged from 28-103 for affiliated group practices, depending on the measure; N for patients ranged from 89-842, depending on the measure. Table T3*
- *Does the Committee concur with the developer that the performance score reliability testing at the affiliated group level reflects reliability at the intended Level of Analysis, which is health plan and Population: state?*
- *Do the results for each of the 10 FECC measures independently demonstrate sufficient reliability so that differences in performance can be identified?*
- *The developer reports FECC 3 and FECC 15 did not achieve good reliability, which is attributed to sample size. Should the Committee vote separately on these two measures at the in-person meeting? In general, NQF would otherwise ask only for a single vote on the Reliability for all measures. Are there any other measures the Committee wishes to vote on separately?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The goal of the measure is to assess the quality of care coordination received by children with medical complexity.

- As noted in the Evidence section, the developer provides information for each measure.

Question for the Committee:

- *Are the specifications consistent with the evidence?*
- *Does the Committee wish to comment on the relationship between the evidence and specifications (as noted in the Evidence section questions above) for FECC 5, FECC 7, FECC 9, and FECC 14?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer reports the following:

- Validity testing was done at the critical data element level (denominator) and the performance measure score levels.
- Empirical validity testing at the data element testing was performed for the denominator data elements (PMCA) for all measures using the data sources. Specifically, the algorithm-determined classifications of 700 children (no chronic disease, non-complex chronic disease, or complex chronic disease) were compared to a classification determined by clinician chart review, which NQF considers as a gold standard. The results were:
  - Seattle Children's Hospital data: Sensitivity (95% CI) = 84 (80-88); Specificity (95% CI) = 92 (89-94)
  - WA Medicaid data: Sensitivity (95% CI) = 89 (85-82); Specificity (95% CI) = 85 (81-89)
- Content/face validity of the measures was established using the RAND-UCLA modified Delphi method. Aggregated among the factors the expert panel was asked to consider is whether they would consider providers who adhere more consistently to the quality measure to be providing higher quality care. We interpret this to indicate an assessment of the performance score level, as required by NQF, although it was aggregated with other factors such as evidence. **Per the NQF Algorithm for Validity**, the highest rating based on this alone is MODERATE.
- The measures underwent additional testing through cognitive interviews (9) in English or Spanish to establish understandability by families.
- The developers field tested for convergent validity against two previously validated measures: the Clinician and Group (CG) Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Child 12-month Survey, and a measure adapted from the Adult Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Heath Plan 4.0 supplemental item on care coordination.
  - The developer used linear regression to examine the association between measure scores and the two CG-CAHPS measures and the one adapted CAHPS measure described above, unadjusted and adjusted for caregiver education and assigned survey mode. The analysis was carried out for each FECC measure.
  - Table T6, Table T7, and Table T8 present the results for each measure against the validation metrics. The developer reports unadjusted and adjusted (for mode of administration and caregiver education). In summary, the developer reports:
    - All 10 FECC measures in this submission were associated with better experience in terms of access to care in both unadjusted and adjusted analyses.
    - All but FECC 15 was significantly associated with overall provider rating in both unadjusted and adjusted analysis. FECC 15 was significantly associated only in the unadjusted analysis, likely due to smaller sample size, per the developer.
    - All but FECC 15 was associated with getting all the care coordination help the family needed in the adjusted analyses. Again, FECC 15 was not, likely due to smaller sample size, per the developer.
    - The developer notes FECC 15 had the highest content/face validity.
    - Overall, the developer concludes all 10 measures are valid representation of quality, measuring what they purport to measure.

- **Per the NQF Algorithm for Validity**, empirical testing at the computed performance measure score may be rated HIGH, MODERATE, or LOW. For content/face validity testing, the highest rating is MODERATE.

Questions for the Committee

- *For the empirical testing at the performance score, is the test sample for each FECC measure adequate to generalize for widespread implementation?*
- *Do the results for each FECC measure demonstrate sufficient validity so that conclusions about quality can be made?*
- *Based on the results from convergent testing for FECC 15, should the Committee vote separately on it at the in-person meeting? [If relying on content/face validity, the highest possible rating is MODERATE. For all other measures, because empirical testing was at the computer performance score, the highest possible rating is HIGH.)*
- *Do you agree that the score for each FECC measure, as specified, is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The survey was not sent to families to whom exclusions apply, so the developers were unable to test the impact of exclusions. The developers note, however, that 1.1% of the identified potential sample was excluded.

Questions for the Committee (as appropriate):

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment:

- The measure is risk-adjusted using case mix adjustment. The developer recommends adjusting for survey mode (telephone-only vs. mixed mode) and for respondent education level.
- Case-mix adjustment is via linear regression for continuous measures and logistic regression for binary measures and uses the method of covariance adjustment.
- If a "clinically-adjusted" model that does not include sociodemographic variables (i.e., education) is desired, education may be omitted from the model and survey mode may be retained. To stratify clinically-adjusted scores by education, the case-mix model with survey mode as a covariate could be fit separately within each education category.
- The detailed risk model specifications are included.

Questions for the Committee (as appropriate):

- *Is the case-mix-adjustment strategy included in the measure appropriate?*
- *Are the variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all the adjustment variables present at the start of care? If not, describe the rationale provided and whether you concur with it.*
- *Does the risk adjustment model include any factors related to disparities of care? Is this appropriate? The developer indicates and "option" to exclude education from the model. To achieve comparable results across measured entities for the purpose of accountability, such options are not advisable. Does the Committee recommend inclusion or exclusion of the education variable in the case-mix adjustment?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer identified statistically significant differences by state Medicaid agency for FECC 7, FECC 9, FECC 16, FECC 17. Minimum clinically important differences have not been established for any FECC measures, but for FECC 9,

FECC 16, and FECC 17, the difference was at or close to 10 points on a 100-point scale, which the developer concludes is almost certainly of clinical import.

- The developer noted that the remaining six measures did not show differences, but hypothesized this could be due to limited samples.
- The developer identified racial/ethnic and linguistic disparities in FECC measure scores. Many of these differences were also in the realm of 10 points or more, which would be of clear clinical import.

Questions for the Committee

- *Does each of the 10 FECC measures identify meaningful differences about quality?*
- *Given the disparities in FECC scores based on race/ethnicity and language, does the Committee wish to discuss case-mix adjustment for these variables with the developers?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

- The Washington and Minnesota Medicaid IRBs did not permit the use of demographic data for non-respondents, so the developers were unable to compare respondents and non-respondents. The developers state that one would expect lower response rates for low SES and non-English speaking caregivers, but that they did achieve meaningful participation from these groups.
- The overall response rate was 40% and the sample was racially and linguistically diverse.
- The developers recommended a number of strategies for reducing non-response bias.
- The developers opted to score measures only if all component items were answered, and to only score measures where caregivers had provided a definitive response, to avoid making assumptions.

Question for the Committee

- *Does the Committee have concerns about the impact on validity given the inability to examine demographic data of non-respondents?*

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- Unclear as to individuals with developmental disabilities. Complexity conceptually has disability imbedded in it but the use of codes doesn't capture this. People with DD have high rates of unmet need and often have several comorbid conditions. At risk for worse health outcomes.
- Social factors aren't included in the modeling but do impact the outcomes of patients.
- It is not clear that the measure numbers directly correlate with the survey question numbers. For example, FECC 1 talks about a "designated care coordinator" and survey Q1 refers to "provider". My assumption is that the care coordinator and the provider are not necessarily one in the same.
- Given that the developer states that these measures can be used independently of each other, should they be voted on as such?
- Thank you for the clarification regarding the use of ICD-9 in order to have 12 month retrospective data (so that ICD-10 could be used 10/16.) This was a bit confusing until the reference on the algorithm code list.
  Regarding the questions for the committee, after this clarification I agree that the appropriate codes are included in the denominator and the logic is clear. I think that all of the measures can be "consistently implemented."
- Validity testing is weak, only comparing to a couple domains CG CAHPS and one question HCAHPS.

2a2. Reliability testing

- There is an issue with the move from ICD-9 to 10. Also the PMCA hasn't been comprehensively tested across states and other payors.
- Thank you for the clarification that the measures in each domain are "not meant to function as a scale...but instead

measure separate aspects of care coordination." I am concerned with the results for FECC-3 and FECC-15 due to small sample size.
- Regarding questions for the committee, I think that the "test sample is adequate to generalize for widespread implementation" except for FECC-3 an FECC-15. Other than those two, it seems that the reliability is "at the intended level" for the measures. Differences in performance could be determined, again except for measures FECC-3 an FECC-15. Unless additional information is forthcoming on these two measures, they should be voted on separately.

2b1. Validity Specifications

- Small samples, have opportunity with test-retest with the small sample for measures such as #5 and 8.
- The test sample was adequate to generalize.
- Performance at the group level does reflect reliability at the intended level of analysis.
- It is unclear why no ICC and Spearman Brown were reported for FECC 5 and 8. Is this also because of sample size?
- Not sure I accept the sample size reason for lack for reliability for FECC 3 and FECC 15. They do not even report the results.
- I am not comfortable with the authors saying that FECC 5 and 8 clearly demonstrate reliability just because of the internal consistency alpha for the sub items of these measures.
- These questions are ideal for test / retest reliability yet this was not done even on a small scale. I believe this would have been a better assessment of reliability and would have proven favorable for FECC 5 and 8
- Not sure that I agree that these measure don't measure one or more constructs. Perhaps a principal components assessment to see if they do group in ways that make sense.
- Am on the fence as to whether FECC5 and 8 should be evaluated separately just on reliability grounds.
- It appears that the specifications are consistent with the evidence/
- Please see comments in evidence above for FECC-5, FECC-7, FECC-9. FECC-5 needed clarification on caregiver health and FECC-7 needed clarification on complex care scheduling. I'm most concerned with the missing components of FECC-9 (allergies, hospitalizations, emergency letters.) Again, I agree that for FECC-14, more evidence is needed on outcomes of provider-school communications. Otherwise the "specifications are consistent with the evidence."
- Overall samples were very small so unclear why a sample of cases were used here as opposed to all cases.

2b2. Validity Testing

- I worry about the small sample size. Agree that #15 should be separately looked at.
- Validity testing was very comprehensive and covered content, face and convergent validity.
- The results for each measure do seem adequate to generalize to widespread implementation.
- All validity values were high and demonstrated clear validity for each measure.
- The validity assessment was quite good and it would be unfair to deny a high score due to the lack of convergent validity for FECC15. Access to a medical interpreter is the most likely question not to correlate with the other measures of coordination.
- Each of these measures assess the quality of care coordination.
- It does appear that the measure results are valid. As far as being an indicator of quality, linking these results to an outcome of interest (such as fewer admissions or office visits) would get at the true impact of care coordination.
- I appreciate the additional information on empirical validity testing, content/face validity (RAND-UCLA modified Delphi method), and "understandability for families." I understand that convergent validity was tested using CAHPS (Consumer Assessment of Healthcare Providers and Systems) and a "measure adapted from the Adult" version on care coordination. Here again there is difficulty with validity testing of FECC-15 due to small sample size though there is face validity.
- Regarding questions for the committee, all of the test sample, with the exception of FECC-15, was adequate.. The results for each measure, except FECC-15, showed sufficient validity. Unless more information is forthcoming, I agree that a separate vote on FECC-15 is needed. I agree that the scores, except fo FECC-15, indicates quality.

2b3-2b7. Threats to Validity

- Survey data that tends to have low response rates probably has lower response rates among those at the greatest risk.

- Case adjustment doesn't necessarily capture all the areas that would help risk adjust. This is a general weakness that won't be overcome.
- Exclusion criteria was appropriate and minimal
- The methodology used for risk adjustment was appropriate as well as how missing data was treated. There was really no rationale as to why they chose education and mode of administration for risk adjustment. Also when risk adjustment is appropriate was not addressed (ie just for reporting of aggregate means and SDs). Finally information on the impact of risk adjustment would be helpful (the coefficients look small for items with a range of 0 to 100)
- Adjustment variables should be present at the start of care, for education level but not mode of administration which is okay.
- Education level is related to disparities in care. The coefficients are not that large so see no reason not to recommend excluding.
- Meaningful difference is hard to identify with only two state comparisons. However there does seem to be sufficient variability for each of the measures to be able to measure significant differences.
- No concerns about validity due to inability to examine demographic data for non-responders
- Missing data do not constitute a threat as there were so few
- It appears that the developers did all they could given their IRB limitations.
- For 2.b.3. I understand that only 1.1% were excluded. Regarding questions for the committee, it seems that the exclusions are "consistent with the evidence." It doesn't appear that there were any inappropriate exclusions with the possible exception of allowing caregiver assistance with responses for some individuals who may need clarification of survey questions (e.g. developmentally disabled) instead of just noting inability to complete. Indeed a recent study indicated health disparities are highest for individuals with developmental disabilities (seehttp://www.cdc.gov/ncbddd/disabilityandhealth/features/unrecognizedpopulation.html.) There is a high correlation (see http://www.sciencedirect.com/science/article/pii/S1936657410000373) between developmental disabilities and secondary comorbid conditions so this could affect a sizable portion of the sample population. Otherwise, the exclusions seem to be of "sufficient frequency and variation."
- For 2.b.4. I understand that a case mix adjustment was used. Regarding questions for the committee, I disagree with "adjusting for survey mode" e.g. phone vs. mixed mode. The variables are "adequately described" and "present at the start." The adjustment would address disparities (e.g. access to the survey) but I strongly disagree to exclude education.
- For 2.b.5 Here again FECC-17 is a problem as "minimum clinically important differences have not been established." There is also a concern of six measures not showing differences "due to limited samples." Regarding questions for the committee, again the concern is FECC-17 for the reasons above. The committee should discuss case-mix adjustment as it relates to race, ethnicity, and language.
- For 2.b.6 n/a
- For 2.b.7 The lack of demographic data is concerning. Although it was reassuring that there was "meaningful participation" for low SES and non-English speaking caregivers. I agree with scoring measures "only if all component items were answered." Regarding questions for the committee, I agree that lack of demographic data will adversely affect validity.
- Yes as measure relies on caregiver survey which may results in higher non-response rates among lower SES/educational level. This is same population one would predict is at higher risk for uncoordinated care if they have a CMC

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer reports the following:

- The data are currently collected via caregiver survey, which is expensive and time-consuming; however it is currently the most valid approach for collecting data on care quality for children with medical complexity. Administrative data (billing data) are used to identify children eligible for the FECC.
- Because this is new measure, limited data are available on feasibility. The developer notes it achieved a good response rate (40%) during field testing. The developer recommends a mixed mode approach (contact by mail and telephone), since this had the lowest refusal rate.

Question for the Committee

- *Since the developer recommends a mixed mode approach, does the Committee wish to discuss with the developer limiting administration specifications to this approach (which also obviates the need to adjust for this variable)?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Note small scale testing has been done. maybe be challenging to operationalize
- I am not sure limiting to mail and phone is necessary given the different contexts and patient groups for which the survey is appropriate. It may also be feasible to administer to parents during a child's visit or hospitalization so in person can also be an option.
- Administration is feasible given the types of questions.
- Survey tools/measures can be difficult (time consuming, expensive, etc.) to implement. However, the information that could be gained through use of this tool to improve care coordination (and potentially outcomes) for this specific population could be very important.
- Clarification is needed on if respondents were assured of confidentiality or ability to skip a question if concerned. It is also unclear if there were any incentives, even intrinsic motivation, used to increase response rate. Regarding the questions for the committee, the committee should discuss the mixed mode approach for clarification purposes.
- No data provided. Overall these are cumbersome measures with denominators relying on complex algorithms using ICD codes and numerators relying on surveys. Given these hurdles, predict there would be minimal uptake of these measures. On top of these significant feasibility concerns, the evidence behind these measures and the broad scale testing is weak.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

The developer reports the following:

- This measure is currently in use for internal quality improvement by a number of organizations, including children's hospitals, universities, and health plans.
- This is a relatively new measure and is not currently used for public reporting, but the developer reports that it is being widely distributed and they expect it will be used for public reporting in the next few years.
- The developer is unaware of any unintended consequences.

Questions for the Committee

- *Will performance results on each of the 10 FECC measures further the goal of high-quality, efficient healthcare?*
- *Do the benefits of each of the 10 FECC measures outweigh any potential unintended consequences?*

'Committee pre-evaluation comments

Criteria 4: Usability and Use

- I see no issues with usability
- Given the limited data available, the benefits of this survey tool and they associated measures could outweigh the potential consequences. Outside of the "wide distribution" mentioned, are there other formal plans to implement/test to tool and measures?
- Not currently in use likely given poor feasibility, minimal evidence and small scale testing to date.

Criterion 5: Related and Competing Measures

The following measures are related and not harmonized:

- 0009 : CAHPS Health Plan Survey v 3.0 children with chronic conditions supplement
- 0718 : Children Who Had Problems Obtaining Referrals When Needed
- 0719 : Children Who Receive Effective Care Coordination of Healthcare Services When Needed

According to the developer:

- The currently available NQF-endorsed measures related to care coordination and care for children with chronic conditions are related to, but fundamentally different from, the quality measures addressed in the FECC measure set.
  - The measures differ with regard to target population. The currently-endorsed measures address children with chronic conditions (0009), children who have received a referral to specialty services (0718), and children who received care from at least 2 types of health care services (0719), while the FECC measures address children with medical complexity. While the other measures likely apply to CMC (in addition to many other children), the FECC measures are specific to CMC.
  - The FECC measures differ from currently-endorsed measures with regard to focus. The currently-available measures largely focus on whether families who needed specialized services for their child found it easy or difficult to obtain them and whether anyone in their health plan or child's doctor's office/clinic helped them to get that service. The FECC measures focuses more on the quality of services provided by a family's self-identified care coordinator, delving into the specific care coordination attributes and processes that have been associated with better outcomes in the literature.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Family Experiences with Coordination of Care (FECC) Measure Set

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/29/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention

(with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: quality of care coordination process measures for children with medical complexity

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

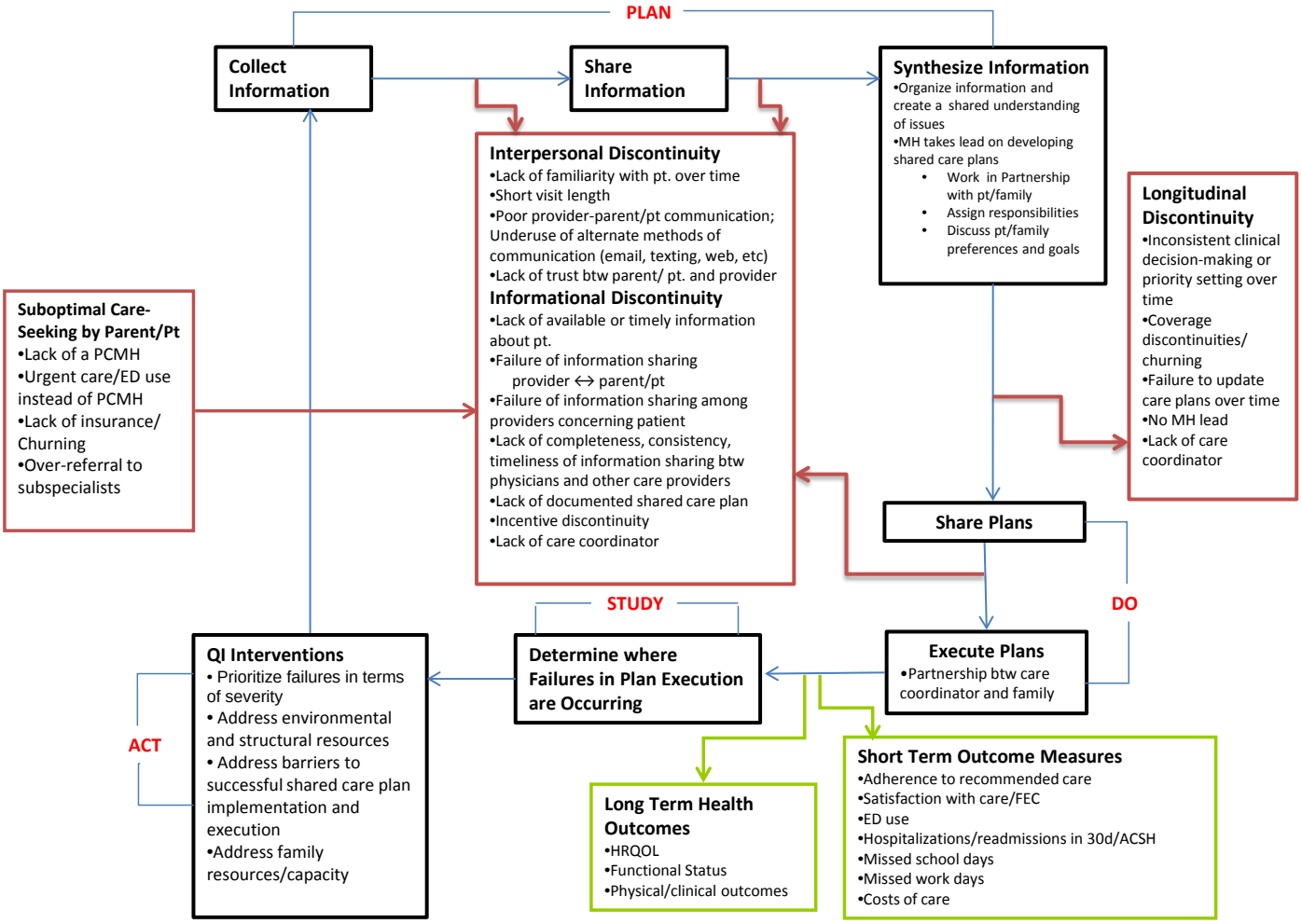
Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The conceptual framework below diagrams the relationships between care coordination/fragmentation and both long- and short-term health outcomes for children with medical complexity (CMC).

Conceptual Framework for Care Coordination/Fragmentation in the Context of the PCMH for Children with Complex Needs



The 10 quality measures included in this submission set assess care coordination processes (outlined in the black boxes above) associated with better outcomes for CMC (outlined in the green boxes above). The specific relationships between each quality measure and the care coordination processes included in this conceptual framework are detailed here:

ID	Indicator description	Importance to Outcomes
FECC 1	Has a care coordinator	Related to all actions outlined in black boxes above
FECC 3	Care coordinator helped to obtain community services	Related to executing plans
FECC 5	Care coordinator asked about concerns and health changes	Related to collecting information
FECC 7	Care coordinator assisted with	Related to executing plans

	specialist service referrals	
FECC 8	Care coordinator was knowledgeable, supportive & advocated for child's needs	Related to collecting information, synthesizing information, sharing plans, executing plans, determining where failures occur, and QI interventions
FECC 9	Appropriate written visit summary content	Related to synthesizing information and sharing plans
FECC 14	Health care provider communicated with school staff about child's condition	Related to sharing information and sharing plans
FECC 15	Caregiver has access to medical interpreter when needed	Related to all actions outlined in black boxes, above
FECC 16	Child has shared care plan	Related to sharing and synthesizing information and sharing and executing plans
FECC 17	Child has emergency care plan	Related to sharing and synthesizing information and sharing and executing plans

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

For development of the FECC Survey quality measures, we began by first developing a conceptual framework for care coordination/fragmentation for children with medical complexity (see 1a.3, above). The framework indicates events that may lead to fragmented care, such as interpersonal discontinuity, where providers lack familiarity with the child's health issues, or informational discontinuity, where information needed to care adequately for the child is missing. The framework also illuminates how care coordination relates to both short- and long-term outcomes, such as emergency department utilization and health-related quality of life.

Based on the conceptual framework, we identified 6 topics for evidence review: shared care plans, goal setting, information exchange, care coordination services, continuity of care, and the patient-centered medical home. We then conducted a focused review of the relevant literature in each topic area, summarizing the evidence assessing links between care coordination structures, processes and outcomes for children with medical complexity. From those reviews, we developed draft quality measures that were supported by the identified evidence.

1a.8.2. Provide the citation and summary for each piece of evidence.

<u>FECC 1. Caregivers of CMC should report that their child has a designated care coordinator.</u>

One randomized control study (RCT), one cohort study, and five case series, case-control, or historically-controlled studies demonstrate that outcomes improve when caregivers of children with medical complexity (CMC) report that their child has a designated care coordinator.

Specifically, Farmer, et al (2011; see Evidence Form for list of references) conducted a randomized control trial with intervention for 100 children with chronic illness on Medicaid (6-month intervention supporting 32 primary care provider [PCP] offices) wherein the care coordinator worked with the family to develop a written health plan for the child to

provide access to services and coordination with doctors and home visit/ telephone support. In between-group analyses, participants in the intervention reported significantly higher satisfaction with mental health services and specialized therapies as measured by a family survey adapted from the Shared Responsibilities Tool Kit—Version 1.0, and significantly lower need for information as measured by the Family Needs Survey (FNS). In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs as measured by the FNS. There was a significant improvement in satisfaction with specialty care and care coordination as measured by the Shared Responsibilities Tool Kit—Version 1.0. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status as measured by the Functional Status II (Revised)—14 item version. There was a significant decrease in personal and family strain, as measured by the Impact on Family Scale (IFS).

Wood et al (2008) compared nurse-led practice-based care coordination (intervention) for 144 children enrolled in Title V in three practices with pre-existing agency-based care coordination (control) in three practices. Participants in the intervention reported significantly fewer barriers to getting services, as measured by survey items in which they identified needed services and then reported barriers to obtaining them. They were also significantly more likely to report improved support from the care coordinator and improved satisfaction with care coordination services.

Finally, five case series/case-control/historically-controlled studies also support the measure (Gordon et al., 2007; Palfrey et al., 2004; Farmer et al., 2005; Cady et al. 2009; McAllister et al., 2009). These studies, ranging in size from 43 to 227 children, documented that reporting that the child has a designated care coordinator is associated with (depending on the study), significant increase in Emergency Department (ED) use, significant decrease in hospitalization and length of stay, significant increase or decrease (depending on study) in outpatient visits; decreased cost of care, significant increase in satisfaction, significant decrease in lost work days, significant decrease in school absence, significant decrease in unmet needs, and/or decrease in family strain.

FECC 3. Caregivers of CMC who report having a designated care coordinator and who require community services should also report that their care coordinator helped their child to obtain needed community services in the last year.

One randomized controlled trial and two uncontrolled intervention studies demonstrate that outcomes improve when care coordinators assist families with obtaining needed community services.

In the randomized controlled trial by Farmer et al. (2011; see Evidence Form for list of references) which included 100 children with chronic illness on Medicaid (6-month intervention supporting 32 PCP offices), the care coordination intervention included, among other components, a) facilitating communication among families, primary and specialty care providers, and community service agencies and b) providing information to help the family access needed educational and community resources. While it is not possible to determine which elements of this bundled intervention resulted in improved outcomes, Farmer had several important findings that could be related to obtaining community services such as mental health services and therapies, and also found decreased unmet needs, some of which may be met by community services. In between-group analyses, participants in the intervention reported significantly higher satisfaction with mental health services and specialized therapies as measured by a family survey adapted from the Shared Responsibilities Tool Kit—Version 1.0 and significantly lower need for information as measured by the Family Needs Survey (FNS). In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs as measured by the FNS. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status as measured by the Functional Status II (Revised)—14 item version. There was a significant decrease in personal and family strain, as measured by the Impact on Family Scale (IFS).

In a pre- and post-intervention analysis of a care coordination intervention by Gordon et al. (2007) including 227 children with medical complexity, a pediatric nurse case manager and a special needs physician worked with community agencies, such as nursing and durable medical equipment companies, and insurance to ensure that children received needed services. Gordon et al. found a significant decrease in the number of hospitalizations and length of stay and an increase in the use of outpatient services. There was also a decrease in tertiary care center payments.

Finally a small study (n=43) by Cady et al. (2009) examining a telephone care coordination intervention that included coordination with community services found a significant reduction in unplanned hospitalizations in the first year, which then stabilized over subsequent years.

FECC 5. Caregivers of CMC who report having a care coordinator and who report that their care coordinator has contacted them in the last 3 months should also report that their care coordinator asked them about the following:

a. caregiver concerns

b. health changes of the child

One randomized controlled trial demonstrates that having a care coordinator that asks about the child with medical complexity's progress is associated with improved outcomes.

In the randomized controlled trial by Farmer et al (2011; see Evidence Form for list of references) which included 100 children with chronic illness on Medicaid (6-month intervention supporting 32 PCP offices), the care coordination intervention included telephone contact to discuss the child's progress at least once each month as one component of the intervention. Farmer et al. had several important findings that suggest that caregiver concerns were being addressed; namely a significant decrease in personal and family strain, as measured by the Impact on Family Scale (IFS). There was also a significant improvement in satisfaction with specialty care and care coordination as measured by the Shared Responsibilities Tool Kit—Version 1.0. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status as measured by the Functional Status II (Revised)—14 item version.

FECC 7. Caregivers of CMC who report having a care coordinator for their child should also report that the care coordinator assists them with specialty service referrals by ensuring that the appointment with the specialty service provider occurs within 3 months of referral initiation.

One randomized controlled trial and three uncontrolled intervention studies demonstrate that outcomes improve when care coordinators assist families with making sure specialty service referrals are successfully completed.

In the randomized controlled trial by Farmer et al (2011; see Evidence Form for list of references) which included 100 children with chronic illness on Medicaid (6-month intervention supporting 32 PCP offices), the care coordination intervention included a) facilitating communication among families, primary and specialty care providers, and community service agencies and b) direct advocacy for needed care, as required. While Farmer et al. did not track the completion of appointments; there were several findings that suggest that families who received the intervention were receiving needed services. In between-group analyses, participants in the intervention reported significantly higher satisfaction with mental health services and specialized therapies as measured by a family survey adapted from the Shared Responsibilities Tool Kit—Version 1.0. In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs as measured by the FNS. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status as measured by the Functional Status II (Revised)—14 item version.

In a pre- and post-intervention analysis of a care coordination intervention by Gordon et al. (2007) including 227 children with medical complexity, the pediatric nurse case manager and special needs physician worked with specialists involved in the child's care, prepared a plan of care and facilitated communication among specialists and PCPs. Gordon et al. did not track the completion of appointments, but it is plausible that these activities facilitated appointments. Gordon et al. found a significant decrease in the number of hospitalizations and length of stay and an increase in the use of outpatient services. There was also a decrease in tertiary care center payments.

A second pre- and post-intervention analysis of a care coordination intervention conducted by Palfrey et al. (2004) included 117 children with complex and/or chronic medical conditions from 6 primary care practices. The intervention

was structured to improve coordination and communication among primary care providers, specialists and families and included actions to expedite referrals and communication with specialists. Palfrey et al. found a statistically significant decrease in parents' missed work days and hospitalizations. Families also reported that it was significantly easier to obtain services as measured by survey items developed by New England SERVE.

Finally a small study (n=43) by Cady et al. (2009) examining a telephone care coordination intervention that included a component related to facilitating communication between the family and tertiary care found a significant reduction in unplanned hospitalizations in the first year, which then stabilized over subsequent years.

FECC 8. Caregivers of CMC who report having a care coordinator should also report that their care coordinator:

a. is knowledgeable about their child's health

b. supports the caregiver

c. advocates for the needs of their child

One randomized controlled trial and two uncontrolled intervention studies demonstrate that outcomes improve when care coordinators are knowledgeable, supportive, and good advocates for the child's needs.

In the randomized controlled trial by Farmer et al (2011; see Evidence Form for list of references) which included 100 children with chronic illness on Medicaid (6-month intervention supporting 32 PCP offices), the care coordination intervention included a) facilitating communication among families, primary and specialty care providers, and community service agencies and b) direct advocacy for needed care, as required; and c) telephone contact to discuss the child's progress at least once each month as one component of the intervention. These activities were intended to ensure that the care coordinator was informed about the child's health and could support the caregiver and advocate for the child's needs. In between-group analyses, participants in the intervention reported significantly lower needs for information as measured by the Family Needs Survey (FNS). In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs as measured by the FNS. There was a significant improvement in satisfaction with specialty care and care coordination as measured by the Shared Responsibilities Tool Kit—Version 1.0. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status as measured by the Functional Status II (Revised)—14 item version. There was a significant decrease in personal and family strain, as measured by the Impact on Family Scale (IFS).

In a pre- and post-intervention analysis of a care coordination intervention by Gordon et al. (2007) including 227 children with medical complexity, the pediatric nurse case manager was a single point of contact for families and attended appointments, often advocating for the child and family. They also provided psychosocial support and care coordination education. These activities likely would result in the case manager having knowledge of the child's health and supporting and advocating for the child and family. Gordon et al. found a significant decrease in the number of hospitalizations and length of stay and an increase in the use of outpatient services. There was also a decrease in tertiary care center payments.

A second pre- and post-intervention analysis of a care coordination intervention conducted by Palfrey et al. (2004) included 117 children with complex and/or chronic medical conditions from 6 primary care practices. The intervention was structured to maximize family participation in care and care decisions, eliciting family goals in order to integrate healthcare with other aspects of life such as education, social services and recreation. The designated pediatric nurse practitioner visited each child at home to get an understanding of the context of the child's life, and conducted sick visits at home. This nurse practitioner also developed systems to streamline the ordering of medications and supplies and coordinated patient appointments to minimize the burden for families. It is plausible that these activities would result in the nurse practitioner acquiring knowledge of the child's health and being able to support and advocate for the child and family. Palfrey et al. found a statistically significant decrease in parents' missed work days and hospitalizations. Families also reported that it was significantly easier to obtain services as measured by survey items developed by New England SERVE.

Finally, Farmer et al. (2005) conducted a small uncontrolled intervention that included 51 children with complex chronic disease. The intervention provided care coordination, information about resources and services, emotional support, and empowerment for families to advocate for their children. There was a statistically significant decrease in specialty care, an increase in satisfaction with care coordination services and a significant decrease in missed work days and missed school days. There was also a significant decrease in unmet needs and family strain.

FECC 9. Caregivers/patients of CMC should report receiving a written visit summary following all outpatient visits in the last 12 months (or report access to a patient portal that provides a visit summary) and it should contain the following elements: (a) current problem list; (b) current medication list; (c) drug allergies; (d) specialists involved in the child's care; (e) planned follow-up; (f) what to do for problems related to the outpatient visit.

One study that used a pre/post intervention comparison design and two other expert consensus sources (medical home standards from the NCQA and guidelines from the American Academy of Pediatrics) support that caregivers of CMC should report receiving a written visit summary following all outpatient visits.

Palfrey et al. (2004; see Evidence Form for list of references) evaluated the medical home model in Massachusetts through six pediatric practices that introduced interventions to operationalize the medical home for children with special health care needs (n=117). One of the outcomes measured was receipt of a written care plan. After the intervention, more families reported that their PCP gave them a written health care plan (30% at before and 47% after, $p<0.01$). In addition, there were fewer hospitalizations (58% at baseline versus 43% after the intervention, $p>0.01$), and a decrease in parents missing > 20 days of work (26% at baseline versus 14% after the intervention, $p=0.02$). There was no change in emergency department use or school absences. However, since receipt of a written plan was itself an outcome, conclusions are limited to noting correlation of the receipt of a written care plan with other improvements in outcomes.

Additional guidance for receipt of a written plan as an important component of care coordination for CMC is supported by standards published by the NCQA in 2011. These standards call for written information to be given to patients and families as an important element of care management. More specifically, the NCQA recommends that patients/family should be given a written plan of care and a clinical summary at each relevant visit. The NCQA notes that relevant visits are "determined by the practice and the clinician" but would include visits related to important or chronic conditions (this would include well-child visits for pediatric patients), visits that result in a change in treatment plan or goals, visits that result in additional instructions or information for the patient/family, and visits associated with transitions of care. The elements of the clinical summary are not specified by the NCQA. The American Academy of Pediatrics guidelines (2005) on care coordination in the medical home for children with special health care needs states that "the medical home physician should share information among the child, family, and consultants." However, no more specific guidance (e.g. the form of communication; the content of the communication) is given.

FECC 14. Caregivers of CMC should report that one of their child's health care providers (i.e., primary care physician, specialist physician, care coordinator, NP, nurse, social worker, etc.) communicated with school staff at least once a year about the educational impacts of the child's condition.

One paper that synthesizes the authors' experience and provides guidance supports that caregivers should report that their child's health care providers communicated with school staff about the educational impacts of the child's condition.

Savage and colleagues (2004; see Evidence Form for list of references) conducted a small study (n=66) involving the treatment and recovery of children with a traumatic brain injury (TBI), and synthesized their experience to provide guidance for transitioning back into school. The authors identify the importance of having a representative of the patient-centered medical home share suggestions for easing transitions between school and medical facilities and request training for school staff working with the student regarding the condition, best practices, and related educational impacts.

FECC 15: Caregivers of CMC or CMC who self-identify as having a preference for conducting medical visits in a language other than English should have access to a professional medical interpreter (live or telephonic) at all visits for which an interpreter is needed.

One systematic review, one randomized controlled trial, two non-randomized controlled interventions, and one retrospective cohort study support that provision of professional interpreter services improves patient outcomes. While these studies do not examine outcomes among medically complex children specifically, they all included patients with a heterogeneous mix of conditions. We would expect that an intervention to improve communication, associated with improved outcomes, would be at least as beneficial in patients with greater complexity as in those without complex conditions, if not more so.

Specifically, Karliner et al (2007; see Evidence Form for list of references) conducted a systematic review of the literature to determine if the use of professional interpreters improves medical care for patients with limited English proficiency (LEP). The review included 1 randomized controlled trial and 27 cohort studies comparing professional interpreter use to another group (no interpreter use, bilingual provider use, or different types of interpreter use), published between 1966 and 2005, and assessing satisfaction, utilization, clinical outcomes, or comprehension. Included study sample sizes of participants/encounters ranged from 13 to 4,146. Use of professional interpretation, compared to ad hoc (family or friend as interpreter) or no interpretation, was consistently associated with better outcomes, generally approaching or equaling those of patients with language concordant physicians (both Spanish speakers with Spanish-speaking physicians, or English-speakers with English-speaking physicians). The review concluded that professional interpreter use was associated with decreased disparities in utilization and adherence to follow-up care, fewer interpretation errors and better patient diagnosis comprehension, better clinical outcomes (fewer obstetrical interventions, better hemoglobin A1C, lipid levels, creatinine levels), and greater patient satisfaction. Study populations and types of outcomes were too varied to permit meta-analysis.

Bagchi et al (2011) conducted a randomized controlled trial of 447 emergency department patients with limited English proficiency, of whom 242 were assigned to professional in-person interpretation and 205 were assigned to usual care (no interpretation, using untrained family members or friends, or telephone interpretation), based on randomized time blocks of interpreter availability. Assignment to professional in-person interpretation led to significantly greater degrees of patient-reported understanding and satisfaction with communication and to greater satisfaction with communication among ED physicians and nurses.

The two non-randomized, controlled trials (Hampers et al, 2002; Lee et al, 2002) compared outcomes for LEP patients or families who received professional interpretation to those who did not, and compared both groups to English-speakers. Hampers' study, conducted among 4146 children seen in an emergency department (of whom 550 had limited English proficient families) found that professional interpreter use was associated with decreased resource utilization and fewer hospitalizations compared to LEP families who did not receive professional interpretation. Lee's study of 223 English-speaking and 303 Spanish-speaking urgent care patients found that professional interpreter use, compared to none, was associated with increased satisfaction. Lindholm et al (2012) conducted a retrospective cohort study of 3,071 hospitalized patients with limited English proficiency, and found that use of professional interpretation at admission and/or discharge was associated with decreased length of hospital stay and lower risk of 30-day readmission.

FECC 16: Caregivers of CMC should report that the child's main provider created a shared care plan for their child

Seven randomized controlled trials, 3 non-randomized controlled trials, 6 uncontrolled interventions with a pre-post comparison, a non-systematic review including unpublished program evaluations, and a consensus statement from the AAP support that interventions that include a shared care plan are associated with improved outcomes among children and adults with chronic disease or medical complexity. Of note, most identified studies evaluated outcomes associated with shared care plan use in the context of larger care coordination or disease-specific management interventions; however, the shared care plan was generally a central feature of the multi-factorial intervention.

Specifically, Farmer, et al (2011; see Evidence Form for list of references) conducted a randomized controlled trial of an intervention that included 100 children with chronic illness on Medicaid (6-month intervention supporting 32 PCP offices). The intervention included a care coordinator who worked with the family to develop and implement a written health plan for the child to provide coordination with doctors and home visits/ telephone support. In between-group analyses, participants in the intervention group reported significantly lower needs for information as measured by the Family Needs Survey (FNS). In the within-subject analysis comparing pre- and post-intervention, there was a significant decrease in unmet needs as measured by the FNS. There was a significant improvement in satisfaction with specialty care and care coordination as measured by the Shared Responsibilities Tool Kit—Version 1.0. There was a significantly improved overall child health rating as measured by a five-point scale ranging from excellent to poor, and a trend toward improved child functional status as measured by the Functional Status II (Revised)—14 item version. There was a significant decrease in personal and family strain, as measured by the Impact on Family Scale (IFS).

Counsell et al (2007) conducted a cluster RCT of 951 low income seniors with chronic illness, in which the intervention group received two years of home-based care management centered around the development and implementation of an individualized shared care plan. At 24 months, the intervention group demonstrated better general health, vitality, social functioning, and mental health, as measured using the SF-36, along with significantly fewer ED visits overall, and fewer hospital admissions in a pre-defied group at high risk for admission.

Aiken et al (2006) conducted an RCT that enrolled 192 adults with COPD or CHF and estimated 2 year life expectancy. The intervention featured development of a care plan by a nurse case manager, supported by a multidisciplinary team, to provide in-home and telephone support and education to patients. The intervention led to better illness self-management, knowledge of illness-related resources, lower symptom-related distress, greater vitality, better physical functioning and higher self-rated health compared to controls.

In 2004, Lozano et al conducted a multisite cluster RCT with 678 children with mild to moderate persistent asthma (199 controls, 226 in a peer leader intervention, and 213 in a care planning intervention). In the care planning intervention, asthma nurses conducted an assessment, developed individualized shared care plans with the family, and provided self-management support and telephone follow-up. The intervention group had significantly fewer asthma symptom days and fewer oral steroid bursts per year, along with greater controller adherence, by parent report, compared with children receiving usual care.

The randomized controlled trials by Katon et al (2001; n=386), Unutzer et al (2002; n=1801), and Katon et al (2010; n=214) enrolled adults with depression (and co-morbid diabetes or congestive heart disease, for Katon 2010) in variations on an intervention centered on a shared care plan developed with the patient and supported by a multi-disciplinary care team. Intervention patients were found to have improvements in clinical measures of depression, adherence to therapy, quality of life, functional status, and management of co-morbid diseases.

The non-randomized controlled interventions provide additional support for the measure. In 2008, Dorr et al studied 3,432 chronically ill adults >64 years old. Intervention clinic patients were referred by their PCPs to the intervention, which consisted of a nurse care manager using structured protocols and individualized care plans. Intervention patients were age, sex, condition, and utilization-matched to 2 patients from control clinics who received usual care. The intervention group had significantly lower mortality overall and lower hospitalization rates among patients with diabetes. Adam et al (2010) conducted a non-randomized, controlled intervention of 20 adult outpatients with complex chronic illness considered “frequent attenders” at clinic, in which the multi-disciplinary care team developed and implemented an individualized care plan. The intervention was associated with improved satisfaction, patient well-being, and less frequent clinic visits. Rocco et al (2011) studied the impact of a plan of care intervention with 1110 adults with chronic disease. In the intervention, the PCP and patient developed individual problem lists, goals, and actions to be taken, within a medical home. Controls were drawn from a non-medical home model clinic without the plan of care tool. The intervention was associated with improvements in hemoglobin A1C, LDL cholesterol, and diastolic blood pressure.

The six uncontrolled intervention studies, with non-comparable or historical controls (Gordon et al, 2007; Farmer et al, 2005; Palfrey et al, 2004; Casey et al, 2011; Cady et al, 2009; Weiland et al, 2003), had sample sizes ranging from 22 to

227, and found associations between interventions featuring a shared care plan and (depending on the study) decreased hospitalizations, costs of care, unmet needs, work loss and school absences, and increased ED use, outpatient visits, and satisfaction with services.

Chen et al's non-systematic review (2000) included 29 care coordination programs for adults with chronic systemic disease, including some unpublished results. The review found that a written, goal-oriented, individualized care plan was a common element in cost-saving programs. All cost-saving programs had a care coordinator responsible for adjusting plans as needed. Programs that included typical components but had no measured impact (n=5) had less comprehensive, less specific, and less goal-oriented care plans, and/or inflexible reassessment schedules.

FECC 17: Caregivers of CMC should report that the child's main provider created an emergency care plan for their child

A consensus statement from the American Academy of Pediatrics supports the importance of having an emergency care plan for children with complex medical problems for optimizing outcomes.

The American Academy of Pediatrics recommends that children with special health care needs, especially those with complex conditions, have a written emergency care plan, developed with the primary care provider, detailing the child's condition(s), medication(s), and how best to manage the medical condition(s) in an urgent or emergent situation. Because these children have complex conditions and many have unique needs, they are at high risk for suboptimal outcomes in emergency situations.

Evidence Tables

The evidence for each of the 10 quality indicators being submitted here is presented below, in a series of 2 tables. The first table lists the quality measure name, the specific evidence, and the quality of the evidence using the 2011 Oxford Centre for Evidence Based Medicine grading scale (see below Table E1 for key).[1] The second table lists the specific studies and summarizes the findings. The full paper citations are presented at the end.

Table E1: FECC quality measures and their supporting evidence.

Number	Quality Measure	Quality of Evidence*	Supporting Literature
FECC 1	Caregivers of CMC should report that their child has a designated care coordinator.	2 3 4	Farmer et al., 2011[2] Wood et al, 2008[3] Gordon et al., 2007[4] Palfrey et al., 2004[5] Farmer et al., 2005[6] Cady et al, 2009[7] McAllister et al., 2009[8]

Number	Quality Measure	Quality of Evidence*	Supporting Literature
FECC 3	Caregivers of CMC who report having a designated care coordinator and who require community services should also report that their care coordinator helped their child to obtain needed community services in the last year.	2 4	Farmer et al. 2011[2] Gordon et al, 2007[4] Cady et al, 2009[7]
FECC 5	Caregivers of CMC who report having a care coordinator and who report that their care coordinator has contacted them in the last 3 months should also report that their care coordinator asked them about the following: a. caregiver concerns b. health changes of the child	2	Farmer et al. 2011[2]
FECC 7	Caregivers of CMC who report having a care coordinator for their child should also report that the care coordinator assists them with specialty service referrals by ensuring that the appointment with the specialty service provider occurs within 3 months of referral initiation	2 4	Farmer et al. 2011[2] Gordon et al, 2007[4] Palfrey et al, 2004[5] Cady et al, 2009[7]
FECC 8	Caregivers of CMC who report having a care coordinator should also report that their care coordinator: a. is knowledgeable about their child's health b. supports the caregiver c. advocates for the needs of their child	2 4	Farmer et al. 2011[2] Gordon et al, 2007[4] Palfrey et al, 2004[5] Farmer et al., 2005[6]
FECC 9	Caregivers/patients of CMC should report receiving a written visit summary following all outpatient visits in the last 12 months (or report access to a patient portal that provides a visit summary) and it should contain the following elements: a. current problem list b. current medication list c. drug allergies d. specialists involved in the child's care e. planned follow-up f. what to do for problems related to the outpatient visit	2 5	Palfrey et al, 2004[5] AAP 2005; Care Coordination in the Medical Home[9] National Committee for Quality Assurance (NCQA) 2011[10]
FECC 14	Caregivers of CMC should report that one of their child's health care providers (i.e., primary care	5	Savage, 2001[11]

Number	Quality Measure	Quality of Evidence*	Supporting Literature
	physician, specialist physician, care coordinator, NP, nurse, social worker, etc) communicated with school staff at least once a year about the educational impacts of the child's condition.		
FECC 15	Caregivers of CMC or CMC who self-identify as having a preference for conducting medical visits in a language other than English should have access to a professional medical interpreter (live or telephonic) at all visits for which an interpreter is needed.	2 3	Karliner et al, 2007[12] Bagchi et al, 2011[13] Linholm et al, 2012[14] Hampers et al, 2002[15] Lee et al, 2002[16]
FECC 16	Caregivers of CMC should report that the child's main provider created a shared care plan for their child	2 3 4 5	Counsell et al, 2007[17] Lozano et al, 2004[18] Unutzer et al, 2002[19] Katon et al, 2001[20] Katon et al, 2010[21] Aiken et al, 2006[22] Farmer et al, 2011,[23] Dorr et al, 2008[24] Adam et al, 2010[25] Rocco et al, 2011[26] Gordon et al, 2007[27] Farmer et al, 2005,[28] Palfrey et al, 2004[29] Casey et al, 2011[30] Cady et al, 2009[7] Chen et al, 2000,[31] Weiland et al, 2003[32] AAP 2002: The Medical Home[33] AAP 2005: Care Coordination in the Medical Home[9]

Number	Quality Measure	Quality of Evidence*	Supporting Literature
FECC 17	Caregivers of CMC should report that the child's main provider created an emergency care plan for their child	5	AAP 2010: Emergency Information Forms[34]

*Quality of Evidence Codes:

1: Systematic review
2: Randomized controlled trial (RCT)
3: Cohort studies
4: Case series, case-control, or historically-controlled studies
5: Consensus or mechanism-based reasoning

Level may be graded down on the basis of study quality, imprecision, indirectness, because of inconsistency between studies, or because the absolute effect size is very small; Level may be graded up if there is a large or very large effect size.

Table E2: Details and outcomes of the studies providing support to the FECC quality measures

Source, Study Design, and Population	Program	ED use	Hospitalization	Hospital bed days/ LOS	OP visits	Satisfaction[a]	Work Loss	School Absence	Quality of life	Cost of Care	Family Strain	Unmet Needs	Adherence	Clinical Measures[b]
Adam,[25] 2010 Controlled intervention, non-randomized 20 adult outpatients with chronic, complex illness (12 intervention)	Care team of 4 doctors, a psychologist, a pharmacist and a nurse develop a tentative individualized plan; patient feedback is incorporated, and then the plan is implemented.	↑	nc		↓									↑

Source, Study Design, and Population	Program	ED use	Hospitalization	Hospital bed days/ LOS	OP visits	Satisfaction[a]	Work Loss	School Absence	Quality of life	Cost of Care	Family Strain	Unmet Needs	Adherence	Clinical Measures[b]
Aiken,[22] 2006 RCT 192 adults with COPD or CHF and estimated 2 year life expectancy (101 intervention)	A nurse case manager, supported by a medical director, social worker, and pastor, provided in-home and telephone support, education, and care plan development to patients. Care plan was shared with the PCP and other providers.	nc							↑					
Bagchi et al, 2011[13] RCT 447 emergency department patients with limited English proficiency (242 intervention)	Patients received professional in-person interpretation or usual care (no interpretation or untrained family memebers or friends) based on randomized time blocks of availability					↑								
Cady,[7] 2009 Uncontrolled intervention 43 children with complex chronic disease	Nurse practitioners provided phone-based care coordination between the family, PCP, and specialists, and helped develop a care plan for recurrent acute illnesses (intervention details from Kelly et al.)[35]		↓											
Casey,[30] 2011 Uncontrolled intervention 225 children with complex chronic disease	Multidisciplinary clinic (MD, RN, nutrition, social work) worked with the family to develop an Individual Health Plan; also provided care coordination.		↓		↑					↓				
Chen,[31] 2000 Non-systematic review 29 care coordination	Reviewed care coordination programs associated with decreased hospitalizations or health care expenditures; also reviewed selected programs with no demonstrated impact		↓							↓				

Source, Study Design, and Population	Program	ED use	Hospitalization	Hospital bed days/ LOS	OP visits	Satisfaction[a]	Work Loss	School Absence	Quality of life	Cost of Care	Family Strain	Unmet Needs	Adherence	Clinical Measures[b]
programs for adults with chronic systemic disease	on cost or hospitalizations for comparison. - A written, goal-oriented, individualized care plan was a common element in cost-saving programs. All had a care coordinator responsible for adjusting plans as needed. - Programs that included typical components but had no measured impact (n=5) had less comprehensive, less specific, and less goal-oriented care plans. Two also had inflexible reassessment schedules.													
Counsell,[17] 2007 Cluster RCT 951 low income seniors with chronic illness (474 intervention)	Two years of home-based care management by a nurse practitioner and social worker, collaborating with the PCP and interdisciplinary team to develop and implement an individualized care plan.	↓	↓						↑					
Dorr,[24] 2008 Controlled intervention, non-randomized 3432 chronically ill adults >64 years (1144 in the intervention)	Intervention clinic patients were referred by their PCPs to the intervention: a nurse care manager using structured protocols and individualized care plans. Intervention patients were age-, sex-, condition-, and utilization-matched to 2 patients from control clinics who received usual care.	↑	↓											
Farmer,[28] 2005 Uncontrolled intervention 51 children with complex chronic disease	Nurse practitioner-led care coordination involving a home visit, assessment, referral to resources, and an individualized written plan with specific goals. The NP served as a consultant to the PCPs.				↓	↑	↓	↓				↓		

Source, Study Design, and Population	Program	ED use	Hospitalization	Hospital bed days/ LOS	OP visits	Satisfaction[a]	Work Loss	School Absence	Quality of life	Cost of Care	Family Strain	Unmet Needs	Adherence	Clinical Measures[b]
Farmer,[23] 2011 RCT with crossover to intervention 100 children with chronic illness on Medicaid (50 intervention)	6-month intervention supporting 32 PCP offices. Care coordinator worked with the family to develop a written health plan for the child, provide access to services, coordination with doctors and home visit/ telephone support.					↑	nc	nc			↓	↓		
Gordon,[27] 2007 Uncontrolled intervention 227 children with medical complexity	Depending on complexity, patients were assigned to an NP only or NP and MD, who developed a care plan with the family, interfaced with the PCP and other services, and provided support.	↑	↓	↓	↑					↓				
Hampers et al, 2002[15] Controlled intervention, non-randomized 4146 children seen in an emergency department, of whom 550 had limited English proficient families	Limited English proficient families received care from a bilingual physician, through a professional interpreter, or without the aid of either, depending on availability; resource utilization was also assessed for English proficient families. Results shown are for professional interpretation compared to interpreter needed but not available		↓							↓				
Karliner et al, 2007[12] Systematic review Review included 1 RCT and 27 cohort studies comparing professional interpreter use to another group from 1966-2005, and assessing	Use of professional interpretation, compared to ad hoc or no interpretation, was consistently associated with better outcomes, generally approaching or equaling those of patients with language concordant physicians. Study populations and types of outcomes were too varied to permit meta-analysis.		↓			↑				↓			↑	↑

Source, Study Design, and Population	Program	ED use	Hospitalization	Hospital bed days/ LOS	OP visits	Satisfaction[a]	Work Loss	School Absence	Quality of life	Cost of Care	Family Strain	Unmet Needs	Adherence	Clinical Measures[b]
satisfaction, utilization, clinical outcomes, or comprension														
Katon,[20] 2001 RCT 386 adults with major depression (194 intervention)	Intervention included 2 visits with a depression specialist in which a written personal relapse prevention plan was devised and then shared with the PCP, 3 follow up phone calls, and medication refill monitoring.												↑	↑/ nc
Katon,[21] 2010 RCT 214 adults with poorly controlled diabetes mellitus, congestive heart disease, or both, and depression (106 intervention)	12 month intervention in which a nurse care coordinator, supervised by a psychiatrist, the PCP, and a psychologist, worked with patients to develop and implement an individualized treatment plan.					↑			↑					↑
Lee et al, 2002[16] Controlled intervention, non-randomized 223 English-speaking and 303 Spanish speaking urgent care patients	Spanish-speaking patients received care from bilingual physicians, via professional telephone interpreter, or through untrained interpreters based on availability and preference; English-speakers enrolled as additional comparison. Results presented reflect use of professional interpreter compared to untrained interpreter.					↑								

Source, Study Design, and Population	Program	ED use	Hospitalization	Hospital bed days/ LOS	OP visits	Satisfaction[a]	Work Loss	School Absence	Quality of life	Cost of Care	Family Strain	Unmet Needs	Adherence	Clinical Measures[b]
Linholm et al, 2012[14] Restrospective cohort study 3071 hospitalized patients with limited English proficiency	Among hospitalized patients with limited English proficiency, use of professional interpretation at admission and/or discharge was assessed, and association between professional interpreter use and length of stay determined			↓										
Lozano,[18] 2004 Multisite cluster RCT 678 children with mild to moderate persistent asthma (199 control, 226 in a peer leader intervention, and 213 in a care planning intervention)	Asthma nurses conducted assessment, developed individualized care plan with family, provided self-management support and phone follow-up. There was also an MD peer leader to champion office-wide change.													↑
McAlister, 2009[8] Uncontrolled intervention 82 children, in 10 practices, with special health care needs	Ten practices participated in a 3-year medical home improvement process that involved engaging families and providing care coordination to children with special health care needs		↓	↓	↓			↓			↓			
Palfrey,[29] 2004 Uncontrolled intervention 117 children with complex and/or chronic medical conditions in 6 practices	A nurse practitioner as care coordinator within a medical home provided home visits (including sick visits), family support, and care coordination, and worked with the family to develop a written care plan.	nc	↓			↑	↓	nc						

Source, Study Design, and Population	Program	ED use	Hospitalization	Hospital bed days/ LOS	OP visits	Satisfaction[a]	Work Loss	School Absence	Quality of life	Cost of Care	Family Strain	Unmet Needs	Adherence	Clinical Measures[b]
Rocco,[26] 2011 Controlled retrospective cohort 1110 adults with chronic disease (593 intervention)	Plan of care intervention: PCP and patient develop individual problem list, goals, and actions to be taken, within a medical home. Controls were drawn from a non-medical home model clinic without the plan of care tool.													↑
Unutzer,[19] 2002 RCT 1801 adults >59 years old with major depression or dysthymic disorder (906 intervention)	Intervention included 12 months of depression care management by a care manager, under the supervision of PCP and a psychiatrist, beginning with development of an individualized care plan guided by algorithms.					↑			↑				↑	↑
Weiland,[32] 2003 Intervention with non-comparable controls 22 adolescents with cystic fibrosis (17 intervention)	The intervention consisted of an individualized daily inpatient schedule that the adolescent developed with care team. The control group was made up of patients who declined to participate in the intervention.					↑								
Wood et al, 2008[3] Controlled intervention, non-randomized 144 children Children enrolled in Title V	Compared nurse led practice-based care coordination (intervention) in 3 practices with pre-existing agency-based care coordination (control) in 3 practices					↑						↓		

ED = Emergency department

OP = Outpatient

ED = Emergency department
LOS=Length of stay
↑ = increase in any outcome measure within column domain

↑ = significant increase in any outcome measure within column domain

↓ = decrease in any outcome measure within column domain

↓ = significant decrease in any outcome measure within column domain

nc = no change

[a]Multiple different measures of satisfaction were used within and between studies. A positive indicator in this column reflects improvement in any measure of satisfaction.

[b]Examples of clinical measures include depressive symptoms, asthma symptom days, hemoglobin A1C levels, and LDL cholesterol levels

[c]This category includes self- or parent-reported overall health status

References:

1. The Oxford 2011 Levels of Evidence. 2011. (Accessed September 27, 2015, at http://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf.)
2. Farmer J, Clark M, Drewel E, Swenson T, Ge B. Consultative Care Coordination Through the Medical Home for CSHCN: A Randomized Controlled Trial. Maternal and Child Health Journal 2011;15:1110-8.
3. Wood D, Winterbauer N, Sloyer P, et al. A longitudinal study of a pediatric practice-based versus an agency-based model of care coordination for children and youth with special health care needs. Maternal & Child Health Journal 2009;13:667-76.
4. Gordon JB, Colby HH, Bartelt T, Jablonski D, Krauthoefer ML, Havens P. A tertiary care-primary care partnership model for medically complex and fragile children and youth with special health care needs. Archives of Pediatrics & Adolescent Medicine 2007;161:937-44.
5. Palfrey JS, Sofis LA, Davidson EJ, et al. The Pediatric Alliance for Coordinated Care: evaluation of a medical home model. Pediatrics 2004;113:1507-16.
6. Farmer JE, Clark MJ, Sherman A, Marien WE, Selva TJ. Comprehensive Primary Care for Children With Special Health Care Needs in Rural Areas. Pediatrics 2005;116:649-56.
7. Cady R, Finkelstein S, Kelly A. A telehealth nursing intervention reduces hospitalizations in children with complex health conditions. J Telemed Telecare 2009;15:317-20.
8. McAllister JWB, MS, MHA; Sherrieb, Kathleen MS, DrPH; Cooley, W. Carl MD. Improvement in the Family-Centered Medical Home Enhances Outcomes for Children and Youth With Special Healthcare Needs.

. Journal of Ambulatory Care Management 2009;32:9.
9. AAP. Care coordination in the medical home: integrating health and related systems of care for children with special health care needs. Pediatrics 2005;116:1238-44.
10. National Committee for Quality Assurance. NCQA's Patient-Centered Medical Home (PCMH) 20112011.
11. Savage RC, Pearson S, McDonald H, Potoczny-Gray A, Marchese N. After hospital: working with schools and families to support the long term needs of children with brain injuries. NeuroRehabilitation 2001;16:49-58.
12. Karliner LS, Jacobs EA, Chen AH, Mutha S. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. Health Serv Res 2007;42:727-54.
13. Bagchi AD, Dale S, Verbitsky-Savitz N, Andrecheck S, Zavotsky K, Eisenstein R. Examining effectiveness of medical interpreters in emergency departments for Spanish-speaking patients with limited English proficiency: results of a randomized controlled trial. Ann Emerg Med 2011;57:248-56 e1-4.
14. Lindholm M, Hargraves JL, Ferguson WJ, Reed G. Professional Language Interpretation and Inpatient Length of Stay and Readmission Rates. J Gen Intern Med 2012.

15. Hampers LC, McNulty JE. Professional interpreters and bilingual physicians in a pediatric emergency department: effect on resource utilization. Arch Pediatr Adolesc Med 2002;156:1108-13.
16. Lee LJ, Batal HA, Maselli JH, Kutner JS. Effect of Spanish interpretation method on patient satisfaction in an urban walk-in clinic. J Gen Intern Med 2002;17:641-5.
17. Counsell SR, Callahan CM, Clark DO, et al. Geriatric care management for low-income seniors: a randomized controlled trial. JAMA 2007;298:2623-33.
18. Lozano P, Finkelstein JA, Carey VJ, et al. A multisite randomized trial of the effects of physician education and organizational change in chronic-asthma care: health outcomes of the Pediatric Asthma Care Patient Outcomes Research Team II Study. Arch Pediatr Adolesc Med 2004;158:875-83.
19. Unutzer J, Katon W, Callahan CM, et al. Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. JAMA 2002;288:2836-45.
20. Katon W, Rutter C, Ludman EJ, et al. A randomized trial of relapse prevention of depression in primary care. Arch Gen Psychiatry 2001;58:241-7.
21. Katon WJ, Lin EHB, Von Korff M, et al. Collaborative care for patients with depression and chronic illnesses. The New England Journal of Medicine 2010;363:2611-20.
22. Aiken LS, Butner J, Lockhart CA, Volk-Craft BE, Hamilton G, Williams FG. Outcome evaluation of a randomized trial of the PhoenixCare intervention: program of case management and coordinated care for the seriously chronically ill. Journal of Palliative Medicine 2006;9:111-26.
23. Farmer JE, Clark MJ, Drewel EH, Swenson TM, Ge B. Consultative care coordination through the medical home for CSHCN: a randomized controlled trial. Matern Child Health J 2011;15:1110-8.
24. Dorr DA, Wilcox AB, Brunker CP, Burdon RE, Donnelly SM. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. J Am Geriatr Soc 2008;56:2195-202.
25. Adam P, Brandenburg DL, Bremer KL, Nordstrom DL. Effects of team care of frequent attenders on patients and physicians. Fam Syst Health 2010;28:247-57.
26. Rocco N, Scher K, Basberg B, Yalamanchi S, Baker-Genaw K. Patient-centered plan-of-care tool for improving clinical outcomes. Qual Manag Health Care 2011;20:89-97.
27. Gordon JB, Colby HH, Bartelt T, Jablonski D, Krauthoefer ML, Havens P. A tertiary care-primary care partnership model for medically complex and fragile children and youth with special health care needs. Arch Pediatr Adolesc Med 2007;161:937-44.
28. Farmer JE, Clark MJ, Sherman A, Marien WE, Selva TJ. Comprehensive primary care for children with special health care needs in rural areas. Pediatrics 2005;116:649-56.
29. Palfrey JS, Sofis LA, Davidson EJ, Liu J, Freeman L, Ganz ML. The Pediatric Alliance for Coordinated Care: evaluation of a medical home model. Pediatrics 2004;113:1507-16.
30. Casey PH, Lyle RE, Bird TM, et al. Effect of hospital-based comprehensive care clinic on health costs for Medicaid-insured medically complex children. Arch Pediatr Adolesc Med 2011;165:392-8.
31. Chen A, Brown R, Archibald N, Aliotta S, Fox PD. Best Practices in Care Coordination. Baltimore, MD: Health Care Financing Administration, Division of Demonstration Programs, Center for Health Plans and Providers 2000.
32. Weiland J, Schoettker PJ, Byczkowski T, Britto MT, Pandzik G, Kotagal UR. Individualized daily schedules for hospitalized adolescents with cystic fibrosis. J Pediatr Health Care 2003;17:284-9.
33. American Academy of Pediatrics. The medical home. Pediatrics 2002;110:184-6.
34. American Academy of Pediatrics. Policy statement--emergency information forms and emergency preparedness for children with special health care needs. Pediatrics 2010;125:829-37.
35. Kelly AM, Kratz B, Bielski M, Rinehart PM. Implementing transitions for youth with complex chronic conditions using the medical home model. Pediatrics 2002;110:1322-7.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_Evidence_FECC_submit.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Increasing numbers of children in the United States are living with medical complexity.(2) Although these children with medical complexity (CMC) comprise only 13% of the pediatric population, they account for a disproportionately high 26-49% of hospital days(3,4) and 70% of overall health expenditures.(5) Given the cost and complexity of caring for these children, optimizing the quality of their care is likely to yield significant health and economic benefits.

Comprehensive, well-coordinated care in a medical home improves patient and family experiences of care6-8 and patient medical outcomes.(6,7,9,10) Care coordination interventions among CMC have also been associated with decreased unmet specialty care need11 and improved utilization of health care services, decreasing hospitalizations and cost.8,9,12-14 Improving care coordination for CMC is likely to improve many aspects of care received by these children and their families.

Little is known about the quality of care coordination received by CMC. Present assessments of care coordination are generally limited to whether care coordination was received or not, without any attempt to identify potentially beneficial components of care coordination or the manner in which they were delivered. The evidence that is available suggests that 29-41% of parents of children with special health care needs report not getting needed help with care coordination;(15,16) little is known about the quality of the help that is being received.

While limited information on quality of care coordination exists, data do demonstrate disparities in receipt of care coordination. Latino and black children have been found to be more likely to have unmet care coordination needs compared to non-Hispanic white children.(16) In addition, children from families with limited English proficiency have reported higher unmet care coordination needs and greater difficulty getting needed referrals compared to English proficient families.(15) These data suggest that there may also be disparities in quality of care coordination received by race/ethnicity and language. The FECC Survey can be collected with data on child and parent race, ethnicity and language, which will allow for tracking of disparities in care coordination quality over time.

references:

2. Bethell CD, Read D, Blumberg SJ, Newacheck PW. What is the prevalence of children with special health care needs? Toward an understanding of variations in findings and methods across three national surveys. Matern Child Health J. 2008;12(1):1-14.
3. Berry JG, Hall M, Hall DE, et al. Inpatient growth and resource use in 28 children´s hospitals: a longitudinal, multi-institutional study. JAMA Pediatr. 2013;167(2):170-177.
4. Simon TD, Berry J, Feudtner C, et al. Children with complex chronic conditions in inpatient hospital settings in the United States. Pediatrics. 2010;126(4):647-655.
5. Ireys HT, Anderson GF, Shaffer TJ, Neff JM. Expenditures for care of children with chronic illnesses enrolled in the Washington State Medicaid program, fiscal year 1993. Pediatrics. 1997;100(2 Pt 1):197-204.
6. Farmer JE, Clark MJ, Sherman A, Marien WE, Selva TJ. Comprehensive primary care for children with special health care needs in rural areas. Pediatrics. 2005;116(3):649-656.
7. Farmer JE, Clark MJ, Drewel EH, Swenson TM, Ge B. Consultative care coordination through the medical home for CSHCN: a randomized controlled trial. Matern Child Health J. 2011;15(7):1110-1118.
8. Palfrey JS, Sofis LA, Davidson EJ, Liu J, Freeman L, Ganz ML. The Pediatric Alliance for Coordinated Care: evaluation of a medical home model. Pediatrics. 2004;113(5 Suppl):1507-1516.
9. Counsell SR, Callahan CM, Clark DO, et al. Geriatric care management for low-income seniors: a randomized controlled trial. JAMA. 2007;298(22):2623-2633.

10. Rocco N, Scher K, Basberg B, Yalamanchi S, Baker-Genaw K. Patient-centered plan-of-care tool for improving clinical outcomes. Qual Manag Health Care. 2011;20(2):89-97.
11. Boudreau AA, Perrin JM, Goodman E, Kurowski D, Cooley WC, Kuhlthau K. Care coordination and unmet specialty care among children with special health care needs. Pediatrics. 2014;133(6):1046-1053.
12. Casey PH, Lyle RE, Bird TM, et al. Effect of hospital-based comprehensive care clinic on health costs for Medicaid-insured medically complex children. Arch Pediatr Adolesc Med. 2011;165(5):392-398.
13. Dorr DA, Wilcox AB, Brunker CP, Burdon RE, Donnelly SM. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. J Am Geriatr Soc. 2008;56(12):2195-2202.
14. Gordon JB, Colby HH, Bartelt T, Jablonski D, Krauthoefer ML, Havens P. A tertiary care-primary care partnership model for medically complex and fragile children and youth with special health care needs. Arch Pediatr Adolesc Med. 2007;161(10):937-944.
15. Zickafoose JS, Davis MM. Medical home disparities are not created equal: differences in the medical home for children from different vulnerable groups. J Health Care Poor Underserved. 2013;24(3):1331-1343.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

The following performance scores were derived from our field-testing of the FECC measure set, in which we sampled 1500 caregivers from each of 2 states who had a child with medical complexity covered by Medicaid. We administered the survey from July to November 2013 via both mixed mode (mail with phone follow-up) and phone only; the survey was available in English and Spanish. We obtained 600 completed surveys in Washington and 609 in Minnesota. Of caregiver respondents, 95% were female, 60% were non-Hispanic white, and 80% were English proficient (defined as speaking English "very well"). Thirty-six percent had completed high school as their highest education, and 53% had completed college. Please see Table T1, in section 1.6 of the testing attachment, for compete demographic characteristics of our sample.

Measure: FECC 1
Description: Has a care coordinator
Respondents (N): 841
Mean (SD): 72.5 (44.7)
Min: 0
25th percentile: 0
Median: 100
75th percentile: 100
Max: 100

Measure: FECC 3
Description: Care coordinator helped to obtain community services
Respondents (N): 279
Mean (SD): 50.5 (50.1)
Min: 0
25th percentile: 0
Median: 100
75th percentile: 100
Max: 100

Measure: FECC 5
Description: Care coordinator asked about concerns and health changes
Respondents (N): 267
Mean (SD): 81.0 (25.7)
Min: 0
25th percentile: 66.7
Median: 100
75th percentile: 100
Max: 100

Measure: FECC 7

Description: Care coordinator assisted with specialist service referrals
Respondents (N): 455
Mean (SD): 73.2 (44.4)
Min: 0
25th percentile: 0
Median: 100
75th percentile: 100
Max: 100

Measure: FECC 8
Description: Care coordinator was knowledgeable, supportive & advocated for child's needs
Respondents (N): 558
Mean (SD): 84.3 (17.9)
Min: 8.3
25th percentile: 75
Median: 83.3
75th percentile: 100
Max: 100

Measure: FECC 9
Description: Appropriate written visit summary content
Respondents (N): 709
Mean (SD): 81.1 (20.5)
Min: 0
25th percentile: 70.8
Median: 83.3
75th percentile: 100
Max: 100

Measure: FECC 14
Description: Health care provider communicated with school staff about child's condition
Respondents (N): 657
Mean (SD): 28.5 (45.1)
Min: 0
25th percentile: 0
Median: 0
75th percentile: 100
Max: 100

Measure: FECC 15
Description: Caregiver has access to medical interpreter when needed
Respondents (N): 117
Mean (SD): 83.5 (23.0)
Min: 0
25th percentile: 66.7
Median: 100
75th percentile: 100
Max: 100

Measure: FECC 16
Description: Child has shared care plan
Respondents (N): 1095
Mean (SD): 43.7 (49.6)
Min: 0
25th percentile: 0
Median: 0

75th percentile: 100
Max: 100

Measure: FECC 17
Description: Child has emergency care plan
Respondents (N): 1138
Mean (SD): 20.3 (40.2)
Min: 0
25th percentile: 0
Median: 0
75th percentile: 0
Max: 100

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
not applicable

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Disparities data are derived from the same field-testing data described above in section 1b.2, and in greater detail in section 2b2.2 of the testing attachment. Child and caregiver race/ethnicity, and caregiver languages are given below (see section 1.6 of the testing attachment for full demographic details of the respondents):

Caregiver race/ethnicity:
Non-Hispanic white: 722 (60%)
Hispanic: 250 (21%)
African American: 92 (8%)
Other: 119 (10%)
Missing: 26 (2%)

Caregiver English proficiency
Speaks very well: 972 (80%)
Speaks well: 78 (6%)
Does not speak well: 82 (7%)
Does not speak at all: 52 (4%)
Not answered: 25 (2%)

Language of survey completion
English: 1048 (87%)
Spanish: 161 (13%)

Child race/ethnicity
Non-Hispanic white: 585 (48%)
Hispanic: 308 (26%)
African American: 94 (8%)
Other: 195 (22%)
Missing: 27 (2%)

We evaluated differences in FECC quality measure scores by child race/ethnicity. Child race/ethnicity was categorized on the basis of 2 questions: “Is your child of Hispanic or Latino origin or descent?” and “What is your child’s race? Please choose one or more from this list: White; Black or African American; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native; Other.” Based on these questions, children were categorized as one of the following: non-Hispanic white, Hispanic, black or other (including multiple races). Individual groups within the “other” category were too small to evaluate separately.

FECC quality measure scores were evaluated by race/ethnicity, both in unadjusted and adjusted analyses. In unadjusted analyses, some variability was seen on the basis of race/ethnicity. Compared to caregivers of non-Hispanic white children, caregivers of black children reported better scores on 4 FECC measures, and caregivers of Hispanic children reported better scores on 3 measures and worse scores on 2 measures.

In analyses adjusting for caregiver education and assigned survey mode, the results for black children remained unchanged, with better scores for 4 FECC measures. However, many of the findings for Hispanic children were no longer statistically significant after adjustment, with only 1 measure still showing a positive difference, and 1 measure showing a negative difference. Black and Hispanic children remained significantly more likely than non-Hispanic white children to have a shared care plan.

For full results stratified by race/ethnicity, please see Tables T10 (unadjusted) and T11 (adjusted) in the testing attachment (section 2b5.2). Results in which differences were found in adjusted analyses are presented here:

Measure: FECC 8
Description: Care coordinator was knowledgeable, supportive and advocated for child's needs
Non-Hispanic white (n=585): 85.6
Hispanic (n=308): 81.3*
Black (n=94): 81.1
Other (n=222): 86.3

Measure: FECC 9
Description: Appropriate written visit summary content
Non-Hispanic white (n=585): 81.0
Hispanic (n=308): 80.8
Black (n=94): 86.5*
Other (n=222): 78.2

Measure: FECC 14
Description: Health care provider communicated with school staff about child's condition
Non-Hispanic white (n=585): 25.6
Hispanic (n=308): 28.2
Black (n=94): 39.6*
Other (n=222): 32.8

Measure: FECC 16
Description: Child has shared care plan
Non-Hispanic white (n=585): 38.3
Hispanic (n=308): 47.4*
Black (n=94): 65.5***
Other (n=222): 41.7

Measure: FECC 17
Description: Child has emergency care plan
Non-Hispanic white (n=585): 17.1
Hispanic (n=308): 21.9
Black (n=94): 43.8***
Other (n=222): 14.3

Compared to white reference group using linear or logistic regression: *p<0.05 **p<0.01 ***p<0.001

We also evaluated the FECC quality measures by caregiver English proficiency. We identified the LEP population using the methodology described by Karliner et al (2008), based on a combination of the US Census question regarding self-reported English proficiency (How well do you Speak English? Very well; Well; Not well; Not at all) and preferred language for health care

conversations. Those who report speaking English very well are considered English proficient. Those who report speaking English not well or not at all are considered LEP. Those who report speaking English well are classified as English proficient if their preferred language for medical care is English, and LEP if it is another language. Compared to using the US Census question alone, this methodology better identifies families who are likely to benefit from interpretation in the medical setting, or conversely those most likely to suffer harm from lack of professional interpretation. The vast majority (147 out of 154) of the LEP respondents were Spanish-speaking, as the FECC Survey was available during field-testing in English and Spanish only.

Unadjusted analyses and analyses adjusting for caregiver education and survey mode were generally similar (see Tables T12 and T13 in section 2b5.2 of the testing attachment). While LEP was positively associated with having a shared care plan, having a care coordinator and receiving help with access-related aspects of care coordination, it was negatively associated with communication-related care coordinator attributes. Results for which a disparity was found in adjusted analyses are presented here:

Measure: FECC 5
Description: Care coordinator asked about concerns and health changes
English proficient (n=1094): 82.9
Limited English proficient (n=154): 69.2*

Measure: FECC 8
Description: Care coordinator was knowledgeable, supportive and advocated for child's needs
English proficient (n=1094): 85.3
Limited English proficient (n=154): 77.5**

Measure: FECC 9
Description: Appropriate written visit summary content
English proficient (n=1094): 81.8
Limited English proficient (n=154): 75.6*

Measure: FECC 14
Description: Health care provider communicated with school staff about child's condition
English proficient (n=1094): 26.4
Limited English proficient (n=154): 48.6**

Measure: FECC 16
Description: Child has shared care plan
English proficient (n=1094): 41.7
Limited English proficient (n=154): 55.1*

Compared to English proficient reference group using linear or logistic regression: *p<0.05 **p<0.01 ***p<0.001

Because the field test was restricted to children receiving Medicaid, there was limited variability in socioeconomic status. We are therefore unable to comment on the FECC quality measures' ability to identify disparities based on socioeconomic status. The quality measures in the FECC measure set apply exclusively to children with medical complexity, and so are not intended to identify disparities between those who do and do not have special health care needs.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
not applicable

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Children with medical complexity (CMC) comprise only 13% of the pediatric population, yet they account for a disproportionately high 26-49% of hospital days(3,4) and 70% of overall health expenditures.(5)

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

2. Bethell CD, Read D, Blumberg SJ, Newacheck PW. What is the prevalence of children with special health care needs? Toward an understanding of variations in findings and methods across three national surveys. Matern Child Health J. 2008;12(1):1-14.
3. Berry JG, Hall M, Hall DE, et al. Inpatient growth and resource use in 28 children´s hospitals: a longitudinal, multi-institutional study. JAMA Pediatr. 2013;167(2):170-177.
4. Simon TD, Berry J, Feudtner C, et al. Children with complex chronic conditions in inpatient hospital settings in the United States. Pediatrics. 2010;126(4):647-655.
5. Ireys HT, Anderson GF, Shaffer TJ, Neff JM. Expenditures for care of children with chronic illnesses enrolled in the Washington State Medicaid program, fiscal year 1993. Pediatrics. 1997;100(2 Pt 1):197-204.
6. Farmer JE, Clark MJ, Sherman A, Marien WE, Selva TJ. Comprehensive primary care for children with special health care needs in rural areas. Pediatrics. 2005;116(3):649-656.
7. Farmer JE, Clark MJ, Drewel EH, Swenson TM, Ge B. Consultative care coordination through the medical home for CSHCN: a randomized controlled trial. Matern Child Health J. 2011;15(7):1110-1118.
8. Palfrey JS, Sofis LA, Davidson EJ, Liu J, Freeman L, Ganz ML. The Pediatric Alliance for Coordinated Care: evaluation of a medical home model. Pediatrics. 2004;113(5 Suppl):1507-1516.
9. Counsell SR, Callahan CM, Clark DO, et al. Geriatric care management for low-income seniors: a randomized controlled trial. JAMA. 2007;298(22):2623-2633.
10. Rocco N, Scher K, Basberg B, Yalamanchi S, Baker-Genaw K. Patient-centered plan-of-care tool for improving clinical outcomes. Qual Manag Health Care. 2011;20(2):89-97.
11. Boudreau AA, Perrin JM, Goodman E, Kurowski D, Cooley WC, Kuhlthau K. Care coordination and unmet specialty care among children with special health care needs. Pediatrics. 2014;133(6):1046-1053.
12. Casey PH, Lyle RE, Bird TM, et al. Effect of hospital-based comprehensive care clinic on health costs for Medicaid-insured medically complex children. Arch Pediatr Adolesc Med. 2011;165(5):392-398.
13. Dorr DA, Wilcox AB, Brunker CP, Burdon RE, Donnelly SM. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. J Am Geriatr Soc. 2008;56(12):2195-2202.
14. Gordon JB, Colby HH, Bartelt T, Jablonski D, Krauthoefer ML, Havens P. A tertiary care-primary care partnership model for medically complex and fragile children and youth with special health care needs. Arch Pediatr Adolesc Med. 2007;161(10):937-944.
15. Zickafoose JS, Davis MM. Medical home disparities are not created equal: differences in the medical home for children from different vulnerable groups. J Health Care Poor Underserved. 2013;24(3):1331-1343.
16. Toomey SL, Chien AT, Elliott MN, Ratner J, Schuster MA. Disparities in unmet need for care coordination: the national survey of children´s health. Pediatrics. 2013;131(2):217-224.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.seattlechildrens.org/research/child-health-behavior-and-development/mangione-smith-lab/measurement-tools/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_detailed_specs_FECC_092915_submit.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
not applicable

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerators for each of the 10 FECC quality measures included within the FECC measures set are specified in the Detailed Measure Specifications (see S.2b). A brief description of each numerator is laid out in Table 1 in section De.3, and a more detailed description follows:

FECC-1: Caregivers of CMC should report that their child has a designated care coordinator.

FECC-3: Caregivers of CMC who report having a designated care coordinator and who require community services should also report that their care coordinator helped their child to obtain needed community services in the last year.

FECC-5:Caregivers of CMC who report having a care coordinator and who report that their care coordinator has contacted them in the last 3 months should also report that their care coordinator asked them about the following:
- Caregiver concerns
- Health changes of the child

FECC-7: Caregivers of CMC who report having a care coordinator for their child should also report that the care coordinator assists them with specialty service referrals by ensuring that the appointment with the specialty service provider occurs

FECC-8: Caregivers of CMC who report having a care coordinator should also report that their care coordinator:
- Was knowledgeable about their child's health
- Supported the caregiver
- Advocated for the needs of the child

FECC-9: Caregivers of CMC who report receiving a written visit summary during the last 12 months from their child's main provider's office should report that it contained the following elements:
- Current problem list
- Current medication list
- Drug allergies
- Specialists involved in the child's care
- Planned follow-up
- What to do for problems related to outpatient visit

FECC-14: Caregivers of CMC who report their child's condition causes difficulty learning, understanding, or paying attention in class should also report that one of their child's health care providers (i.e., primary care physician, specialist physician, care coordinator, nurse practitioner, nurse, social worker, etc.) communicated with school staff at least once a year about the educational impacts of the child's condition.

FECC-15: Caregivers of CMC who self-identify as having a preference for conducting medical visits in a language other than English should have access to a professional medical interpreter (live or telephonic) at all visits for which an interpreter is needed.

FECC-16: Caregivers of CMC should report that their child's primary care provider created a shared care plan for their child.

FECC-17: Caregivers of CMC should report that their child's main provider created an emergency care plan for their child.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The recall period for the majority of these caregiver-reported measures is 12 months, although measure FECC 5 specifies a 3-month recall period. Calculation of the denominator uses the PMCA, which uses up to 3 years' worth of retrospective ICD-9 codes to identify children with complex, chronic disease.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerators for each of the care coordination quality measures included within the FECC measure set are specified in the Detailed Measure Specifications (S.2b).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The eligible population of caregivers for the FECC Survey overall is composed of those who meet the following criteria:

1. Parents or legal guardians of children 0-17 years of age
2. Child classified as having a complex, chronic condition using the Pediatric Medical Complexity Algorithm (PMCA) (see Simon TD, Cawthon ML et al. 2014)
3. Child had at least 4 visits to a healthcare provider over the previous year

While some of the FECC measures only apply to a subset of the overall eligible population for the survey (e.g., measures related to the quality of care coordination services provided are only scored for those caregivers who endorse having a care coordinator), eligibility for these quality measures can only be gleaned from responses to the FECC Survey itself. This is analogous to the situation with many H-CAHPS measures, where, for example, measures about blood draws and laboratory testing are scored only for those who had the relevant service performed during the time frame or hospitalization in question.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk : Individuals with multiple chronic conditions

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The details for denominator identification are provided in S.2b, including the ICD-9 codes used for determining the PMCA. The PMCA SAS programming code is available at:
http://www.seattlechildrens.org/research/child-health-behavior-and-development/mangione-smith-lab/measurement-tools/

The process of converting the ICD-9 codes to ICD-10 codes for calculating the PMCA is underway, and should be complete and available within 6-9 months. However, because the PMCA uses up to 3 years' worth of retrospective administrative data, the ICD-10 code version is not expected to be needed for widespread use immediately.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

Denominator exclusions:
1. Child had died
2. Caregiver spoke a language other than English or Spanish

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Please see S2.b.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Please see the response to S.14, below, for details about producing a clinically-adjusted model that could be stratified by caregiver education (the sociodemographic factor we recommend adjustment for). The specifications for those models are also included in S.2b.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Other
If other: case mix adjustment

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Case-mix adjustment is completed via linear regression for continuous measures and logistic regression for binary measures and uses the method of covariance adjustment. We recommend adjusting for survey mode (if applicable) and respondent education. Survey mode is an administrative variable created during survey fielding and respondent education is a self-reported item collected with the FECC survey. Because education was rarely missing among survey respondents (2.2%), cases with missing data were excluded from the case-mix adjustment model. In data with higher rates of item missingness, missing values could be imputed with the mean within the relevant unit of analysis, such as practice. This method avoids losing large numbers of cases due to item missingness.

Recommended Case-Mix Adjustors
Survey mode is coded with an indicator for whether the respondent was randomized to the phone-only study arm as opposed to the mixed-mode study arm (mail survey with phone follow-up), irrespective of the mode in which the survey was actually completed (for example, if the survey was completed by phone but the participant was randomized to mixed-mode, the survey mode indicator would be “mixed-mode”).

Education is coded as a series of six indicators for the six response categories to the education item from the survey, with one indicator left out of the regression model as the reference category. The choice of reference category is arbitrary and does not affect results. Categories with very small numbers of respondents may need to be combined for modeling purposes. Alternatively, the ordinal education variable could be used (1 df) if it is not feasible to include five education category indicators in a given model.

What is the highest grade or level of school that you have completed?
1=8th grade or less
2=Some high school, but did not graduate
3=High school graduate or GED
4=Some college or 2-year degree
5=4-year college graduate
6=More than 4-year college degree

If a “clinically-adjusted” model that does not include sociodemographic variables (i.e., education) is desired, education may be omitted from the model and survey mode may be retained. To stratify clinically-adjusted scores by education, the case-mix model with survey mode as a covariate could be fit separately within each education category.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate

worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Specific coefficients are included in the file in S.2b.

The case-mix adjustment model is similar to the one used for CAHPS surveys (see pages 54-57 of the Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1, available at https://cahps.ahrq.gov/surveys-guidance/docs/2015_instructions_for_analyzing_data.pdf). The form of the case-mix adjustment model is as follows for linear regressions for continuous (not binary yes/no) measures transformed to a 0-100 scale:
y_ipj= ß_i´x_ipj + µ_ip + e_ipj

The form of the case-mix adjustment model is similar for logistic regressions for binary measures.
logit{P(Y_ipj=1)}= ß_i´x_ipj+ µ_ip+ e_ipj

In both cases, y_ipj is the response to measure i of respondent j from unit p (e.g., state Medicaid program), ß_i is the vector of regression coefficients, x_ipj is the vector of covariate adjustor variables (mode and education category indicators described in S.14 above), µ_ip is an intercept parameter for unit p, and e_ipj is an error term. Adjusted scores at the unit level are estimated by predicted population margins. For continuous measures, adjusted unit scores are constructed as least squares means from the linear regression model (described on pages 54-57 of the Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1, available at https://cahps.ahrq.gov/surveys-guidance/docs/2015_instructions_for_analyzing_data.pdf). For binary measures, adjusted unit scores are the within-unit means of predicted probabilities p^_(ipj)given by the inverse logit function

$$\hat{p}_{ipj} = 1/(1+ e^{(-\hat{ß}_i' x_{ipj} - \hat{µ}_{ip})})$$

where ß ^_i and µ ^_ip are estimated coefficients from the logistic regression model. Adjusted binary measures are transformed to a 0-100 scale by calculating p^_(ipj)*100.

S.16. **Type of score:**
Other (specify):
If other: Each of the quality measures is scored on a 0-100 scale, with higher scores indicating better care. For dichotomous measures, a score of 100 indicates the child received the recommended care; a score of 0 indicates that they did not. Please see Detailed Measure Specifications (see S.2b) for additional measure-specific scoring information.

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
To produce scores for the FECC quality measure set, the following steps were taken, in order:
1. Identify children 0-17 years of age
2. Include only those with parent or legal guardian contact information
3. Run the PMCA algorithm, and retain only those children classified as having complex chronic disease
4. Retain children with at least 4 health care provider visits in the past year
5. Exclude caregivers who speak only a language other than English or Spanish
6. Exclude caregivers if child had died
7. Administer FECC Survey to remaining sample, over the telephone or via mail
8. Score each measure according to detailed measure specifications in S.2b
9. For comparisons between health plans, states, or by demographic groups, adjust scores for caregiver education level (and assigned survey mode, if applicable) using linear or logistic regression.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF a PRO-PM, identify whether (and how) proxy responses are allowed.
We recommend sending the FECC survey to a simple random sample of eligible caregivers. Depending on the size of the population of CMC in question, in some cases it may be appropriate to send the survey to all eligible caregivers.

Regarding minimum sample size recommended, we provide guidance below based on the level of measurement.

State or other geographic level: For comparing state or other entity performance to a national benchmark, we recommend collecting a minimum of 199 responses to detect a small effect size (Cohen's d of 0.2), 34 responses to detect a medium effect size (Cohen's d of 0.5), and 15 responses to detect a large effect size (Cohen's d of 0.8). Cohen's d is calculated as the difference in the state mean and the national mean, divided by the standard deviation of the error. It can be calculated separately for each quality measure in order to determine the sample size needed to detect a specific difference in scores in the particular measure.

For comparing the performances of two states or other entities to one another, we recommend collecting a minimum of 394 responses per state to detect a small effect size (Cohen's d of 0.2), 64 responses per state to detect a medium effect size (Cohen's d of 0.5), and 26 responses per state to detect a large effect size (Cohen's d of 0.8). In this case, Cohen's d is calculated as the difference in the two states' means, divided by the standard deviation of the common error.

Medicaid or CHIP payment model: Recommended minimum sample sizes are the same as those listed for the state level.

Health plan: Recommended minimum sample sizes are the same as those listed for the state level.

Individual provider: These measures cannot be used to compare individual providers, because most individual providers will not have sufficient numbers of children with medical complexity within their patient panels to make meaningful comparisons. In our field-testing, the average number of participating patient families per provider was 2.5, and the median was 1.

Hospital: Not recommended. Care coordination is generally provided within the context of an outpatient primary care or subspecialty medical practice, so it would not make sense for hospitals to measure the quality of care coordination being provided to CMC.

Practice, group, or facility: These measures will likely not be useful for most groups or facilities, because most groups will not have sufficient numbers of children with medical complexity within their patient panels to make meaningful comparisons. To compare between groups, the sample sizes listed above for state apply. However, these measures could potentially be used by a group or facility over time to drive QI efforts, given a large enough population of CMC. We recommend obtaining a minimum of 199 responses per time period from the same group of caregivers, to detect a small effect size (Cohen's d of 0.2), 34 responses per time period to detect a medium effect size (Cohen's d of 0.5), and 15 responses per time period to detect a large effect size (Cohen's d of 0.8). In this case, Cohen's d is calculated as the difference in the mean value at the two measurement time points, divided by the standard deviation of the common error. These calculations assume a correlation between time points of 0.5; with higher correlation (as one might expect when surveying the same caregivers at multiple time points), a larger effect size is detectable for any given sample size.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
The FECC survey can be administered over the telephone or via a mailed version, although we recommend a mixed-mode approach (mailing followed by telephone interview for mail non-responders). A copy of the survey is attached with this submission.

In the mixed-mode approach, two mailings were sent to participants prior to transferring to telephone mode, at which time a maximum of 10 attempts were made to complete the survey by telephone. The telephone survey was administered by trained research assistants using a computer-assisted telephone interview script. Both the mailed and telephone surveys were offered in English and Spanish.

Regarding minimum response rate, we suggest a target of 40% (achieved in our field testing) and a minimum of 25%, primarily on the basis of face validity.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
In general, if a caregiver failed to respond to a survey question required for calculating either the numerator or denominator of a

quality metric, they were excluded from that metric. In the case where there were multiple components used to score a given measure, we required all components to be non-missing in order to score the item.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Patient Reported Data/Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The overall FECC-eligible population is identified using ICD-9 codes and administrative data. Data for the measure numerators and some denominator elements come from caregiver responses to the FECC Survey (attached). The survey was administered via mail and telephone, in English and Spanish.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Other
If other: The FECC quality measures concern care coordination that occurs across the spectrum of health care settings, from inpatient to outpatient to home health. However, the majority of care coordination services assessed were provided by the outpatient clinici

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_FECC_testing_submit.docx

FAMILY EXPERIENCES WITH COORDINATION OF CARE SURVEY
TELEPHONE INTERVIEW VERSION

1.		Your child's main provider is the doctor, physician assistant, nurse or other health care provider who knows the most about your child's health, and who is in charge of your child's care overall.	
1A.	OPEN TEXT (100 CHARACTERS)	What is the name of your child's main provider?	
1B.	EMPTY	The questions in this survey will refer to [FILL 1A.] as "your child's main provider." Please think of that person as you answer the questions.	
2-INTRO	EMPTY	This first set of questions are about the people who help you manage care, treatment and services for your child.	
2.	0=NO (GO TO 17-INTRO) 1=YES (GO TO 3A) 8 = DON'T KNOW (GO TO 3A) 9 = REFUSED (GO TO 3A)	In the last 12 months, did your child visit more than one doctor's office or use more than one kind of health care service, such as physical or speech therapy, or community service, such as home health care or transportation services? IF NEEDED: Other examples of community services are early intervention programs, respite care, and parent or caregiver support services.	
3A.	0=NO (GO TO 3B) 1=YES (GO TO 4) 8 = DON'T KNOW (GO TO 3B) 9 = REFUSED (GO TO 3B)	Did anyone <u>in</u> the main provider's office help you to manage your child's care or treatment from different doctors or care providers?	
3B.	0=NO (GO to #17 Intro) 1=YES (GO TO 3C)	Did anyone else <u>outside</u> of [1A]'s office help you to	

	8 = DON'T KNOW (GO TO 17 INTRO) 9 = REFUSED (GO TO 17 INTRO)	manage your child's care or treatment from different doctors or care providers?	
3C.	1. Another provider from a different office/clinic 2. A care coordinator who isn't part of [FILL 1A's] office staff 3. A social worker who isn't part of [1A's] office staff 4. A care or case manager who isn't part of [1A's] office staff 5. Someone else who isn't part of [1A's] office staff 6.	Who was it that helped you? If more than one person helped you, we want to know the person who helped you most often in the last 12 months.	ALL GO TO 5A
4.	1. Your child's main provider 2. Another doctor or nurse in the main provider's office 3. A clerk or receptionist in the main provider's office 4. A care coordinator in the main provider's office 5. A social worker in the main provider's office 6. A care manager or case manager in the main provider's office 7. Someone else in the main provider's office 8 = DON'T KNOW 9 = REFUSED	Who in the main provider's office helped you? If more than one person helped you, we want to know the person who helped you most often in the last 12 months.	
5a.	1 = Yes, definitely 2 = Yes, somewhat 3 = No 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did the person who helped you with managing your child's care know the important information about your child's health and care needs? Would you say:	
5b.	1 = Yes, definitely 2 = Yes, somewhat 3 = No	In the last 12 months, did the person who helped you with managing your child's care seem informed	

	8 = DON'T KNOW 9 = REFUSED	and up-to-date about the care your child got from other providers? Would you say:	
5c.	1 = Yes, definitely 2 = Yes, somewhat 3 = No 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did the person who helped you with managing your child's care support your decisions about what is best for your child's health and treatment?	
5d	1 = Yes, definitely 2 = Yes, somewhat 3 = No 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did the person who helped you with managing your child's care help you to get appointments to visit other providers?	
5e	1 = Yes, definitely 2 = Yes, somewhat 3 = No 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did the person who helped you with managing your child's care help you to get special medical equipment your child needed like a special bed, wheelchair, or feeding tube supplies?	
6.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did you know how to contact the person who helped you with managing your child's care when you needed help or had a question?	
7.	0=NO (GO TO 11.) 1=YES 8 = DON'T KNOW 9 = REFUSED	<u>In the last 3 months</u>, did this person contact you without you getting in touch with them first?	
8.	CHOOSE ALL 1. During a visit to the main provider's office 2. By telephone 3. By email 4. By mail 5. Some other way 8 = DON'T KNOW 9 = REFUSED	How did he or she contact you? Please tell me all the ways you were contacted. Was it:	
9.	1. Never 2. Sometimes 3. Usually 4. Always 8= DON'T KNOW	In the last 3 months, when the who helped you with managing your child's care contacted you, how often did he or she ask if you had any concerns	

	9 = REFUSED	about your child's health or treatment?	
10.	1. Never 2. Sometimes 3. Usually 4. Always 8= DON'T KNOW 9 = REFUSED	In the last 3 months, when the person who helped you with managing your child's care contacted you, how often did he or she ask if your child's health had changed in any way?	
11.	1. Never 2. Sometimes 3. Usually 4. Always 8= DON'T KNOW 9 = REFUSED	Overall, how often did you get the help you needed to manage your child's care or treatment from different doctors or care providers in the last 12 months?	
12.	1. Very satisfied 2. Somewhat satisfied 3. Somewhat dissatisfied 4. Very dissatisfied	Overall, how satisfied or dissatisfied were you with the help you received in managing your child's care or treatment in the last 12 months?	
13-INTRO		The next few questions ask about your experiences with getting care for your child from specialists.	
13.	0=NO (GO TO 15-INTRO) 1=YES 8 = DON'T KNOW 9 = REFUSED	Specialists are doctors like surgeons, heart doctors, allergy doctors, mental health doctors, and other doctors who specialize in one area of health care. During the last 12 months, did the main provider tell you that your child needed to see a specialist?	
14.	1. Never 2. Sometimes 3. Usually 4. Always 8= DON'T KNOW 9 = REFUSED	Did the person who helped you with managing your child's care contact you to make sure your child got an appointment to see a specialist?	
15-INTRO		The next few questions ask about your experiences with getting community services for you or your child.	
15.	0=NO (GO TO 17-INTRO)	Community services are	

	1=YES 8 = DON'T KNOW 9 = REFUSED	services to help maintain your and your child's health and well-being, which may or may not be ordered by one of your child's doctors. This can include things like home health care, early intervention programs, respite care, help with transportation, and parent or caregiver support services. In the last 12 months, did you or your child need or use community services?	
16.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Did the person who helped you with managing your child's care help you to get the community services you or your child needed?	
17-Intro		The next set of questions asks about different ways in which you might get information about the care your child is receiving. We are interested in summaries you might have received after visiting the main provider's office or after your child was in the hospital.	
17.	0=NO (GO TO 21) 1=YES 8 = DON'T KNOW 9 = REFUSED	A written visit summary sums up what happened during your child's visit to a health care provider. A written visit summary can be available on paper, on a web site, through an app, or sent by email. In the last 12 months, did anyone at the main provider's office give you a written visit summary after your child's visits?	
18a.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	How often did the written visit summaries you got from the main provider's office include a list of your child's health problems at the time of the visit?	

18b.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	How often did the written visit summaries you got from the main provider's office include an up-to-date list of all the prescription medicines your child is taking?	
18c.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	How often did the written visit summaries you got from the main provider's office include an up-to-date list of all the over the counter medicines your child is taking?	
18d.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	How often did the written visit summaries you got from the main provider's office include a list of your child's allergies?	
18e.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	How often did the written visit summaries you got from the main provider's office include the names of all the specialist doctors who help care for your child?	
18f.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	How often did the written visit summaries you got from the main provider's office include the plan for follow-up care for your child after the visit?	
18g.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	How often did the written visit summaries you got from the main provider's office include what to do if your child had a problem after the visit?	
	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, how often was the written visit summary you got from the main provider's office easy to understand?	
20.	1. Never 2. Sometimes 3. Always 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, how often was the written visit summary you got from the main provider's office useful to you and your family?	

21.	0=NO (GO TO 26-INTRO) 1=YES 8 = DON'T KNOW 9 = REFUSED	Has your child had an overnight hospital stay in the last 12 months?	
22.	0=NO (GO TO 25) 1=YES 8 = DON'T KNOW 9 = REFUSED	A written hospital stay summary sums up all that happened during your child's hospital stay. A written hospital stay summary can be available on paper, on a web site, through an app, or sent by email. <u>The last time your child was in the hospital</u>, did your child's doctor, nurse, or other hospital staff give you a written hospital stay summary on the day your child left the hospital?	
	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Did the written hospital stay summary you got include a list of the health problems your child had when he or she left the hospital?	
	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Did the written hospital stay summary you got include a list of all the prescription medicines your child was taking when he or she left the hospital?	
	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Did the written hospital stay summary you got include a list of all the over the counter medicines your child was taking when he or she left the hospital?	
	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Did the written hospital stay summary you got include a list of your child's allergies?	
	0=NO 1=YES	Did the written hospital stay summary you got include the names of all the	

	8 = DON'T KNOW 9 = REFUSED	specialist doctors who helped care for your child during the hospital stay?	
	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Did the written hospital stay summary you got include what the planned follow-up care was for your child after the hospital stay?	
	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Did the written hospital stay summary you got include who to call if your child had problems after the hospital stay?	
24.	1. Yes, definitely 2. Yes, somewhat 3. No 8 = DON'T KNOW 9 = REFUSED	Was the information in the written hospital stay summary you got easy to understand?	
25.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Hospital rounds are the daily visits the health care team makes to patients in the hospital to check up on how they are doing and how well the treatment is working, and what the plan for the day will be. Nurses, doctors, medical students and other health care providers may join hospital rounds to discuss the plan for the day for every patient. <u>The last time your child was in the hospital</u>, did any of your child's doctors or nurses invite you to take part in hospital rounds?	
26-Intro		In addition to information you may get after a visit or a hospital stay, some providers make information available through a web site or an app. We are interested in your experiences with this way of getting information about your child's health	

		and health care.	
26.	0=No (GO TO 29-INTRO) 1=Yes 2= Or are you not sure if the main provider's office has a web site or app? (GO TO 29-INTRO) 9 = REFUSED *Yes* *No → If No, go to #29-Intro* *Or are you not sure if the main provider's office has a web site or app? → If not sure, go to #29-Intro*	In the last 12 months, did the main provider's office have a web site or app you could use between visits to look up information about your child's visits and health care? Would you say:	
27.	0=No 1=Yes 2=Or your child did not get any shots or immunizations in the last 12 months? 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did the main provider's web site or app have a list of the <u>shots or immunizations</u> your child has received? Would you say:	
28.	0=No 1=Yes 2=Or your child did not take any medications in the last 12 months? 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did the main provider's web site or app have a list of your child's <u>medications</u>? Would you say:	
29-Intro		The next set of questions asks about three different types of written care plans the main provider may have created for your child: shared care plans, emergency care plans, and transition care plans. We are interested in your experiences, if any, with these different types of plans.	
29.	0=NO (GO TO 32-INTRO) 1=YES 8 = DON'T KNOW 9 = REFUSED	A shared care plan is a written document that contains information about your child's active health problems, medicines he or she is taking, special considerations that all	

		people caring for your child should know, goals for your child's health, growth and development, and steps to take to reach those goals. Has the main provider created a shared care plan for your child?	
30.	0=NO (GO TO 32-INTRO) 1=YES 8 = DON'T KNOW 9 = REFUSED	Do you have a copy of your child's shared care plan?	
31.	0= No 1= Yes 2= Or are there no goals written in your child's shared care plan? 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, has the main provider or anyone from the main provider's office talked with you about the progress your child was making toward the goals written in his or her shared care plan? Would you say:	
32-Intro		An emergency care plan is a written document that contains important information about your child's health, treatment and medications. It also includes special considerations that all people caring for your child should know, for example, how your child lets you know he or she is in pain, or how to communicate with your child if he or she can't hear or speak. Families often bring the emergency care plan when they take a child to an emergency room or urgent care clinic.	
32.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Has the main provider created an emergency care plan for your child?	

33-INTRO		If your child is at least 15 years old, we are interested in your experiences with making plans for your child's care when he or she becomes an adult. This is sometimes called a transition plan.	
	0=NO (GO TO 35-INTRO) 1=YES 8 = DON'T KNOW 9 = REFUSED	Is your child age 15 or older?	
34.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	Has the main provider created a written transition plan that summarizes how your child's care will change and how it will stay the same when he or she becomes an adult?	
35-Intro		The next set of questions asks about your child's experiences in school.	
35.	0=NO (GO TO 38-INTRO) 1=YES 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did your child attend school?	
36.	0=NO (GO TO 38-INTRO) 1=YES 8 = DON'T KNOW 9 = REFUSED	Because of his or her health condition does your child have any difficulty learning, understanding, or paying attention in class?	
37.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did anyone from the main provider's office contact staff at your child's school to make sure they understood how your child's health condition affected his or her ability to learn, understand or pay attention in class?	
38-Intro		This last set of questions is about you and your child. This information will help us to describe the parents and children who take part in this study.	

38.	1. Very well 2. Well 3. Not well 4. Not at all well	How well do you speak English?	
39.	0=NO (GO TO 46) 1=YES 8 = DON'T KNOW 9 = REFUSED	Do you speak a language other than English at home?	
40.	1. SPANISH 2. SOME OTHER LANGUAGE 8 = DON'T KNOW 9 = REFUSED	What is the language you speak at home?	
41.	1. ENGLISH GO TO #46 2. ANOTHER LANGUAGE 8 = DON'T KNOW 9 = REFUSED	Do you prefer to talk with your child's doctors and care providers in English or in another language?	
42.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did the main provider speak to you in the language you prefer?	
43.	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	In the last 12 months, did anyone in the main provider's office speak to you in the language you prefer?	
44.	1. No visits (GO TO #46) 2. Some visits 3. Most visits 4. All visits 8 = DON'T KNOW 9 = REFUSED	A medical interpreter is a professional who helps you talk with doctors and other providers who do not speak your language. The interpreter can do this over the phone or in-person. In the last 12 months, how often did you need an interpreter during a visit to the main provider?	
45.	1. Never 2. Sometimes 3. Usually 4. Always 8 = DON'T KNOW 9 = REFUSED	When you needed a professional interpreter during a visit to the main provider, how often was an interpreter available?	
46.	1. YES, HISPANIC OR LATINO 2. NO, NOT HISPANIC OR LATINO 8 = DON'T KNOW	Is <u>this child</u> of Hispanic or Latino origin or descent?	

	9 = REFUSED		
47.	1. White 2. Black or African American 3. Asian 4. Native Hawaiian or Other Pacific Islander 5. American Indian or Alaska Native 6. Other 8 = DON'T KNOW 9 = REFUSED	What is this child's race? Please choose one or more from this list:	
47a.	1. ONLY ONE (GO TO 48) 2. 2 (GO TO 47B) 3. 3 (GO TO 47B) 4. 4 OR MORE (GO TO 47B) 8 = DON'T KNOW (GO TO 48) 9 = REFUSED (GO TO 48)	Counting all children living in the household, including this child as well as any adult children, how many children live in the household?	
47b.	ENTER NUMBER (RANGE 1-9) 88 = DON'T KNOW 99 = REFUSED	And how many of these [IF 47A=1,2, OR 3, FILL WITH 47A, OTHERWISE LEAVE BLANK] children have special health care needs? IF R ASKS FOR DEFINITION OF 'SPECIAL NEEDS': There's no definition provided. The researchers are interested in the answer you think is most appropriate. IN: INCLUDE CHILD WHO IS FOCUS OF THIS SURVEY	
48.	18 TO 24 25 TO 34 35 TO 44 45 TO 54 55 TO 64 65 TO 74 75 OR OLDER 8 = DON'T KNOW 9 = REFUSED	What is your age?	
49.	1. MALE 2. FEMALE 8 = DON'T KNOW 9 = REFUSED	I'm required to ask, are you male or female?	
50.	1. YES, HISPANIC OR	Are you of Hispanic or	

	LATINO 2. NO, NOT HISPANIC OR LATINO 8 = DON'T KNOW 9 = REFUSED	Latino origin or descent?	
51.	1. White 2. Black or African American 3. Asian 4. Native Hawaiian or Other Pacific Islander 5. American Indian or Alaska Native 6. Other 8 = DON'T KNOW 9 = REFUSED	What is <u>your</u> race? Please choose one or more from this list.	
52.	1. 8th grade or less 2. Some high school, but did not graduate 3. High school graduate or GED 4. Some college or 2-year degree 5. 4-year college graduate 6. More than 4-year college degree 8 = DON'T KNOW 9 = REFUSED	What is the highest grade or level of school that <u>you</u> have completed?	
THANKS	EMPTY	That's the end of the survey. To thank you for your time, we'd like to get your name and address so we can send you your $20 check. I just need to verify your mailing address.	
CHK_ADDR [IF DO_SURVEY=1]	1-7, REF 1=NAME IS WRONG 2=STREET ADDRESS (1ST LINE) IS WRONG 3=STREET ADDRESS (2ND LINE) IS WRONG 4=CITY IS WRONG 5=STATE IS WRONG 6=ZIP CODE IS WRONG 7=INFORMATION ON RECORD IS CORRECT 9=R REFUSES ADDRESS VERIFICATION AND/OR DOESN'T WANT CHECK (GO TO REF_CHECK)	Should we still send that to: [FILL PRELOAD DATA: FIRST & LAST NAME ADDRESS (1ST LINE) ADDRESS (2ND LINE) CITY, STATE & ZIP] INTERVIEWER: CHECK ALL PARTS OF NAME OR ADDRESS THAT NEED CORRECTING (OR CHECK "INFORMATION ON RECORD IS CORRECT" IF NO CORRECTIONS ARE	

		NECESSARY)	
UPD_NAME [IF CHK_ADDR=1]	OPEN TEXT [100 CHAR]	ENTER CORRECT NAME	
UPD_ADDR [IF CHK_ADDR=2]	OPEN TEXT [100 CHAR]	ENTER CORRECT STREET ADDRESS (1ST LINE)	
UPD_ADDR2 [IF CHK_ADDR=3]	OPEN TEXT [50 CHAR]	ENTER CORRECT STREET ADDRESS (2ND LINE)	
UPD_CITY [IF CHK_ADDR=4]	OPEN TEXT [25 CHAR]	ENTER CORRECT CITY	
UPD_STATE [IF CHK_ADDR=5]	DROP-DOWN LIST OF STATES	ENTER CORRECT STATE	
UPD_ZIP [IF CHK_ADDR=6]	OPEN TEXT [5 CHAR]	ENTER CORRECT ZIP CODE	
VERIFY [IF CHK_ADDR=1-6]	1 1=OK (GO TO REC_CHK)	Let me read this back to you to verify that I've entered everything correctly: [FILL: CONF_NAME CONF_ADDR CONF_ADDR2 CONF_CITY CONF_STATE CONF_ZIP] READ BACK TO MAKE SURE CONTACT INFORMATION IS 100% CORRECT AND COMPLETE. IF ANY CORRECTIONS ARE NEEDED, BACK UP TO THE APPROPRIATE SCREEN.	
REC_CHK	EMPTY	You should receive your check in 4 to 6 weeks.	
REF_CHECK	EMPTY	In that case we won't send you your thank you gift. However, I can give you a toll-free number to call, if you decide later that you would like to claim your $20 check. Would you like to write that down?	

		The number is 1-866-862-4636. You'll also need your ID number, which is #######.	
WHY_HIPAA	0=NO 1=YES (GO TO HIPAA) 8 = DON'T KNOW 9 = REFUSED	Thank you for participating in this survey. As part of our research on improving health care coordination for children with disabilities, we plan to gather additional information from medical records and billing data. We will use medical records to learn about referrals your primary doctor makes to other doctors and about how those doctors communicate with each other. Billing data will help us understand when and how the patient is using healthcare services. This survey information will be linked with the medical record and billing data and will be de-identified, meaning your child's name will not be associated with the data. All information will be kept confidential, shared only among the study team, and protected according to strict data security guidelines. Do you give permission for us to gather information from your child's medical record?	
HIPAA	0=NO 1=YES 8 = DON'T KNOW 9 = REFUSED	This is a permission called a "HIPAA authorization." It is required by the "Health Insurance Portability and Accountability Act of 1996" (known as "HIPAA") in order for us to get information from your medical records or health insurance records to use in	

		this research study. _____ _____ 1. If you give your consent to this HIPAA authorization form, you are giving your permission for the following people or groups to give the researchers certain information about you (described below): Any health care providers or health care professionals or health plans that have provided health services, treatment, or payment for you such as physicians, clinics, hospitals, home health agencies, diagnostics centers, laboratories, treatment or surgical centers. 2. If you sign this form, this is the health information about you that the people or groups listed in #1 may give to the researchers to use in this research study: Any information in your medical records that relates to your participation in this research. These records might include information about mental health, drug or alcohol use, HIV/AIDS or other communicable diseases, or genetic testing. Other information includes: referrals your primary doctor makes to other doctors and about how those doctors communicate with each other; services listed in billing data that will help us understand when and how the patient is using healthcare services. 3. The HIPAA protections	

		that apply to your medical records will not apply to your information when it is in the research study records. Your information in the research study records may also be shared with, used by or seen by collaborating researchers, the sponsor of the research study, the sponsor's representatives, and certain employees of the university or government agencies (like the FDA) if needed to oversee the research study. HIPAA rules do not usually apply to those people or groups. If any of these people or groups reviews your research record, they may also need to review portions of your original medical record relevant to the situation. The informed consent document describes the procedures in this research study that will be used to protect your personal information. You can also ask the researchers any questions about what they will do with your personal information and how they will protect your personal information in this research study. For this particular study, information will be linked with the medical record and billing data and will be de-identified, meaning your child's name will not be associated with the data. All information will be kept confidential, shared only among the study team, and protected according to strict data security	

		guidelines. 4. If this research study creates medical information about you that will go into your medical record, you may not be able to see the research study information in your medical record until the entire research study is over. NOTE: This study will not create medical record information about you or your child. 5. If you want to participate in this research study, you must sign this HIPAA authorization form to allow the people or groups listed in #1 on this form to give access to the information about you that is listed in #2. If you do not want to sign this HIPAA authorization form, you cannot participate in this research study. However, not signing the authorization form will not change your right to treatment, payment, enrollment or eligibility for medical services outside of this research study. 6. This HIPAA authorization will not stop unless you stop it in writing. 7. You have the right to stop this HIPAA authorization at any time. You must do that in writing. You may give your written stop of this HIPAA authorization directly to Principal Investigator or researcher or you may mail it to the department mailing address listed at	

		the top of this form, or you may give it to one of the researchers in this study and tell the researcher to send it to any person or group the researcher has given a copy of this HIPAA authorization. Stopping this HIPAA authorization will not stop information sharing that has already happened. Do we have your permission to access your child's medical record for the purposes of this study?	
FOCUS_INTEREST	0=NO (GO TO END) 1=YES 8 = DON'T KNOW 9 = REFUSED	Thank you for participation in the survey. We also wanted to let you know that later this year researchers may be holding focus groups and one-on-one interviews with family caregivers of children with disabilities in your area. Each session would last 60-90 minutes and participants would receive a $50 gift card in recognition of their time. Would you like to be contacted in the future about participating in a focus group or caregiver interview?	
FOCUS_PREF	1=PHONE (GO TO PRIM#) 2=EMAIL (GO TO EMAIL)	Would you prefer to be contacted about these sessions by phone or by email?	
PRIM#	ENTER PHONE	What is the best number to reach you?	
ADDL_NUMS	0 = NO 1 = YES 8= DON'T KNOW 9 = REFUSED	Are there any other numbers we should try? INTERVIEWER: INCLUDE CURRENT NUMBER IF IT IS STILL A GOOD ALTERNATE NUMBER	PROGRAM TO ALLOW 2 ADDL NUMS
EMAIL	OPEN TEXT (50 CHARACTERS)	What is that email address?	
END	EMPTY	That's the end of the	

		survey. IF CHK_ADDR=1-7: Your $20 check should be mailed to you within 4 to 6 weeks. Thank you. Goodbye.	

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Family Experiences with Coordination of Care (FECC) Measure Set
Date of Submission: 9/29/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient

frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the

percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claim-- D	☒ administrative claims--D
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: caregiver survey—N and D	☒ other: caregiver survey—N and D

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

We used the Washington state Medicaid claims data and the Minnesota state Medicaid claims data to identify our eligible population.

1.3. What are the dates of the data used in testing? Survey dates were 7/2013-11/2013; administrative data used for calculating the PMCA were from January 1, 2010 through December 31, 2012.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice

☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan WA and MN state Medicaid	☒ health plan WA and MN state Medicaid
☐ other:	☐ other:

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

We included participants enrolled in state Medicaid in Washington and Minnesota. We sampled 3000 caregivers of children with medical complexity from each, and received 600 completed surveys from Washington and 609 from Minnesota.

The measures were intended for use at the state Medicaid agency (health plan) level, and could also be used at the practice group level for sufficiently large practices (please see S.20 for recommendations on minimum sample size). For conducting reliability analyses for this submission, we did look at intra-class correlation coefficients (ICCs) at the practice grouping level, because our field testing included only 2 states, which would not permit ICC calculation. For the ICC calculation, we included up to 103 practice groupings (57 from WA, 46 from MN); please see **Table T3** in section 2a2.3 for the number of practice groupings included in calculating the ICC for each indicator.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

For the field test (and majority of our analyses), we surveyed caregivers of CMC insured by Medicaid in Washington and Minnesota. To identify these children, we applied the Pediatric Medical Complexity Algorithm, based on International Classification of Diseases, Ninth Revision, Clinical Modification codes, to classify children with chronic disease according to level of medical complexity.[1] CMC were eligible for inclusion in the field test if they were (1) aged 3 months-17 years; (2) had at least 2 Medicaid eligibility months during the three months prior to obtaining the sample; (3) had at least 4 visits to a healthcare provider during the prior 12 months; and (4) had a healthcare provider who participated in Medicaid. Children were excluded if (1) the child had died; (2) the listed household contact was < 18 years of age; or (3) the caregiver spoke a language other than English or Spanish.

We sampled 1500 caregivers in each state and administered the survey from July to November 2013 via both mixed mode (mail with phone follow-up) and phone only; the survey was

available in English and Spanish. We obtained 600 completed surveys in Washington and 609 in Minnesota for the overall FECC field test. See Table e1 for demographic characteristics of participating caregivers and their children.

For our reliability analysis, we used a subset of the overall participants (see section 1.7 below for more details); the characteristics of those participants are also given in **Table T1.** The demographic composition of the reliability analysis subset was nearly identical to the overall population of participants.

Table T1: Characteristics of children and caregivers participating in the FECC measures field-test overall, and the subset included in the reliability analysis

	Respondents overall (N=1209)	Respondents included in ICC analysis (N=889)
Child characteristics		
Female gender (available for MN only)	262 (43%)	194 (43%)
Child age		
< 2 years	127 (10%)	92 (10%)
2-5 years	270 (22%)	197 (22%)
6-10 years	357 (30%)	265 (30%)
11-13 years	207 (17%)	138 (16%)
14-17 years	248 (20%)	197 (22%)
Child race/ethnicity		
White	585 (48%)	445 (50%)
Hispanic	308 (26%)	227 (26%)
African American	94 (8%)	66 (7%)
Other	195 (22%)	146 (16%)
Missing	27 (2%)	5 (1%)
Caregiver (respondent) characteristics		
Female gender	1150 (95%)	863 (97%)
Caregiver relationship to child	1108 (92%)	831 (93%)
Parent	42 (3%)	34 (4%)
Grandparent	5 (0.4%)	3 (0.3%)
Aunt or uncle	1 (0.1%)	0 (0%)
Other relative	21 (2%)	15 (2%)
Legal guardian	32 (3%)	6 (1%)
Other or Missing		
Caregiver age		
18-24	60 (5%)	48 (5%)
25-34	433 (36%)	318 (36%)
35-44	417 (34%	314 (35%)
45-54	150 (12%)	112 (13%)

55-64	41 (3%)	34 (4%)
65-74	9 (0.7%)	6 (1%)
75+	3 (0.3%)	3 (0.3%)
Other/Unknown	96 (8%)	54 (6%)
Caregiver race/ethnicity		
White	722 (60%)	541 (61%)
Hispanic	250 (21%)	195 (22%)
African American	92 (8%)	64 (7%)
Other	119 (10%)	87 (10%)
Missing	26 (2%)	2 (0.2%)
Caregiver education (highest level completed)		
8th grade or less	70 (6%)	52 (6%)
High school	435 (36%)	318 (36%)
College	639 (53%)	484 (54%)
More than 4-year college degree	38 (3%)	33 (4%)
Not answered or don't know	27 (2%)	2 (0.2%)
Caregiver English language proficiency		
Speaks very well	972 (80%)	727 (82%)
Speaks well	78 (6%)	55 (6%)
Does not speak well	82 (7%)	61 (7%)
Does not speak at all	52 (4%)	42 (5%)
Not answered	25 (2%)	4 (0.5%)
Language of survey completion		
English	1048 (87%)	769 (87%)
Spanish	161 (13%)	120 (14%)
Mode of survey completion		
Mail	435 (36%)	301 (34%)
Telephone only	544 (45%)	416 (47%)
Telephone following mailing	230 (19%)	172 (19%)
State of residency		
WA	600 (50%)	451 (51%)
MN	609 (50%)	438 (49%)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

The only testing for which the analytic sample differed from the overall sample was for calculating the ICC by practices, which we performed as part of our reliability testing. We were able to identify the child's main provider's practice location for all participants from Minnesota, but for only

39% of participants from Washington state, due to differences in IRB stipulations. We therefore compared mean FECC quality measure scores for Washington participants for whom we could and could not identify the main provider's practice location. Given that scores differed significantly for only one out of the 17 total FECC measures when comparing those two groups, we felt confident in proceeding with ICC calculations for only those Washington participants where their child's main provider's practice location could be identified. In addition, the subset used for ICC calculation was almost identical to the overall respondent pool from a demographic standpoint.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Caregiver and child sociodemographic variables we used are those listed in Table TI above: child gender, age and race/ethnicity, and caregiver age, race/ethnicity, English proficiency, and educational attainment.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

For the Pediatric Medical Complexity Algorithm (PMCA), used to identify our overall denominator, accuracy was determined empirically; please see section 2b2 for the validity testing results.

We examined several aspects of the reliability of the FECC caregiver-reported survey measures. While measure development was informed by domains identified in the conceptual framework, measures within each domain were not meant to function as a scale, as they do not measure a single underlying construct but instead measure separate aspects of care coordination

quality. We therefore do not present any measurement of the reliability within domains, and quality measures included in the FECC survey may be used independently of one another.

Several of the measures include measure stems and sub-parts that were intended to function together as a scale and are scored together. For those measures, we used a variation on Cronbach's alpha to establish the reliability of the construct measurement. Given the ordinal nature of the measures, we used polychoric ordinal alphas rather than Pearson correlations. Results were calculated only for participants who were eligible for and answered all measure sub-parts.

We also assessed score reliability by calculating the intra-class correlation coefficient (ICC). For calculating ICCs, we grouped participants by caregiver identified main provider practice, then grouped affiliated practices. We excluded participants with a provider who was not associated with an affiliated group of practices (n=19), given the small numbers of patient participants represented in single non-affiliated practices (n=1-5 participants each). While practice level is not the intended level of aggregation of the FECC measures due to small samples of CMC per practice, we did use this level for conducting reliability testing, as our field test included only 2 states (the intended level of aggregation).

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Table T2: Construct measurement reliability testing for multi-part measures

		Number of items	Average inter-item polychoric correlation	Polychoric Ordinal Alpha*
FECC 5	Care coordinator asked about caregiver concerns, health changes of child	2	0.75	0.86
FECC 8	Care coordinator knowledgeable, supportive, advocated for needs of child	3	0.47	0.73
FECC 9	Contents of written visit summary	6	0.50	0.86

*The polychoric ordinal alpha statistic is a disattenuated Cronbach's alpha for ordinal scales; therefore the commonly accepted rules for describing internal consistency may be employed: α ≥ 0.9 = excellent; 0.9 > α ≥ 0.8 = good; 0.8 > α ≥ 0.7 = acceptable; 0.7 > α ≥ 0.6 = questionable; 0.6 > α ≥ 0.5 = poor; 0.5 < α = unacceptable. Similar to Cronbach's alpha, the polychoric ordinal alpha statistic increases with the number of items on the scale; therefore, it may be lower with fewer items on the scale. (Zumbo BD, Gadermann AM, Zeisser C. Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods.* 2007;6(1):4.)

Table T3: Intra-class correlation coefficients for reliability testing by affiliated practice groups

Measure and Description		N, practice groups	N, patients	ICC (95% CI)	Spearman-Brown predicted reliability with N cases per measured entity[1]			
					N=30	N=50	N=100	N=300
FECC 1	Has care coordinator[2]	92	626	0.05 (0.008, 0.23)	0.59	0.71	0.83	0.94
FECC 3	Care coordinator helped to obtain community services[3]	48	203	0	--	--	--	--
FECC 5	Care coordinator asked about concerns and health changes[2]	59	202	0	--	--	--	--
FECC 7	Care coordinator assisted with specialist service referrals[3]	67	343	0.09 (0.02, 0.27)	0.74	0.82	0.90	0.97
FECC 8	Care coordinator was knowledgeable, supportive & advocated for child's needs[2]	71	418	0	--	--	--	--
FECC 9	Appropriate written visit summary content[2]	78	518	0.03 (0.003, 0.20)	0.46	0.60	0.74	0.90
FECC 14	Health care provider communicated with school staff about child's condition[3]	85	495	0.06 (0.01, 0.20)	0.64	0.75	0.85	0.95
FECC 15	Caregiver has access to medical interpreter[2]	28	89	0	--	--	--	--
FECC 16	Child has shared care plan[3]	103	808	0.12 (0.05, 0.25)	0.80	0.87	0.93	0.98

FECC 17	Child has emergency care plan[3]	103	842	0.08 (0.03, 0.19)	0.71	0.80	0.89	0.96

[1]Predicted reliability =(N)(ICC) / [1 + (N-1)(ICC)]

[2]Linear model

[3]Logistic model

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Our results clearly established the reliability of 8 of the 10 measures included in this submission; the remaining 2 were limited by small samples, but demonstrated excellent validity (see below).

Regarding the reliability testing, in all 3 of the multi-item measures (FECC 5, 8 and 9), the alpha was > 0.7, indicating good inter-item construct reliability. Our ICCs by affiliated practice groups showed statistically significant ICCs for 5 of the other measures for which construct reliability was not measured (FECC 1, 7, 14, 16, and 17), demonstrating reliable variation by practice, and predicting good to excellent score reliability with the per-entity sample sizes we recommend (>30, and preferably larger, depending on desired detectable effect sizes).

The 2 measures for which we were unable to establish reliability within our current sample (FECC 3 and FECC 15) were limited by small sample sizes, but demonstrated excellent validity (see next section).

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

For the PMCA, which we used to identify CMC for inclusion in the overall denominator, the algorithm-determined classifications of 700 children (no chronic disease, non-complex chronic disease, or complex chronic disease) were compared to a gold-standard classification determined by clinician chart-review. Sensitivity and specificity of the PMCA were calculated, using first Washington Medicaid billing data, then Seattle Children's Hospital billing data, to determine the PMCA category; see **Table T4,** below. These methods and results are available in detail in Simon TD et al. "Pediatric Medical Complexity Algorithm: A New Method to Stratify Children by Medical Complexity." *Pediatrics*. Volume 133, Number 6, June 2014.

Unlike with some quality measures, no gold standard exists for family experiences with coordination of care, which the FECC Survey aims to capture. Therefore, true criterion validity cannot be established. However, during the process of quality measure development and specification, survey development, cognitive interviewing, and field-testing, many efforts were made to demonstrate the content and construct validity of the quality measures included in the FECC Survey, detailed below.

Content Validity: The RAND-UCLA Modified Delphi Method

The content validity of the quality measures in the FECC measure set was established using the RAND-UCLA Modified Delphi Method. The process began with the nomination of 20 individuals by 10 stakeholder organizations including the American Academy of Pediatrics, the Academic Pediatric Association, the Society for Hospital Medicine, the Children's Hospital Association, the Medicaid Medical Directors Learning Network, Family Voices, the American Academy of Child and Adolescent Psychiatry, the Society for Adolescent Medicine, the National Association of Pediatric Nurse Practitioners, and the Society for Developmental and Behavioral Pediatrics. Nine of the nominees agreed to be members of our multi-stakeholder Delphi panel. All panelists were people deemed by the nominating organizations to have substantial expertise and/or experience related to care coordination for CMC (see Ad.1 for a list of panel members). The panel read the literature reviews written by project staff and reviewed and scored each proposed quality measure on validity. This method is a well-established, structured approach to measure evaluation that involves two rounds of independent panel member scoring, with group discussion in between.[2] After reviewing literature reviews and draft quality measures, panel members were asked to rate each measure's validity on a scale from 1 (low) to 9 (high). Validity was assessed by considering whether there was adequate scientific evidence or expert consensus to support its link to better outcomes; whether there would be health benefits associated with receiving measure-specified care; whether they would consider providers who adhere more consistently to the quality measure to be providing higher quality

care; and whether adherence to the measure is under the control of health care providers and/or systems. The Delphi method has been found to be reliable and to have content, construct and predictive validity.[3-7] For a quality measure or measure component to move to the next stage of measure development, it had to have a median validity score $\geq$ 7 (1-9 scale) and be scored without disagreement based on the mean absolute deviation from the median after the second round of scoring. This process ensures that only measures widely judged to be valid moved forward into measure specification. See **Table T5** for scores by measure.

Cognitive Interviews

Twenty-one of the 31 quality measures that were endorsed by the Delphi panel were operationalized into survey items. Survey items were developed to specify: 1) the eligible population of CMC for each measure (the denominator) and 2) whether the indicated care was received among those eligible (the numerator). Survey items underwent cognitive interviews with 9 parents, in Spanish or English, to establish understandability by families. By using cognitive interviews prior to field testing, team members identified questions that required revision that might otherwise have impacted survey validity. For example, caregivers interviewed could not reliably explain what was meant by the term "care coordination." Thus this terminology was removed from the FECC survey and the phrase, "…help with managing your child's care," was used instead, due to its better understandability by the interviewed caregivers.

Convergent Validity: Field Testing

The construct validity of the measures in the FECC Survey was established by demonstrating convergent validity with 2 previously validated measures of outpatient care experiences from the Clinician and Group (CG) Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Child 12-month Survey,[8] and a measure adapted from the Adult Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Heath Plan 4.0 supplemental item on care coordination. [9]

For the field test, we surveyed caregivers of CMC insured by Medicaid in Washington and Minnesota. To identify these children, we applied the Pediatric Medical Complexity Algorithm, based on International Classification of Diseases, Ninth Revision, Clinical Modification codes, to classify children with chronic disease according to level of medical complexity.[1] CMC were eligible for inclusion if they were (1) aged 3 months-17 years; (2) had at least 2 Medicaid eligibility months during the three months prior to obtaining the sample; (3) had at least 4 visits to a health care provider during the prior 12 months; and (4) had a health care provider who participated in Medicaid. Children were excluded if (1) the child had died; (2) the listed household contact was < 18 years of age; or (3) the caregiver spoke a language other than English or Spanish.

We sampled 1500 caregivers in each state and administered the survey from July to November 2013 via both mixed mode (mail with phone follow-up) and phone only; the survey was available in English and Spanish. We obtained 600 completed surveys in Washington and 609 in Minnesota.

FECC Survey Questions

The FECC Survey was comprised of 45 questions, including 6 questions related to care coordination outcomes, and the CG CAHPS experience measures described below. Of the outcome measures, 3 were newly developed, 1 was adapted from the National Survey of Children with Special Health Care Needs,[10] and 2 were adapted from the adult CAHPS Health Plan Survey (V4.0) Supplemental Items; one of the adapted measures that was used as a validation metric is described in greater detail below.[9] All measures were on a 0-100 scale and were scored such that higher scores indicate better care. For binary measures, 100 indicated receipt of the recommended care, 0 indicated non-receipt.

CG CAHPS Experience Measures

Caregiver experience was measured using the overall provider rating and 4 questions concerning access to care (the Access Composite) from the CG-CAHPS Child 12-month Survey.[11] Responses to the access questions were scored on a 0-100 scale (Never = 0, Sometimes = 33.3, Usually = 66.7, Always = 100); caregivers that answered at least 1 of the 4 questions received an Access Composite score calculated as the mean of the non-missing responses.

Adapted Adult CAHPS Health Plan Supplemental Care Coordination Outcome Measure

Receipt of needed care coordination was assessed using an adapted version of the Adult CAHPS Health Plan Supplemental Care Coordination Outcome Measure.[9] The measure was adapted to facilitate a caregiver responding in relation to their child rather than an adult responding in relation to themselves. The question asked caregivers, “Overall, how often did you get the help you needed to manage your child’s care or treatment from different doctors or care providers in the last 12 months?” Responses were scored on a 0-100 scale (Never = 0, Sometimes = 33.3, Usually = 66.7, Always = 100).

Analyses

We used linear regression to examine the association between measure scores and the two CG-CAHPS measures and the one adapted CAHPS measure described above, unadjusted and

adjusted for caregiver education and assigned survey mode. This analysis was carried out for each quality measure.

Tables T6-T8 show results of these validation analyses using the CG-CAHPS Access Composite, Overall Provider Rating, and Health Plan CAHPS Getting Needed Help with Managing Care measure.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Table T4: PMCA sensitivity and specificity for correctly designating a child as having complex chronic disease, using WA Medicaid data and Seattle Children's Hospital (SCH) data

	Sensitivity (95% CI)	Specificity (95% CI)
Seattle Children's Hospital data	84 (80–88)	92 (89–94)
WA Medicaid data	89 (85–92)	85 (81–89)

Table T5: Content validity of FECC Survey measures, based on Delphi panel validity scores; rating scale is 1-9, with 9 indicating highest validity

	Measure	Median	Mean

		validity score	absolute deviation from median
FECC 1	Has care coordinator	8.0	0.7
FECC 3	Care coordinator helped to obtain community services	7.0	0.6
FECC 5	Care coordinator asked about:		
FECC 5a	Caregiver concerns	8.0	0.7
FECC 5b	Health changes in the child	8.0	0.7
FECC 7	Care coordinator assisted with specialist service referrals	7.0	0.8
FECC 8	Care coordinator was:		
FECC 8a	Knowledgeable	7.0	0.6
FECC 8b	Supportive	7.0	0.4
FECC 8c	Advocated for child's needs	8.0	0.7
FECC 9	Appropriate written visit summary content:		
FECC 9a	Current problem list	7.0	0.9
FECC 9b	Current medications	8.0	1.0
FECC 9c	Drug allergies	8.0	1.1
FECC 9d	Specialists involved in child's care	8.0	1.1
FECC 9e	Planned follow-up	8.0	1.0
FECC 9f	What to do for problems related to outpatient visit	8.0	0.7
FECC 14	Health care provider communicated with school staff about child's condition	7.0	1.1
FECC 15	Caregiver has access to medical interpreter when needed	8.0	0.6
FECC 16	Child has shared care plan	7.0	1.1
FECC 17	Child has emergency care plan	7.0	0.7

Table T6: Validation of developed measures using CG-CAHPS Access Composite as validation metric

		Access Composite (0-100)			
		Unadjusted		Adjusted[1]	
		N	β (95%CI)	N	β (95%CI)
CC 1	Has care coordinator	840	0.08 (0.05, 0.11)***	771	0.07 (0.04, 0.11)***
CC 3	Care coordinator helped to obtain community services	278	0.06 (0.02, 0.1)**	250	0.06 (0.02, 0.11)**
CC 5	Care coordinator asked about concerns and health changes	267	0.3 (0.22, 0.38)***	244	0.29 (0.2,1 0.37)***
CC 7	Care coordinator assisted with specialist service referrals	454	0.06 (0.02, 0.1)**	417	0.06 (0.02, 0.10)**
CC 8	Care coordinator was knowledgeable, supportive and advocated for child's needs	557	0.21 (0.12, 0.29)***	513	0.20 (0.12, 0.29)***
CC 9	Appropriate written visit summary content	706	0.26 (0.19, 0.33)***	649	0.25 (0.18, 0.32)***
CC 14	Health care provider communicated with school staff about child's condition	652	0.06 (0.03, 0.1)***	601	0.06 (0.03, 0.1)***
CC 15	Caregiver has access to medical interpreter when needed	115	0.27 (0.09, 0.46)**	113	0.28 (0.09, 0.46)**
CC 16	Child has shared care plan	1089	0.06 (0.04, 0.09)***	998	0.06 (0.03, 0.08)***
CC 17	Child has emergency care plan	1132	0.06 (0.03, 0.09)***	1042	0.06 (0.03, 0.09)***

<0.05; **p<0.01, ***p<0.001

djusted for mode of survey administration (Randomized to mixed mode or phone only mode) and caregiver education

Table T7: Validation of developed measures using CG-CAHPS Overall Provider Rating as validation metric

		Overall Provider Rating (0-100)			
		Unadjusted		Adjusted[1]	
		N	β (95%CI)	N	β (95%CI)
CC 1	Has care coordinator	828	0.07 (0.04, 0.09)***	768	0.06 (0.04, 0.09)***
CC 3	Care coordinator helped to obtain community services	275	0.06 (0.02, 0.09)**	250	0.06 (0.02, 0.10)**
CC 5	Care coordinator asked about concerns and health changes	263	0.16 (0.11, 0.21)***	244	0.17 (0.11, 0.22)***
CC 7	Care coordinator assisted with specialist service referrals	448	0.08 (0.05, 0.1)***	416	0.07 (0.05, 0.10)***
CC 8	Care coordinator was knowledgeable, supportive and advocated for child's needs	551	0.25 (0.19, 0.31)***	513	0.25 (0.19, 0.32)***
CC 9	Appropriate written visit summary content	705	0.16 (0.11, 0.21)***	648	0.15 (0.10, 0.2)***
CC	Health care provider communicated with school staff about child's condition	654	0.06 (0.03, 0.09)***	601	0.05 (0.02, 0.08)***
CC	Caregiver has access to medical interpreter when needed	117	0.08 (-0.01, 0.15)*	114	0.07 (0.0, 0.14)
CC	Child has shared care plan	1089	0.07 (0.05, 0.09)***	996	0.06 (0.04, 0.08)***
CC	Child has emergency care plan	1132	0.06 (0.04, 0.09)***	1040	0.06 (0.03, 0.08)***

<0.05; **p<0.01, ***p<0.001
djusted for mode of survey administration (Randomized to mixed mode or phone only mode) and
regiver education

Table T8: Validation of developed measures using Got Needed Help adapted Health Plan CAHPS measure as validation metric

		Got Needed Help Coordinating Care			
		Unadjusted		Adjusted[1]	
		N	β (95%CI)	N	β (95%CI)
CC 1	Has care coordinator		N/A: outcome measure eligibility required having a care coordinator		N/A: outcome measure eligibility required having a care coordinator
CC 3	Care coordinator helped to obtain community services	277	0.05 (-0.01, 0.1)	255	0.06 (0.01, 0.12)*
CC 5	Care coordinator asked about concerns and health changes	267	0.35 (0.25, 0.45)***	250	0.36 (0.26, 0.46)***
CC 7	Care coordinator assisted with specialist service referrals	453	0.1 (0.05, 0.15)***	424	0.11 (0.05, 0.16)***
CC 8	Care coordinator was knowledgeable, supportive and advocated for child's needs	555	0.55 (0.44, 0.66)***	522	0.55 (0.44, 0.66)***
CC 9	Appropriate written visit summary content	405	0.36 (0.24, 0.48)***	383	0.36 (0.24, 0.48)***
CC	Health care provider communicated with school staff about child's condition	348	0.04 (-0.02, 0.1)	332	0.06 (0, 0.12)*
CC	Caregiver has access to medical interpreter when needed	64	0.16 (-0.17, 0.48)	64	0.16 (-0.19, 0.52)
CC	Child has shared care plan	555	0.08 (0.03, 0.12)***	522	0.09 (0.04, 0.13)***
CC	Child has emergency care plan	571	0.07 (0.02, 0.12)**	541	0.07 (0.02, 0.12)**

<0.05; **p<0.01, ***p<0.001
djusted for mode of survey administration (Randomized to mixed mode or phone only mode) and
egiver education

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Our results demonstrated the validity of the FECC quality measures. The PMCA, used for identifying the denominator, demonstrated excellent sensitivity and specificity compared to a gold-standard population, revealing both reliability and validity. All of the measures demonstrated excellent content validity, with median validity scores $\geq$7 (out of 9) following the Delphi panel. All 10 of the FECC measures set measures were associated with better experience in terms of access to care, in both unadjusted and adjusted analyses. Nine of the measures were significantly associated with overall provider rating in both unadjusted and adjusted analyses, while the remaining measure (FECC 15) was significantly associated in only the unadjusted analysis, likely due to a smaller sample size.

Nine of the measures were also associated with getting all of the care coordination help the family needed in adjusted analyses, and the one without a significant association was, again, quite limited by sample size (FECC 15), but had demonstrated the strongest content validity rating of all of the measures. These additional results demonstrate convergent validity between the quality measures included in the FECC Survey and the CAHPS items that we would also expect to be influenced by the quality and degree of care coordination assistance a parent receives for a CMC. These results demonstrate that the measures in the FECC Survey are indeed measuring what they purport to measure.

2b3. EXCLUSIONS ANALYSIS
NA no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We did not test the impact of exclusions on our FECC measure set scores, as the survey was not sent to families to whom the exclusions applied; therefore, the data needed to test the impact of exclusions was not available. However, our exclusions (the child had died or the caregiver spoke a language other than English or Spanish) involved only 33 families out of 3000 identified for sampling, or 1.1%, so we do not expect that they would have substantially impacted our results.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Not applicable—see above

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Not applicable—see above

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

To identify meaningful differences in FECC quality measure scores, we compared mean measure scores by state Medicaid agency, after adjusting for caregiver education and assigned survey mode. Raw scores were stratified by state, adjusted using linear or logistic regression as outlined in S.14, then compared. Statistical significance was determined using the risk adjustment regression models, and was defined as having a P-value < .05.

We also sought to identify processes in which disparities exist by race/ethnicity and caregiver English proficiency, as described in 1b.4. For those analyses, logistic (for dichotomous outcomes) and linear (for continuous outcomes) regressions were used to identify statistically significant differences in process measure scores by race/ethnicity or English proficiency, in both unadjusted and adjusted models.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Table T9: Mean FECC measure scores by state, adjusted for caregiver education and assigned survey mode.

Measure ID	Measure description	Adjusted mean score (95% CI)[1]	
		Minnesota Medicaid	Washington Medicaid
FECC 1	Has care coordinator	71.4 (67.2, 75.7) N=421	73.03 (68.7, 77.3) N=403
FECC 3	Care coordinator helped to obtain community services	52.4 (44.7, 60.2) N=154	46.4 (37.4, 55.3) N=117
FECC 5	Care coordinator asked about	82.8 (78.5, 87.1)	78.2 (73.5, 82.9)

	concerns and health changes	N=142	N=117
FECC 7	Care coordinator assisted with specialist service referrals	67.6 (61.3, 73.9)* N=210	77.3 (71.9, 82.6)* N=236
FECC 8	Care coordinator was knowledgeable, supportive and advocated for child's needs	84.9 (82.7, 87.0) N=280	83.5 (81.3, 85.7) N=265
FECC 9	Appropriate written visit summary content	82.9 (80.9, 84.9)** N=407	78.4 (76.0, 80.7)** N=289
FECC 14	Health care provider communicated with school staff about child's condition	29.5 (24.6, 34.3) N=334	27.7 (22.8, 32.5) N=317
FECC 15	Caregiver has access to medical interpreter when needed	85.3 (79.6, 91.0) N=66	80.8 (74.2, 87.4) N=50
FECC 16	Child has shared care plan	48.8 (44.7, 52.8)*** N=547	38.0 (34.0, 42.0)*** N=536
FECC 17	Child has emergency care plan	24.5 (21.0, 28.0)*** N=574	15.5 (12.4, 18.5)*** N=552

*p<0.05; **p<0.01, ***p<0.001

[1] Adjusted for mode of survey administration (Randomized to mixed mode or phone only mode) and caregiver education

Table T10: Unadjusted FECC quality measure scores, by child race/ethnicity

FECC Measure	Measure ID	White (n=585)	Hispanic (n=308)	Black (n=94)	Other (n=222)
Has care coordinator	FECC 1	70.3	82.1**	61.0	70.3
Care coordinator helped to obtain community services	FECC 3	43.5	58.0	63.6	54.0
Care coordinator asked about concerns and health	FECC 5	83.7	75.9*	83.3	82.1

changes					
Care coordinator assisted with specialist service referrals	FECC 7	69.2	75.8	81.5	77.6
Care coordinator was knowledgeable, supportive and advocated for child's needs	FECC 8	85.4	81.7*	81.5	86.5
Appropriate written visit summary content	FECC 9	81.0	80.5	86.8*	78.7
Health care provider communicated with school staff about child's condition	FECC 14	24.7	30.5	39.2*	32.2
Caregiver has access to medical interpreter when needed	FECC 15	N/A[1]	N/A[1]	N/A[1]	N/A[1]
Child has shared care plan	FECC 16	36.4	52.4***	65.9***	41.2
Child has emergency care plan	FECC 17	16.5	23.8*	44.6***	14.9

Compared to white reference group using linear or logistic regression: *p<0.05 **p<0.01 ***p<0.001

[1] N/A: only one child with race/ethnicity other than Hispanic was eligible for this measure

Table T11: FECC quality measure scores by child race/ethnicity, adjusted for caregiver education and assigned study mode

FECC Measure	Measure ID	White (n=585)	Hispanic (n=308)	Black (n=94)	Other (n=222)
Has a care coordinator	**FECC 1**	71.8	78.8	60.5	69.9
Care coordinator helped to obtain community services	**FECC 3**	45.2	51.7	68.4	52.9
Care coordinator asked about concerns and health changes	**FECC 5**	83.6	75.8	82.8	80.6

Care coordinator assisted with specialist service referrals	**FECC 7**	69.9	73.4	81.4	76.8
Care coordinator was knowledgeable, supportive and advocated for child's needs	**FECC 8**	85.6	81.3*	81.1	86.3
Appropriate written visit summary content	**FECC 9**	81.0	80.8	86.5*	78.2
Health care provider communicated with school staff about child's condition	**FECC 14**	25.6	28.2	39.6*	32.8
Caregiver has access to medical interpreter when needed	**FECC 15**	N/A[1]	N/A[1]	N/A[1]	N/A[1]
Child has shared care plan	**FECC 16**	38.3	47.4*	65.5***	41.7
Child has emergency care plan	**FECC 17**	17.1	21.9	43.8***	14.3

Compared to white reference group using linear or logistic regression: *p<0.05 **p<0.01 ***p<0.001

[1] N/A: only one child with race/ethnicity other than Hispanic was eligible for this measure

Table T12: Unadjusted FECC measure scores by English proficiency

Measure	Measure ID	English proficient (n=1094)	LEP (n=154)
Has care coordinator	**FECC 1**	70.0	88.3***
Care coordinator helped to obtain community services	**FECC 3**	46.8	67.7*
Care coordinator asked about concerns and health changes	**FECC 5**	82.6	71.3**
Care coordinator assisted with specialist service referrals	**FECC 7**	71.2	82.8
Care coordinator was knowledgeable, supportive and advocated for child's needs	**FECC 8**	85.1	79.1**

Appropriate written visit summary content	**FECC 9**	81.7	76.1*
Health care provider communicated with school staff about child's condition	**FECC 14**	26.2	50***
Caregiver has access to medical interpreter when needed	**FECC 15**	N/A[1]	N/A[1]
Child has shared care plan	**FECC 16**	40.8	61.8***
Child has emergency care plan	**FECC 17**	19.3	27.2*

Compared to English proficient reference group using linear or logistic regression: *p<0.05 **p<0.01 ***p<0.001

[1] No English proficient respondents are eligible for this measure

Table T13: FECC measure scores by English proficiency, adjusted for caregiver education and assigned study mode

Measure	Measure ID	English proficient (n=1094)	LEP (n=154)
Has care coordinator	**FECC 1**	71.0	83.4
Care coordinator helped to obtain community services	**FECC 3**	48.7	55.4
Care coordinator asked about concerns and health changes	**FECC 5**	82.9	69.2*
Care coordinator assisted with specialist service referrals	**FECC 7**	71.3	81.6
Care coordinator was knowledgeable, supportive and advocated for child's needs	**FECC 8**	85.3	77.5**
Appropriate written visit summary content	**FECC 9**	81.8	75.6*
Health care provider communicated with school staff about child's condition	**FECC 14**	26.4	48.6**
Caregiver has access to medical interpreter when needed	**FECC 15**	N/A[1]	N/A[1]

Child has shared care plan	**FECC 16**	41.7	55.1*
Child has emergency care plan	**FECC 17**	19.5	23.8

Compared to English proficient reference group using linear or logistic regression: *p<0.05 **p<0.01 ***p<0.001
[1] No English proficient respondents are eligible for this measure

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

We identified statistically significant differences by state Medicaid agency in quality of care coordination for CMC for 4 of our measures. While minimum clinically important differences (MCIDs) have not been established for these measures, for 3 of the measures, the difference was at or close to 10 points (on a 100-point scale), which is almost certainly of clinical import. While we did not identify differences in all of our measures, our field-test involved only 2 states, and some of our measures applied only to a subset of CMC whose caregivers completed the FECC Survey (e.g., needing an interpreter); we expect that future work is likely to detect differences by entity in other FECC measures as well.

We also identified racial/ethnic and linguistic disparities in FECC measure scores, as discussed in section 1b.4. Many of these differences were also in the realm of 10 points or more, which would be of clear clinical import.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note*: This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a*

method; what statistical analysis was used)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Survey non-response: We tracked survey non-response and failure to contact during field-testing. However, we were significantly limited in our ability to conduct non-response analyses by the Washington and Minnesota Medicaid IRBs, who would not permit us to use the demographic data available through administrative sources from non-respondents. We were therefore unable to meaningfully compare our respondents to non-respondents.

However, we would generally expect to have lower response rates for low SES and non-English speaking caregivers based on other survey results (Elliott, Zaslavsky et al. 2009, Elliott, Edwards et al. 2005, Kahn, Liu et al. 2003, Klein, Elliott, et al. 2011). We were reassured by our ability to achieve meaningful participation from these groups, likely due to our mixed mode of administration.

There are several strategies for reducing nonresponse bias in survey estimates. These include increasing the response rate, weighting respondents so that the distribution of respondents' characteristics is more representative of the distribution in the sample frame with respect to observable characteristics, and patient-mix adjustment.

Strategies to increase response rates include using more concerted tracing, incentives, or follow-up efforts (Fowler, Gallagher et al. 2002, Gallagher, Fowler et al. 2005, Andresen, Machuga et al. 2008). Increasing response rates will not necessarily increase the representativeness of the sample, however. For example, two studies of telephone surveys found that efforts to enlist cooperation from more respondents resulted in only small increases in response rate and did not

increase representativeness to a significant degree (Keeter, Miller et al. 2000, Curtin, Presser et al. 2005). There is evidence, however, that multimodal approaches, similar to what we used in the FECC field test with mail followed by telephone follow-up for mail non-responders, reduce nonresponse bias because different members of the population are more likely to respond to each mode of data collection (Fowler, Gallagher et al. 2002, Beebe, Davern et al. 2005, Peytchev, Baxter et al. 2009). For example, older Medicare beneficiaries are more likely to respond by mail than by telephone (Zaslavsky, Zaborski et al. 2002, Elliott, Zaslavsky et al. 2009).

When surveys adjust for differences across comparison units in patient-mix, as typically is done with CAHPS surveys and as we recommend, any nonresponse bias associated with these the patient characteristics used for patient-mix adjustment comparisons is reduced (Farley, Elliott et al. 2011). For example, two CAHPS studies found that patient-mix adjustment accounted for any nonresponse bias that could have been addressed through weighting (Elliott, Edwards et al. 2005, Elliott, Zaslavsky et al. 2009). When patient-mix adjustment suffices to address nonresponse bias, it generally does so with greater statistical efficiency than nonresponse weighting, resulting in estimates of equal reliability and precision with smaller sample sizes than would be required with nonresponse weighting. So, while we were not able to test directly for bias in our sample related to non-response, we feel confident based on previous work that we have taken all reasonable steps to minimize the risk of bias related to non-response.

Missing data from survey respondents: Regarding missing responses to particular survey questions on otherwise completed surveys, we tracked missingness for each quality measure, but due to the overall low levels of missing data among those who completed the survey, we did not formally evaluate for bias.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Survey non-response: We achieved an overall survey response rate of 40% (1209 out of 3000), which was quite good given that 632 of the original 3000 (21%) were unable to be contacted (bad phone number or undeliverable mail); only 285 (9.5%) actively refused participation, and another 525 (17.5%) passively refused by non-response. As mentioned above, due to IRB constraints, we were unable to formally evaluate the non-respondents in comparison to he respondents.

However, we successfully achieved a racially and linguistically diverse sample from a generally low-income population (as all of the children were insured by Medicaid), with over 50% of the children from a racial or ethnic background other than non-Hispanic white, and 20% of caregivers reporting speaking English less that very well.

Missing data from survey respondents: For missing data where someone was eligible for the measure but did not respond to any of the survey question components, we didn't feel comfortable imputing responses (see below for rationale). For missing data where someone was eligible and responded to some but not all of the components we considered a) imputation b) taking the average of the non-missing components and c) only scoring them if all components were answered. Imputation would have provided complete data and maintained consistent weighting for all components for all participants, but it would have required making assumptions about response patterns and patterns of care coordination provision that we did not feel we had the data to justify. Using the average of non-missing components would allow us to use all available data, but would mean that in some cases, different components would be weighted differently across participants (e.g., for a question with 3 components, for most participants, each component would make up 1/3 of the score, but if a participant answered only 1 of the 3 questions, that question would make up the entire score on the measure). Only scoring measures if all components were answered would result in excluding some data that caregivers had provided, but would avoid making incorrect assumptions and differential weighting of particular care processes. We opted to score measures only if all component items were answered.

The frequency of missing data, by measure, is listed below (Table T14).

Table T14: Missing responses from otherwise completed FECC Surveys, by measure

ID	Measure description	N	Missing responses*
FECC 1	Has care coordinator	840	13
FECC 3	Care coordinator helped to obtain community services	278	4
FECC 5	Care coordinator asked about concerns and health changes	267	3
FECC 7	Care coordinator assisted with specialist service referrals	454	1
FECC 8	Care coordinator was knowledgeable, supportive and advocated for child's needs	557	52
FECC 9	Appropriate written visit summary content	706	90

FECC 14	Health care provider communicated with school staff about child's condition	652	1
FECC 15	Caregiver has access to medical interpreter when needed	115	0
FECC 16	Child has shared care plan	1089	114
FECC 17	Child has emergency care plan	1132	71
*Missing responses are a combination of questions skipped on the mail survey, refused on the telephone survey, or questions to which the respondent said "I don't know". Indicators with a stem and multiple sub-parts, such as FECC 9, had more opportunities for a caregiver to skip or refuse a sub-question and so generally had greater numbers of total missing responses.			

We do not have data related to the frequency of missing data by patient's provider, given that these measures were not specified for use at the provider level, and the median number of patients per provider was 1.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Survey non-response: We were reassured that we achieved a racially and linguistically diverse sample from a generally low-income population of caregivers. Although we were unable to conduct a formal analysis of non-respondents, using a mixed-mode approach we captured a diverse study population that included representatives from groups that are historically harder to reach with surveys, such as individuals with limited English proficiency. In addition, for the reasons discussed above in 2b.7.1, we feel confident that case-mix adjustment should adequately address a majority of the bias introduced by non-response.

Missing data from survey respondents: Overall, there were low levels of missing data for the majority of FECC quality measures, and the number of missing responses did not increase over the course of the survey (i.e., there was no evidence to suggest that missing responses were the result of respondent fatigue). Because these are caregiver-reported process measures, and in some cases a caregiver may genuinely not know whether a particular process had or had not occurred for a given child, we opted to score only those measures for which caregivers gave a definitive response.

Because the vast majority of respondents had different providers and so were presumably interacting with different care coordinators and different local health care systems, we did not feel comfortable imputing missing responses, as doing so would have required making assumptions about patterns of care coordination processes for which we have no empirical evidence. We similarly opted to score multi-component items only if all components were completed, to avoid the situation in which components would receive different weighting between respondents based on how many components they answered.

1. Simon TD, Cawthon ML, Stanford S, et al. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. Pediatrics 2014;133:e1647-54.
2. Brook RH. The RAND/UCLA appropriateness method. In: McCormick KA, Moore SR, Siegel RA, eds. Clinical practice guidelines development:methodology perspectives. Rockville, MD: Agency for Health Care Policy and Research; 1994.
3. Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. N Engl J Med 1998;338:1888-96.
4. Shekelle PG, Chassin MR, Park RE. Assessing the predictive ability of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy. Int J Technol Assess Health Care 1998;14:707-27.
5. Kravitz RL, Park RE, Kahan JP. Measuring the clinical consistency of panelists' appropriateness ratings: the case of coronary artery bypass surgery. Health Policy 1997;42:135-43.
6. Hemingway H, Crook AM, Feder G, et al. Underuse of coronary revascularization procedures in patients considered appropriate candidates for revascularization. N Engl J Med 2001;344 645-54.
7. Selby JV, Fireman BH, Lundstrom RJ, et al. Variation among hospitals in coronaryangiography practices and outcomes after myocardial infarction in a large health maintenance organization. N Engl J Med 1996;335:1888-96.
8. Clinician and Group Surveys. at https://cahps.ahrq.gov/surveys-guidance/cg/instructions/index.html.)
9. Hays RD, Martino S, Brown JA, et al. Evaluation of a Care Coordination Measure for the Consumer Assessment of Healthcare Providers and Systems (CAHPS) Medicare survey. Med Care Res Rev 2014;71:192-202.
10. Blumberg SJ, Welch EM, Chowdhury SR, Upchurch HL, Parker EK, Skalland BJ. Design and operation of the National Survey of Children with Special Health Care Needs, 2005-2006. Vital Health Stat 1 2008:1-188.
11. . at https://cahps.ahrq.gov/surveys-guidance/cg/cgkit/1309_CG_Measures.pdf.)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Other

If other: Caregiver report via survey

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

Administrative data are used to identify children eligible for the FECC Survey, using billing data (ICD-9 codes) for the Pediatric Medical Complexity Algorithm. Such billing data are readily available to practices, hospitals, and insurers. However, the caregiver-reported measures on the FECC Survey must be collected prospectively.

In our field test, we determined that it was feasible to collect information on care coordination quality from parents and caregivers of CMC. We achieved an overall survey response rate of 40% (1209 out of 3000), which was quite good given that 632 of the original 3000 (21%) were unable to be contacted (bad phone number or undeliverable mail); only 285 (9.5%) actively refused participation, and another 525 (17.5%) passively refused by non-response. Caregivers are currently the best source of information for assessing the quality of care coordination services being provided to CMC. We attempted to compare caregiver report to medical records data for a subset of the FECC quality measures for which such comparison would be relevant. We found that very few medical records (paper or electronic) contained the necessary information to assess eligibility and scoring for this subset of FECC care coordination quality measures. For example, among respondents with medical records data available, 39% of parents reported having a shared care plan, while such a plan was identified in 2% of their children's medical charts.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The FECC Survey was completed by 1209 parents of CMC in the states of Washington and Minnesota during field testing in 2013. In the context of the field testing and validation study, patients and families were identified from Medicaid enrollment data. The surveys were administered by the RAND Corporation Survey Research Group (RAND SRG), Santa Monica, CA, and included children served by a range of pediatric practice types, including small group, multi-specialty, urban, and rural practices. The average number of participating families per identified provider was 2.5, while the median was 1. The maximum number of participating families per

provider was 26. Given the low average and median number of eligible CMC per provider, we determined that it would be nearly impossible to make meaningful comparisons on the FECC measures at the provider level, and only possible to do so at the practice level for practices meeting the minimum sample sizes discussed above in S.20.

We achieved an overall survey response rate of 40% (1209 out of 3000), which was quite good given that 632 of the original 3000 (21%) were unable to be contacted (bad phone number or undeliverable mail); only 285 (9.5%) actively refused participation, and another 525 (17.5%) passively refused by non-response. In our field-testing, we randomized participants to either mixed mode (mailings followed by telephone contact) or telephone only arms. The response rate among those assigned to the mixed mode was 45.5% (7.3% refusal rate) and was 35.9% (10.3% refusal rate) among those assigned to telephone only mode. Compared to respondents randomized to telephone mode, mixed mode mail respondents (and their children) were significantly more likely to be non-Hispanic white and English proficient, while mixed mode telephone respondents (and their children) were more likely to be of a minority race/ethnicity and limited English proficient. We therefore recommend a mixed mode approach, given the higher overall response rate and the different approaches to maximize participation of a range of demographic groups.

During initial field testing, one measure (of the original 21) and 11 sub-parts were dropped from analysis and removed from the FECC Survey due to low eligibility and/or ceiling effects. For example, the initial shared care plan measure included four sub-parts, specifying that (a) a shared care plan was created; (b) the caregiver participated in creating it; (c) the caregiver participated in updating it within the last year, if it was first created >1 year ago; and (d) the caregiver received a copy of it. Given that less than half of respondents endorsed having a shared care plan, and that measure sub-parts (b), (c), and (d) exhibited both low eligibility and ceiling effects, only measure sub-part (a) was retained in the final survey.

We also determined that caregiver survey is the only way to identify the use of tools like shared care plans at the present time. We attempted to compare caregiver report to medical record abstraction for a subset of the FECC measures for which such comparison would be relevant. We found that very few medical records (paper or electronic) contained the necessary information to assess eligibility and scoring for this subset of FECC care coordination quality measures. For example, among respondents with medical record data available, 39% of parents reported having a shared care plan, while such a plan was identified in 2% of their children's medical charts.

Survey administration is expensive and time consuming; while it is currently the most valid approach for assessing care coordination quality for CMC, further work should investigate alternate modes of administration, including electronic survey data collection at the point of care using portable devices such as tablet computers.

The FECC survey quality measures are currently being used by a number of groups across the country (see below), but so far additional data related to feasibility and implementation are not available.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Not applicable

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement (Internal to the specific organization)

	Children's Healthcare of Atlanta unknown Boston Children's Hospital unknown Children's Hospital of Wisconsin unknown School of Nursing, University of Minnesota, Minneapolis, MN unknown Department pf Pediatric and Communicable Diseases, University of Michigan Hospital and Health Systems, Ann Arbor, MI unknown Oregon Health & Science University, Portland, OR unknown Meridian Health Plan unknown Health Resources and Services Administration, Maternal and Child Health Bureau, Washington, DC unknown James B. Fahner MD Pediatric Hospice Program, Hospice of Michigan, Ada, MI unknown Mathematica Policy Research, Inc., Ann Arbor, MI unknown National Research Corporation, Lincoln, NE unknown Cleveland Clinic Children's Hospital, Cleveland, OH unknown

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Below is a list of requested user and uses of the FECC Survey, along with date of request. We do not have information beyond the information presented here, including about the number of patients or accountable entities included

1. Javier Tejedor-Sojo, MD; Children's Healthcare of Atlanta, Atlanta, GA; 4/25/15; Tracking work with medically complex patients.

2. Eli Sprecher, MD, MPP, General Academic Pediatric Fellow; Boston Children's Hospital, Boston, MA; 5/1/15; Improving their internal process measures and patient experience measurement for their population of children with medical complexity.

3. John Gordon, MD, Medical Director, Special Need Program; Children's Hospital of Wisconsin Milwaukee, WI; 5/11/15; Adding FECC care coordination survey questions to current parent report questionnaires for medically complex patients.

4. Wendy Looman, PhD, APRN, CNP, Associate Professor; School of Nursing, University of Minnesota, Minneapolis, MN; 5/12/15; Tracking quality of care coordination for their newly developed complex care program

5. Katie Freundlich, MD, Clinical Instructor; Department of Pediatric and Communicable Diseases, University of Michigan Hospital and Health Systems, Ann Arbor, MI; 5/12/15; Currently conceptualizing a hospital-based complex care program.

6. Colleen Peck Reuland, MS, Director – Oregon Pediatric Improvement Partnership; Oregon Health & Science University, Portland, OR; 5/12/15; Assessing care coordination for medically complex children in their network of practices

7. Elzbieta Rozmiej, MD, Medical Director; Meridian Health Plan, Michigan; 5/12/15; Assessing quality of care coordination services for their pediatric clients with medical complexity

8. Marie Mann, MD, MPH; Division of Services for Children with Special Needs, Health Resources and Services Administration, Maternal and Child Health Bureau, Washington, DC; 5/12/15; Sharing with state integration/care coordination grantees for implementation into their work.

9. Mary Spicketts, MSN, RN, CHPN, CHPPN, Director, Pediatric Program; James B. Fahner MD Pediatric Hospice Program, Hospice of Michigan, Ada, MI; 5/12/15; Potential use as grant-writing/Research opportunity within pediatric hospice/palliative care and CMC patient populations.

10. Joe Zickafoose, MD, MS, Health Researcher; Mathematica Policy Research, Inc., Ann Arbor, MI; 5/12/15; No reason given.

11. Sarah Fryda, MS, Senior Research Associate; National Research Corporation, Lincoln, NE; 5/15/15; Interested in amending their current child HCAHPS survey to incorporate FECC.

12. Skyler Kalady, MD, Medical Director, Pediatric Complex Care Clinic; Cleveland Clinic Children's Hospital, Cleveland, OH; 5/18/15; Currently crafting metrics for the Pediatric Complex Care Clinic.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The FECC quality measure set is not currently used in public reporting due to its relatively recent development. There are no policies or other restrictions in place preventing more wide-spread use.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
The FECC Survey measures are being widely distributed to complex care programs across the country, with rapid uptake and application to program evaluation and quality improvement efforts. We expect some of these efforts to be publicly reported in the future. We also know of at least one Centers for Medicare and Medicaid Innovation (CMMI) grant that is using the FECC Survey measures to evaluate the impact of different approaches to care coordination for children with medical complexity; we expect the results of this Pediatric Partners in Care program to be publicly reported within the next 5 years. In addition, the Advancing Care for Exceptional Kids Act (ACE Kids Act; Senate bill 298; House bill 546), if it is approved, may use the FECC Survey measures to document current state and track improvements in care coordination for children with medical complexity, which would also lead to public reporting on a large scale.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

not applicable

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Measurement and reporting on processes related to care coordination for children with medical complexity would be expected to drive improvements in those processes, which, based on our evidence reviews, would in turn be expected to improve patient outcomes. While the population of CMC is small compared to the population of children overall, they consume a great deal of resources and require more services than most children, putting them at increased risk for failing to receive all needed care.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of**

unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.
No negative consequences or unintended effects have occurred to our knowledge as a result of FECC quality measure implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0009 : CAHPS Health Plan Survey v 3.0 children with chronic conditions supplement
0718 : Children Who Had Problems Obtaining Referrals When Needed
0719 : Children Who Receive Effective Care Coordination of Healthcare Services When Needed

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The currently available NQF-endorsed measures related to care coordination and care for children with chronic conditions are related to, but fundamentally different from, the quality measures addressed in the FECC measure set. To begin with, the measures differ with regard to target population. The currently-endorsed measures address children with chronic conditions (0009), children who have received a referral to specialty services (0718), and children who received care from at least 2 types of health care services (0719), while the FECC measures address children with medical complexity. While the other measures likely apply to CMC (in addition to many other children), the FECC measures are specific to CMC. In addition, the FECC measure set differs from currently-endorsed measures with regard to focus. The currently-available measures mostly focus on whether families who needed specialized services for their child found it easy or difficult to obtain them and whether anyone in their health plan or child's doctor's office/clinic helped them to get that service. In contrast, the FECC measure set focuses more on the quality of services provided by a family's self-identified care coordinator, delving into the specific care coordination attributes and processes that have been associated with better outcomes in the literature. For example, the measures regarding care coordination for children with chronic conditions (0009) ask about whether a particular child needed a given type of services, how difficult they were for the family to obtain, and if anyone helped them, which provides valuable information about the family experience and whether they received help. While there is some overlap between those types of measures and some of the measures within the FECC measure set (for example, FECC 3: care coordinator helped to obtain needed community services), those questions within the FECC measure set are predicated upon having a designated care coordinator (a care structure we found to be important for CMC based on the literature), and are assessing the functioning of that care coordinator, rather than just whether a service was provided to the family. The remaining measures within the FECC measure set are similarly focused on specific actions and attributes of the care coordinator and/or main medical provider, and would be expected to provide clearly actionable items for quality improvement intervention. For example, identifying that families are not receiving help with accessing recommended community services is important, but leaves open to interpretation why that may be; using the FECC measure set would help to separate out whether the problem was due to

not having a care coordinator, or whether it was due to having a care coordinator not adequately doing their job. In addition, the FECC measure set addresses other aspects of care coordination beyond the quality of services provided by the care coordinator, as they also assess quality of written communication between providers and families, and between providers and the child's school, along with the quality of care planning with the family. Therefore, the FECC measure set should be seen as complementary to, and enhancing the currently available measures.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Please see discussion above (5a.2) for a description of how the FECC measures complement, focus, and extend the information provided by the currently-endorsed measures.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** FECC_SURVEY_Telephone_Interview_Version.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Seattle Children´s Research Institute
Co.2 **Point of Contact:** Rita, Mangione-Smith, rita.mangione-smith@seattlechildrens.org, 206-884-8242-
Co.3 **Measure Developer if different from Measure Steward:** Seattle Children´s Research Institute
Co.4 **Point of Contact:** Rita, Mangione-Smith, rita.mangione-smith@seattlechildrens.org, 206-884-8242-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
WORK GROUP MEMBERS:

1. Rita Mangione-Smith, MD, MPH; Seattle Children's Research Institute/ University of Washington, Seattle, WA; Oversaw entire project (study PI), including literature reviews, measure development, Delphi panel, measure specification, field testing, and analysis.
2. K. Casey Lion, MD, MPH; Seattle Children's Research Institute/ University of Washington, Seattle, WA; Literature review, measure development, analytic team
3. Courtney Gidengil, MD, MPH; Boston Children's Hospital/ Harvard Medical School/ RAND Corporation, Boston, MA; Literature review, measure development, analytic team
4. Eric Schneider, MD, MSc; RAND Corporation, Boston, MA (now Commonwealth Fund); Provided oversight and participated in all aspects of measure development and testing
5. Elizabeth McGlynn, PhD; Center for Effectiveness and Safety Research, Kaiser Permanente, Pasadena, CA; Provided oversight and participated in all aspects of measure development and testing
6. Layla Parast, PhD; RAND Corporation, Santa Monica, CA; Biostatistician and analytic team lead
7. Q Burkhart, MS; RAND Corporation, Santa Monica, CA; Data analyst, analytic team
8. Marc Elliott, PhD; RAND Corporation, Santa Monica, CA; Biostatistician and analytic team
9. Kimberly Arthur, MPH; Seattle Children's Research Institute, Seattle, WA; Literature review and measure development
10. Julie A. Brown; RAND Corporation, Santa Monica, CA; Survey design and data collection
11. Adam Carle, MA, PhD; Cincinnati Children's Hospital Medical Center, Cincinnati, OH; Measure development
12. Laurie Cawthon, MD, MPH; WA State Department of Social and Health Services, Olympia, WA; Field testing, data acquisition and

analysis
13. Carol Roth, RN, MPH; RAND Corporation, Santa Monica, CA; Quality measure operationalization and survey development
14. Justine Nelson, PhD; Minnesota State Medicaid, Minneapolis, MN;Field testing, data acquisition and analysis
15. Laura Richardson, MD, MPH; Seattle Children's Research Institute/ University of Washington, Seattle, WA; Literature review, measure development
16. Trina Colburn, PhD; Seattle Children's Research Institute, Seattle, WA; Literature review, measure development
17. Jean Popalisky, DNP, RN; Seattle Children's Research Institute, Seattle, WA; Literature review, measure development
18. Maria Britto, MD, MPH; Cincinnati Children's Hospital Medical Center, Cincinnati, OH; Literature review, measure development

DELPHI PANEL MEMBERS:

1. Richard Antonelli, MD, MS
Medical Director of Integrated Care and Strategic Partnerships
Medical Director Physician Relations and Outreach
Boston Children's Hospital
Assistant Professor of Pediatrics
Harvard Medical School
Nominated by American Academy of Pediatrics (AAP)

2. Allison Ballantine, MD, MEd
Assistant Professor of Pediatrics
University of Pennsylvania School of Medicine
Section Chief of Education
Medical Director, Integrated Care Services
Division of General Pediatrics
Attending Physician Palliative Care Team
Attending Physician Inpatient General Pediatrics
The Children's Hospital of Philadelphia
Nominated by Society of Hospital Medicine (SHM)

3. Jennifer Bolden-Pitre, MA, JD
Director of Integrated Systems,
Statewide Parent Advocacy Network
Family Fellow, Leadership Education in Neurodevelopmental Disabilities
Children´s Hospital of Philadelphia
Nominated by Family Voices

4. Carol A. Ford, MD
Professor of Pediatrics
Orton Jackson Endowed Chair in Adolescent Medicine
University of Pennsylvania
Chief, Craig Dalsimer Division of Adolescent Medicine
The Children´s Hospital of Philadelphia
Nominated by Society for Adolescent Health & Medicine (SAHM)

5. Jason Kessler, MD, FAAP, CHBE
Medical Director
Iowa Medicaid Enterprise
Nominated by Medicaid Medical Directors Learning Network (MMDLN)

6. Karen Kuhlthau, PhD
Associate Professor, Pediatrics
Harvard Medical School
Associate Sociologist, Pediatrics
Center for Child and Adolescent Health Policy
Massachusetts General Hospital for Children

Nominated by Academic Pediatric Association (APA)

7. Dennis Kuo, MD, MHS
Assistant Professor of Health Policy and Management
Fay W. Boozman College of Public Health,
University of Arkansas for Medical Sciences
Assistant Professor of Pediatrics
Section on General Pediatrics
Center for Applied Research and Evaluation,
University of Arkansas for Medical Sciences
Pediatrician
Medical Home Program for Children with Special Needs,
Arkansas Children's Hospital
Nominated by Children's Hospital Association (CHA)

8. Wendy Sue Looman, PhD, RN, CNP
Pediatric Nurse Practitioner
Cleft Palate and Craniofacial Clinic
School of Dentistry, University of Minnesota
Associate Professor
School of Nursing, University of Minnesota
Nominated by National Association of Pediatric Nurse Practitioners (NAPNAP)

9. Karen Pierce, MD, FAPA, FAACAP
Attending Physician
Department of Child and Adolescent Psychiatry
Children's Memorial Hospital, Chicago, Illinois
Clinical Associate Professor
Feinberg School of Medicine, Northwestern University Medical School
Department of Psychiatry and Behavioral Sciences
Nominated by American Academy of Child & Adolescent Psychiatry (AACAP)

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2015

Ad.3 **Month and Year of most recent revision:** 12, 2014

Ad.4 **What is your frequency for review/update of this measure?** every 6 months

Ad.5 **When is the next scheduled review/update for this measure?** 03, 2016

Ad.6 **Copyright statement:**

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2789
De.2. **Measure Title:** Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care
Co.1.1. **Measure Steward:** Center of Excellence for Pediatric Quality Measurement
De.3. **Brief Description of Measure:** The Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care measures the quality of preparation for transition from pediatric-focused to adult-focused health care as reported in a survey completed by youth ages 16-17 years old with a chronic health condition. The ADAPT survey generates measures for each of the 3 domains: 1) Counseling on Transition Self-Management, 2) Counseling on Prescription Medication, and 3) Transfer Planning.
1b.1. **Developer Rationale:** IMPORTANCE OF MEASURING THE QUALITY OF TRANSITION FROM PEDIATRIC TO ADULT-FOCUSED HEALTH CARE
Health care transition (HCT) has been defined as a planned, purposeful process in which adolescents and young adults move from pediatric-focused health care delivery to adult-focused delivery.[1] The goal of HCT is to maximize lifelong functioning and potential through the provision of uninterrupted, high-quality, developmentally-appropriate health care services.[1] The lack of effective transition from pediatric to adult-focused health care may contribute to fragmentation of health care and increased risk for adverse health outcomes. Those at highest risk during this period include youth with special health care needs (YSHCN).[2]

The process of HCT involves 3 key phases: 1) transition planning and preparation; 2) transfer of health care to an adult-focused model; and 3) intake to the adult-focused health system. There is broad consensus that preparation for HCT should start in adolescence and involve individualized planning and ongoing skills development.[3]

In 2002, a consensus statement from the American Academy of Pediatrics, the American Academy of Family Physicians, and the American College of Physicians envisioned the goal that by 2010 "all physicians who provide primary or subspecialty care to young people with special health care needs 1) understand the rationale for transition from child-oriented to adult-oriented health care; 2) have the knowledge and skills to facilitate that process; and 3) know if, how, and when transfer of care is indicated."[1] For youth receiving care in pediatric-focused health care settings, preparation for HCT includes the acquisition of self-care skills and promotion of increased youth responsibility for chronic condition management. For many youth, transition preparation culminates in a transfer to a new health care setting. However, even for youth who do not change care settings (e.g., those in family medicine settings), the shift to adult-oriented health care still requires appropriate preparation. Because transition preparation is primarily a series of interactions with clinicians, obtaining reports directly from youth about their experience is critical to understanding current gaps in health care delivery for this population.

PREPARATION FOR HEALTH CARE TRANSITION: LACK OF STANDARDIZED QUALITY MEASUREMENT
In its 2011 Patient-Centered Medical Home Standards, the National Committee on Quality Assurance included a specific requirement to address care transitions in primary care.[4] The MCHB identified HCT services as a core

outcome for the community-based services required for CSHCN under Title V and Healthy People 2000 and reiterated this priority in the Healthy People 2010 and Healthy People 2020 goals.[1,5-6] However, systematic assessments of transition readiness are rarely incorporated as part of routine health care.[6] Measuring the quality of HCT preparation is intended to drive providers to adopt strategies that foster disease self-management among youth and reliably result in safe and effective transfer to adult care.[7]

DISPARITIES IN HEALTH CARE TRANSITION PREPARATION

Geographic, socioeconomic, racial and ethnic disparities have been documented in the receipt of HCT services.[8-10] In the 2005-2006 NS-CSHCN, fewer African-American and Latino respondents reported having discussed shifting their child's care to an adult-focused provider.[8] In the same survey, the proportion of respondents who met the core performance outcomes for successful transition increased significantly with increasing family income.[5] Additionally, the 2007 SATH revealed that low-income young adults had poorer access to health care than those with higher incomes.[8] Such disparities in transition preparation and access to care are likely to result in adverse health outcomes.

References:

1. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians-American Society of Internal Medicine. A consensus statement on health care transitions for young adults with special health care needs. Pediatrics. 2002;110(6 Pt 2):1304-1306.
2. Lotstein DS, McPherson M, Strickland B, Newacheck PW. Transition planning for youth with special health care needs: results from the National Survey of Children with Special Health Care Needs. Pediatrics. 2005;115(6):1562-1568. doi:10.1542/peds.2004-1262.
3. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians, Transitions Clinical Report Authoring Group, Cooley WC, Sagerman PJ. Supporting the health care transition from adolescence to adulthood in the medical home. Pediatrics. 2011;128(1):182-200. doi:10.1542/peds.2011-0969.
4. Standards for Patient-Centered Medical Home (PCMH), in Washington, DC: National Committee for Quality Assurance, Editor. 2011.
5. Lotstein DS, Ghandour R, Cash A, McGuire E, Strickland B, Newacheck P. Planning for health care transitions: results from the 2005-2006 National Survey of Children With Special Health Care Needs. Pediatrics. 2009;123(1):e145-152. doi:10.1542/peds.2008-1298.
6. McManus MA, Pollack LR, Cooley WC, McAllister JW, Lotstein D, Strickland B, Mann MY. Current status of transition preparation among youth with special needs in the United States. Pediatrics. 2013;131(6):1090-1097. doi:10.1542/peds.2012-3050.
7. Park MJ, Adams SH, Irwin CE. Health care services and the transition to young adulthood: challenges and opportunities. Acad Pediatr. 2011;11(2):115-122. doi:10.1016/j.acap.2010.11.010.
8. Lotstein DS, Kuo AA, Strickland B, Tait F. The transition to adult health care for youth with special health care needs: do racial and ethnic disparities exist? Pediatrics. 2010;126 Suppl 3:S129-136. doi:10.1542/peds.2010-1466F.
9. Richmond N, Tran T, Berry S. Receipt of transition services within a medical home: do racial and geographic disparities exist? Matern Child Health J. 2011;15(6):742-752. doi:10.1007/s10995-010-0635-2.
10. Kane DJ, Kasehagen L, Punyko J, Carle AC, Penziner A, Thorson S. What factors are associated with state performance on provision of transition services to CSHCN? Pediatrics. 2009;124 Suppl 4:S375-383. doi:10.1542/peds.2009-1255H.

Numerator Statement: The ADAPT survey consists of 26 questions assessing the quality of health care transition preparation for youth with chronic health conditions, based on youth report of whether specific recommended processes of care were received. The ADAPT survey generates measures for each of 3 domains: 1) Counseling on Transition Self-Management, 2) Counseling on Prescription Medication, and 3) Transfer Planning. ADAPT measure scores are calculated using the sum of the proportions of positive responses to between 3 and 5 individual items. Complete instructions for measure score calculations are provided in the Detailed Measure Specifications (Appendix A).

1) Counseling on Transition Self-Management:

The numerator is the sum of the proportions of positive responses to the five questions about counseling on transition self-management, among respondents with valid responses to all questions.

2) Counseling on prescription medication:
The numerator is the sum of the proportions of positive responses to the three questions about counseling on prescription medication, among respondents who indicate that they take prescription medication every day and with valid responses to all questions.

3) Transfer planning:
The numerator is the sum of the proportions of positive responses to the four questions about transfer planning, among respondents who report being treated by a pediatric provider and with valid responses to all questions.
S.7. **Denominator Statement:** The target population of the survey is 16- or 17-year-old adolescents with a chronic health condition who are either (a) receiving health care services in a clinical program or (b) enrolled in a health plan or similar defined population.

The denominator for each measure is the number of respondents with valid responses for all of the questions in the measure.
S.10. **Denominator Exclusions:** SURVEY SAMPLE
Exclude patients in the following categories from the ADAPT survey sample frame:

1. "No-publicity" patients (i.e., those who requested that they not be contacted)

2. Court/law enforcement patients

3. Patients with a foreign home address

4. Patients who cannot be surveyed because of local, state, or federal regulations

SURVEY RESPONSE
Exclude survey respondents based on the following clinical and non-clinical criteria:

1. Undeliverable survey, i.e., the survey is returned by US Mail as undeliverable. "Undeliverable" should not be assumed merely because of non-response.

2. The survey is returned with clear indication that the patient does not meet eligibility criteria (e.g., ineligible age or lack of a chronic health condition).

3. Patient unable to complete survey independently: This must be indicated by the appropriate checkbox in the cover letter or equivalent clear indication by the parent/guardian that the patient is unable to complete the survey independently (e.g., due to cognitive limitation).

4. Exclude all respondents who answered "None" to ADAPT question 3 ("In the last 12 months, how many times did you visit this provider?").

Measure Type: PRO
S.23. **Data Source:** Patient Reported Data/Survey
S.26. **Level of Analysis:** Clinician : Group/Practice, Facility, Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: Not applicable

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measures evaluation and voting processes. The preliminary analysis will help to guide the Standing Committee evaluation of each measure by summarizing the measure developer submission, guide measure evaluation discussion, and identify topic areas for additional input. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if health outcomes measures agree the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

The developer provides the following rationale for this patient-reported outcome measure:

- The Level of Analysis for this measure is Clinician: Group/Practice; Facility; Health Plan.
- Health care transition (HCT) is a core aspect of healthcare quality for youth with special health care needs (YSHCN) and a major focus for quality improvement (Lotstein et al. 2005). The goal of HCT is to maximize lifelong functioning and potential through the provision of uninterrupted, high-quality, developmentally appropriate health care services (AAP et al. 2002). The lack of effective transition from pediatric to adult-focused health care can contribute to fragmented or delayed care and increased risk for adverse health outcomes. Improving transition preparation for at-risk youth may decrease costs associated with inappropriate or delayed healthcare utilization.
- Research and consensus agree effective care transitions are important, but limited data exist that link adequate transition preparation and readiness (the desired goal of preparation) with improved adult health outcomes. A small number of quasi-experimental studies, all conducted in patients with type 1 diabetes, have shown that transition preparation interventions were associated with improved frequency of post-transition medical follow-up (Holmes-Walker et al. 2007; Cadario et al. 2009; Van Walleghem et al. 2008), reduced acute diabetes complications and improved hemoglobin A1c levels. While limited, available data suggest that the lack of effective transition from pediatric- to adult-focused care may contribute to fragmentation of young adult health care and increased risk for adverse outcomes.
- Consensus recommendations for transition preparation identify 14-15 years as the ideal age to initiate the development of a patient-specific transition plan (AAP et al. 2011). Querying patients at 16-17 years captures them at a time by which some transition preparation generally should have occurred.
- Because this is a PRO-PM, evidence that the target population values the measured PRO and finds it useful should be provided.
    - The developer states that 11 focus groups in Boston, Chicago, and LA were conducted with adolescent (age 16-18 years) and young adult (age 19-26 years) patients with one or more chronic health conditions, as well as parents/guardians of youth or young adults with chronic health conditions. One young adult group and two parent groups were in Spanish.
        - Focus group findings were synthesized with research and expert interviews and the developers then conducted cognitive interviews with adoslecents (age not provided in the submission) with chronic health conditions to ensure they understood the survey questions as intended. The developer performed four rounds of 26 total cognitive interviews of youth respondents in English and Spanish in Boston, Chicago, and Dallas.
        - Questions were refined based on the focus groups and cognitive interviews to ensure the

survey was useable and useful, but the developer does not expand on its findings specifically on the usefulness/value to the target population.

- Per the **NQF Algorithm for Evidence**, the Committee should assess this on a Pass/No Pass basis (box 1—>)

Questions for the Committee

- *Is the relationship of this measure to patient outcomes presented reasonable?*
- *The measure specifies a target survey population of 16-17 years; is this reasonable?*
- *Is there at least one thing that the provider can do to achieve a change in measure results?*
- *Does the Committee wish to discuss with the developer whether it specifically asked about the value/usefulness of the survey during its focus groups and/or interviews?*
- *The measure is specified for a population of 16-17 years, but the focus groups were adolescents (16-18 years) and young adult (age 19-26 years). Does the Committee wish to discuss further any difference in value/usefulness of the measure's specified population vs. the broader focus group/interview population?*

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The Maternal and Child Health Bureau (MCHB) identified HCT services as a core outcome for the community-based services required for Children with Special Health Care Needs (CSHCN) under Title V and *Healthy People 2000* and reiterated this priority in the *Healthy People 2010* and *Healthy People 2020* goals.
- In the 2005-2006 National Survey of Children with Special Health Care Needs (NS-CSHCN), fewer African-American and Latino respondents reported having discussed shifting their child's care to an adult-focused provider. In the same survey, the proportion of respondents who met the core performance outcomes for successful transition increased significantly with increasing family income. Additionally, the 2007 Survey of Adolescent Transition and Health (SATH) revealed that low-income young adults had poorer access to health care than those with higher incomes.
- NS-CSHCN and SATH both report poor performance around HCT and much room for improvement; there are no data on transition preparation from the adolescents themselves. The developer states that the ADAPT survey addresses this information gap and will provide information for benchmarking and improving care.
- The developer reports that despite different geographic regions and demographic characteristics, the scores and responses were similar across all three sites based on its 2013-2014 testing.
- Higher score = better quality.
- The sample size of the field test for the measure was not designed to provide statistical power to detect differences between racial/ethnic groups or differences between patients with non-complex and complex chronic diseases.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b)

1a. Evidence

- I think this is conceptually compelling but with not great evidence that counseling really influences behavior.
- The age may be an issue as now many hospitals use 26 at the age of transition (ACA response).
- Concern about the cognitively disabled.
- This is clearly a patient outcome measure. The survey assesses readiness for transition and resources available and is well represented using survey data.
- 16-17 seems reasonable though my hospital cares for kids usually till 18 and in some cases beyond.
- The items are very behavior oriented and each has the ability to provide an area where behavior can concretely change if needed.
- Providing additional information on usefulness of survey to patients and parents would be helpful however I think said interviews should have been done with clinicians to determine how useful they thought the survey would be for quality improvement.
- No issues with the age of the focus group participants.
- No evidence provided why the three domains they chose for this survey are appropriate.
- There appears to be enough evidence to support the survey tool/measure. Transition of children with chronic conditions from pediatric to adult care is an important aspect of care for this age range. Did the developer consider expanding the denominator to the upper age ranges as some children/young adults stay with their pediatrician up to age 21? Providers can improve how they prepare children/young adults to transition them to adult care as they take more responsibility for their own condition/care.
- In general I was pleased to see mention of avoiding adverse outcomes. Unfortunately, these can occur during transition after successful treatment in a pediatric program (e.g. renal transplant.) Although some of the evidence cited studies on specific conditions 9e.g. diabetes,) it should be noted that some diseases are progressive despite best efforts. In addition, some with a chronic condition could have a second comorbid condition and it is initially unclear if this was the case with the sample population. Our organization strongly supports transition from pediatric to adult healthcare under the MCH (Maternal/Child Health) six core outcomes. I also strongly supported this in our comments on both Healthy People 2010 and Healthy People 2020. I was pleased to see the mention of disparities as it relates to transition and addressing underserved populations. I understand that the level of analysis could be at the provider/group, facility, or plan level. I support the concept to "maximize lifelong functioning" using effective transition from pediatric to adult care. I also support the age range of 14-15 (as did the American Academy of Pediatrics {AAP});, indeed our organization successfully advocated to have the age to begin transition from special education to adult life begin at age 14, while the federal level is age 16. Regarding questions for the committee, the relationship to outcomes appears "reasonable." I support the eligibility age for the measure as 16 since then "some transition preparation generally should have occurred." Just having providers administer the survey can help raise awareness of the provider role in easing the transition from pediatric to adult care. The committee should ask if the focus groups were asked about the "value/usefulness" of the survey. The committee could discuss the difference with the specified population age of 16-17 and the focus group of young adults ages 19-26.
- Overall there is consensus concepts that transition to adult care should occur and may lead to better outcomes. Where evidence is lacking is that the counseling prescribed in this survey actually accomplishes the goal of transition readiness. This is key issue as perhaps other modalities may be more appropriate such as a transition coordinator or simulations as opposed to the approach suggested in the survey that provider counseling should accomplish transition readiness.
- While the concept is clearly important, perhaps first step should be research into best methods to

achieve transition readiness and then design survey to assess implementation of that methodology. The measure developers provide no evidence that provider counseling is effective tool for ensuring transition readiness.

1b. Performance Gap

- How is this disparities sensitive?
- There is evidence that transitions need to be improved.
- Evidence as provided does warrant a measure to assess care transition.
- The evidence provided indicates that this measure is disparities sensitive in terms of the results but not necessarily in terms of the ability of respondents to understand and complete the survey.
- There does appear to be a gap. It appears that they measures could be disparities sensitive, if the testing was designed to do so.
- As mentioned above, our organization strongly supported transition from pediatric to adult care for both the MCH six core outcomes and in our Healthy People 2010/2020 testimonies. We are aware of the results of the National Survey of Children with Special Health Care Needs regarding underserved populations related to transition and have specifically targeted these diverse families, translating all of our materials into Spanish at a minimum, or various languages due to our multicultural staff. I was pleased to see the similarity found in the responses despite differences geographically. Regarding the questions for the committee, there is definitely a gap in care. I support this measure as "disparities sensitive" and would highly recommend this expanded to people with disabilities. As with measure 2770, individuals with developmental disabilities are more likely to have health disparities (see http://www.sciencedirect.com/science/article/pii/S1936657410000373.)
- Minimal data on performance given with no data comparing how adolescents believed to be well prepared for transition score compared to adolescents who are not well prepared. This means we are unsure if this survey actually detects transition readiness. Therefore cannot comment if this data shows that a gap in care exists. The survey scores were nearly identical across 3 different geographic regions.
- Data provided does not suggest any sensitivity to detect disparities in care.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. The developer states that the measure contains three domains. Scores are reported by domain and not a single, overall score, however it is reported as one measure. • The numerators are: o 1) **Counseling on Transition Self-Management**: The numerator is the sum of the proportions of positive responses to the five questions about counseling on transition self-management, among respondents with valid responses to all questions. o 2) **Counseling on Prescription Medication**: The numerator is the sum of the proportions of positive responses to the three questions about counseling on prescription medication, among respondents who indicate that they take prescription medication every day and with valid responses to all questions.

- 3) **Transfer Planning**: The numerator is the sum of the proportions of positive responses to the four questions about transfer planning, among respondents who report being treated by a pediatric provider and with valid responses to all questions.

- The denominator is: The target population of the survey is 16- or 17-year-old adolescents with a chronic health condition who are either (a) receiving health care services in a clinical program, or (b) enrolled in a health plan or similar defined population. The denominator for each measure is the number of respondents with valid responses for all of the questions in the measure.
- The numerator details explain which survey questions map to which numerator. The denominator details explain who is eligible for the survey and how to score the responses.
- The calculation algorithm appears to be clearly specified.
- The measure can be stratified but it is not required.
- The measure is risk-adjusted using case mix adjustment by age and self-reported health status.
- The measure specifies three domains within a single measure.

The developer reports the following related to survey administration (which goes to both reliability and validity testing, below):

- For eligibility for the Level of Analysis at the group practice level, the developer indicates options based on the "goals for quality measurement," with the use of patient registries, EHRs, or patient panels as options to determine eligibility.
- For a health plan or entity with access to claims data, the developer notes that eligibility can be determined by using the Pediatric Medical Complexity Algorithm.
- The developer indicates the eligible population should "generally" be included in the sample frame, except for the following:
  - Patients who request that they not be contacted
  - Court/law enforcement involved patients (i.e., prisoners); this category does not include those residing in halfway houses
  - Patients with a foreign home address (the US territories – American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and Virgin Islands – are not considered foreign addresses and therefore are not excluded)
  - Patients who cannot be surveyed because of local, state, or federal regulations
- The survey administration mode specifies only mail.

Questions for the Committee

- *Are all ALL the questions for all three measures (Self-Management, Medication, Transfer Planning) clear, unambiguous, and at an appropriate comprehension level? clearly defined?*
- *Is the logic or calculation algorithm clear?*
- *Is the sampling methodology clear? The developer provides options for creating the underlying denominator population. Given the purpose of the measure is standardization for comparative purposes of accountability, does the Committee wish to discuss with the developer whether the sampling should be standardized (either within or among the Levels of Analysis—i.e., be more prescriptive?)*
- *Is it likely this measure can be consistently implemented based on the numerator details, denominator details, algorithm, **and** survey administration details?*

2a2. Reliability Testing **Testing attachment**

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer provides the following information:

- This measure was tested at the critical data element level and the performance measure score level, but data appear only to be provided at the performance measure score level.
- The measure was tested in one hospital and two health plans serving Medicaid enrollees.
- The response rate by setting varies from 21% and 28% for the health plans and 47% for the clinical programs. The initial deployment for the health plans was 1,500 surveys and 623 for the clinical programs.
- Internal consistency reliability tested with ordinal alpha was provided for each of the three measures at each of the three test sites. Results generally ranging from .74-.99 with one exception at one site (.57). These results generally indicate good to excellent reliability. The transfer planning measure had the highest score at 0.99 at each site. Counseling on transition ranged from 0.70 to 0.79. Counseling on prescriptions received 0.57, 0.74, and 0.78.
- **Per the NQF Algorithm for Reliability,** empirical testing at the performance score level is eligible for a rating of HIGH, MODERATE, or LOW, depending on the results.
- The developer concludes that its results indicate good to excellent internal consistency reliability.

Questions for the Committee

- *Is the methodology to test reliability appropriate?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified for the proposed Levels of Analysis of Clinician: Group, Facility, and Health Plan?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The goal of the measure is to assess the quality of preparation for transition from pediatric-focused to adult-focused health care for youth ages 16-17 years old with a chronic health condition.
- The three numerators are *counseling on transition self-management*, *counseling on prescription medication*, and *transfer planning*. The denominator for each measure is the number of respondents to the survey with valid responses for all of the questions in the measure. The target population for the survey is 16- or 17-year-old adolescents with a chronic health condition who are either (a) receiving health care services in a clinical program or (b) enrolled in a health plan or similar defined population.
- Since this is a patient-reported outcome measure (PRO-PM), the evidence should support the relationship of the health outcome to at least one clinical action. The evidence provided by the developer note that research and consensus agree effective care transitions are important, and that available data, while limited, suggest that the lack of effective transition from pediatric- to adult-focused care may contribute to fragmentation of young adult health care and increased risk for adverse outcomes. Expert consensus indicates 14-15 years as the ideal age to initiate the development of a patient-specific transition plan, so by ages 16-17 some planning should have started.

Question for the Committee:

- *Are the specifications consistent with the evidence for each domain?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer provides the following information:

- Empirical validity testing for the performance measure score was performed.
- **Focus groups and cognitive interviews** were conducted to test content validity and to confirm that each question was understandable.
  - The focus groups and interviews generally confirmed that the questions were understandable for their intended meaning, construction, etc.
  - The developers made small revisions in areas that were not as clear and then did additional testing.
- **Confirmatory factor analysis** for the two counseling measures was performed; it could not be performed for the transfer planning measure due to small sample size.
  - Results from these analyses supported the hypothesis that the individual questions within each of the two domains are associated with one another.
  - CFA results were similar across the three sites, and the developer states that this provides further confirmation that questions grouped together on conceptual grounds also are empirically related.
- The developer states that the values of the loading factor estimates within each measure demonstrate that questions are strongly associated with their hypothesized construct. In addition, the association between the two constructs in all three sites is also significant.
- Per the **NQF Algorithm for Validity**, empirical testing for validity at the performance measure score level (boxes 6—>8) is eligible for a rating of HIGH, MODERATE, or LOW.

Questions for the Committee

- *Is the validity testing methodology appropriate?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score for each domain for this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Some populations are excluded during the sampling frame. The following populations are excluded:
  - Patients who request that they not be contacted
  - Court/law enforcement involved patients (i.e., prisoners); this category does not include those residing in halfway houses
  - Patients with a foreign home address (the US territories – American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and Virgin Islands – are not considered foreign addresses and therefore are not excluded)
  - Patients who cannot be surveyed because of local, state, or federal regulations

Questions for the Committee

- *Are the exclusions for the sampling frame appropriate?*

2b4. Risk adjustment:

- The developer reports risk adjustment/case mix for self-reported health status and age.
- The developer assessed variation by education and gender; no variation was found so these were not included in the final risk adjustment model. The developer reports it found variation based on medical complexity and the patient's county of residence.

Questions for the Committee (as appropriate):

- o *Is risk-adjustment strategy appropriate?*
- o *Ares the final variables adequately described for the measure to be implemented?*
- o *Does the Committee wish to discuss the risk adjustment for medical complexity with the developer?*

2b5. Meaningful difference:

The developer provides the following information:

- Differences in population-level scores were compared using t-tests and f-tests based on the case-mix adjustment model estimates. Statistically significant differences in performance were assessed for the case-mix adjusted scores for measures by examining whether the three sites were different. A t-test of means was used for comparing between the two health plan sites. A n f-test of means was used for comparing across the three sites. A level of alpha error of $p < .05$ was set as the criterion for significance.
  - o The developer concludes that the overall scores for all three measures were low in all three populations, which is consistent with national findings on transition readiness.
  - o The developer concludes that the small differences currently seen in the Counseling on Transition Self-Management measure are not clinically meaningful given the overall low performance across all three sites.

Question for the Committee:

- o *Does this measure identify meaningful differences about quality?*

2b7. Missing Data

The developer provides the following information:

- Since the survey uses skip patterns, there is a high percentage of appropriately missing data. All three sites had less than 3% truly missing cases, suggesting results are unlikely to be biased by systemic missing data and there does not appear to be a systemic biase in response on demographic factors.
- The developer also notes there does not appear to be a systematic bias in response to the survey overall based on demograpnic factors of age, race/ethnicity, gender, or chronic condition complexity.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- This may be an area of concern for cognitively impaired individuals and how they are accounted for. Overall logical and straightforward.
- Evidence is consistent with demonstrated need for the measure. NO real specifics on why the need is split by the three domains chosen.
- Questions and logic appear to be clear. If this were to be used for accountability purposes, there should be more details on sampling/eligibility for the survey/measures. Mail only administration will limit survey response. Was there any thought to evaluating point of care survey administration?

- Self-management, medication adherence, and transfer planning are all key to successful transition. However, it was unclear until reading Appendix A that the transfer plan was in written form. Regarding the numerator questions, for self-management there may be instances in which a legal guardian cannot leave the room but the provider could still direct the questions to the youth, enhancing transition, even if that child will never live independently. Under the prescription questions, there needs to be education regarding side effects or drug interactions (e.g. if 2 medications need to be taken hours apart from each other.) In addition, there should be discussion if the child has private and public insurance to ensure that the secondary is billed after the primary, avoiding unnecessary copays. The denominator questions appear appropriate, although there should be allowances for caregiver assistance for those with developmental disabilities. It is confusing that the caregiver can put that the youth can't answer the survey, yet questions 25 and 26 allow for this per Appendix C in English and D in Spanish respectively. As with measure 2770, there is a high correlation (see http://www.sciencedirect.com/science/article/pii/S1936657410000373) between developmental disabilities and secondary comorbid conditions so this could affect a sizable portion of the sample population. Other than the concerns mentioned above, the calculation algorithm appears appropriate. I would support the use of "registries, EHRs, or patient panels to determine eligibility" and for plans to use claims data. I agree with the exclusions due to requests of non-contact, court/law, foreign address, and local/state/national prohibitions. Regarding the questions for the committee, it appears that the questions are clear as is the calculation algorithm. The sampling is also clear. This could be "consistently implemented" if the aforementioned consideration of legal guardians is permissible.
- No evidence or discussion as to whether domains of self management, medications and planning capture the key constructs for transition readiness. Given focus in on adolescents with chronic health conditions, this survey and domains do not address key concept of care coordination among multiple providers (e.g. PCP and specialists). This seems a key area for transition readiness that is not captured in measures.

2a2. Reliability testing

- I am concerned that while the measure was tested at the "critical data element level," that only the performance measure score level appears. I am also concerned about the low response rate, particularly for the health plans at 21% and 28%. It was reassuring that there was "good to excellent reliability." Regarding the questions for the committee, there needs to be clarification on the critical data element score. As mentioned above, there is concern with the test sample due to low response rates from health plans. It is uncertain if there is sufficient reliability without the information on the critical data element level score.

2b1. Validity Specifications

- Appear appropriate given the available data. Repeat testing could be performed.
- Methodology for internal consistency reliability testing was appropriate. Might have done some test/retest on a small sample.
- Test sample was adequate
- Scores indicate room for improvement and sufficient variability to identify differences.
- Specifications appear to align with the available evidence.
- I understand that the "measure specifies three domains" (self-management, prescription medication, and transfer planning) within a single measure as stated in the previous section on reliability. If the goal is the "quality of preparation," there needs to be more education on prescriptions as mentioned above. In addition, there should be encouragement of self-advocacy. I strongly support that this measure is a "patient-reported outcome." Regarding the questions for the committee, other than the concerns

mentioned above (prescription education and guardianship issues,) the specifications are consistent.
- No repeated testing in same population was performed to be able to assess reliability. Testing was done in 3 geographically dispersed areas but no information provided as to how similar these test populations were or were not.

2b2. Validity Testing
- No issues
- The methodology to show that composite items group appropriately is appropriate and sample size for transfer planning measure could be an issue since the average rate was so low.
- Content validity appropriately assessed by interview but would have liked some expert opinions.
- No criterion related validity assessment was done though based on the evidence table on page 117 there are plenty of measures they could have used.
- Validity testing appears t be sufficient.
- It was reassuring that "empirical validity testing" was done. I understand that both "focus groups and cognitive interviews" were completed but would like more information on the revisions to enhance understandability. It is a concern that the confirmatory factor analysis could not be conducted on the transfer planning measure due to small sample size. It does appear that the "two domains are associated with one another." Regarding questions for the committee, it appears that the methodology was appropriate with the exception of the transfer planning measure. It seems that there is sufficient validity for the other two measures. The scores, with the exception of transfer planning due to small sample size, are indicative of quality.
- Validity testing focused on factor analysis as to whether domain group questions measure conceptually similar areas and whether questions understandable. No data or testing done to address whether target population values the concepts captured by the survey as meaningful.

2b3-2b7. Threats to Validity
- Appears to be appropriate
- No issues with exclusions
- Risk adjustment strategy is appropriate especially if you want to compare across health plans and for norming.
- Final variables for risk adjustment seem to be adequate based on the testing done but there will always be users who believe other variables should be included.
- Unclear whether age was put into categories or actual age used in the risk adjustment.
- Committee discussion with the developer is worth it especially if committee is concerned that some variables were not considered for risk adjustment that potentially should be.
- The measure allows for improvement and the ability to assess meaningful differences
- Missing data does not seem to be an issue based on piloted response data though not sure how composites are treated if items are missing.
- Exclusions and risk adjustment appear appropriate.
- Overall results appears to be low, suggesting room for improvement.
    - For 2b3., as stated above the exclusions (by request, court/law, foreign address, local/state/federal regulation) appear appropriate. Regarding the questions for the committee, these exclusions are appropriate.
    - For 2b4., it is understood that there was no variation due to education or gender which were then excluded from the final risk adjustment. It is a concern that language was excluded due to low response. It was unfortunate that there were variations based on medical complexity and geographically. Regarding the questions for the committee, it appears that the risk-adjustment

is appropriate. Clarification is needed on the final variables as it is unclear if the variation was due to medical complexity alone, geographic location alone, or a combination of the two so the committee should discuss the risk adjustment. It would also be worthwhile to know if there are disparities based on a specific condition.

- For 2b5., it is understood that the low scores are consistent with national data. It would appear that the slight differences in self-management due to low scores are "not clinically meaningful" due to low scores "across all three sites." Regarding questions for the committee, it appears that this measure does indicate differences in quality.
- 2b6. n/a
- For 2b7., it appears that since all three sites had less than 3% missing data, that this doesn't represent bias by systemic missing data, despite the use of skip patterns. It also seems that there is no bias based on demographic data.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer provides the following information:

- The ADAPT survey is administered by mail. The rationale for not using electronic sources (e.g., administration by email) is that mail and telephone administration are the best ways to obtain representative samples of patients based on the contact information (mailing address and telephone number) that is most often available for sampling and data collection.
- The survey takes approximiately 10 minutes to complete and is free of charge for users. The developer provides some tips for sending the survey.

Questions for the Committee (as appropriate):

- *Is the data collection strategy ready to be put into operational use?*
- *Does the Committee wish to discuss with the developer the feasibility or advisability of telephone administration?*
- *Does the Committee wish to further discuss with the developer the feasibility or advisability of electronic information, given the target age population?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Short survey could be delivered in multiple settings.
- The data collection strategy is available to put into operational use
- Since they suggest providing a toll free number they could also administer on the phone if someone calls with a question (unless they think there is bias between phone and mail)
- So I am not sure why, when they mail the survey they can't also provide a web link to complete it online as an option. Would save data entry time and errors.
- Would like additional discussion around other modes to collect the survey, outside of direct mail.
- Although this was administered by mail, telephone and electronic methods are also successful. This is a

relatively short survey so other means could be used. Regarding questions for the committee, it appears that the "data collection strategy" is ready for use but the committee should discuss the feasibility of utilizing telephone administration at the very least, if not electronic means.

- Developer reports focuses on feasibility of survey administration but doe not address feasibility of first identifying eligible denominator pool of adolescents. Denominator criteria are cumbersome and many and may hamper feasibility.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

The developer provides the following information.

- This is a newly developed measure that is not currently in use. The developer intends it to be available for public use.
- Planned use includes internal QI and QI with benchmarking but the specifics are not included.
    - *The NQF usability and use criteria states that developers should "provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement."*

Questions for the Committee

- *Will performance results further the goal of high-quality, efficient healthcare?*
- *The developer indicates use of the measure for internal QI and QI with benchmarking. Is the measure appropriate for accountability purposes?*
    - *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Perfomance results could further the goal of high-quality efficient healthcare. Improvements in this measure could also result in more efficient care
- The measure is appropriate for accountability purposes.
- Unintended consequences may be that the parent/patient realizes they are totally unprepared to transition into adult care but at least there will be no surprises and it will induce dialogue with their primary provider.
- Does the developer have a specific plan outlined to move this measure toward use for accountability?
- Measures not currently in uses so no data provided. Overall survey addresses a key concept but does not address whether provider counseling is best strategy for transition readiness nor where all key aspects of transition readiness are captured in the 3 domains.

Criterion 5: Related and Competing Measures

Related to 0005 : CAHPS Clinician & Group Surveys (CG-CAHPS)-Adult, Child

- Not completely harmonized. The developer indicates that CG-CAHPS is intended to be completed by parents and ADAPT is intended to be completed by adolescents. The developer states that "the

ADAPT survey complements the CG CAHPS survey well and has the potential to be administered concurrently."

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☒Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Health care transition (HCT) is a core aspect of health care quality for youth with special health care needs (YSHCN) and a major focus for quality improvement.[1] The goal of HCT is to maximize lifelong functioning and potential through the provision of uninterrupted, high-quality, developmentally appropriate health care services.[2] The lack of effective transition from pediatric to adult-focused health care may contribute to fragmented or delayed care and increased risk for adverse health outcomes. Improving transition preparation for at-risk youth may well decrease costs associated with inappropriate or delayed health care utilization.

References:

1. Lotstein DS, McPherson M, Strickland B, Newacheck PW. Transition planning for youth with special health care needs: results from the National Survey of Children with Special Health Care Needs. *Pediatrics*. 2005;115(6):1562-1568. doi:10.1542/peds.2004-1262.

2. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians-American Society of Internal Medicine. A consensus statement on health care transitions for young adults with special health care needs. *Pediatrics*. 2002;110(6 Pt 2):1304-1306.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

TRANSITION PREPARATION AS A FOCUS OF QUALITY MEASUREMENT

Measurement of transition preparation is essential to assess and improve the quality of transition care in the US and beyond. Research findings (Table 1) underscore the need for more purposeful transition planning across the spectrum of pediatric chronic conditions and have led to consensus regarding the importance of improving transition preparation. This consensus is manifest in recommendations outlined by national organizations such as the American Academy of Pediatrics.[1] Consequences of failure to provide effective preparation for transition from pediatric to adult-centered care have been described, including high rates of emergency care utilization among adults ages 20-29 in the US[2] and pediatric hospitalizations for young adults with chronic conditions.[3,4] However, at present, there is a paucity of data linking adequate transition preparation and readiness (the desired goal of preparation) with improved adult health outcomes. A small number of quasi-experimental studies, all conducted in patients with type 1 diabetes, have shown that transition preparation interventions were associated with improved frequency of post-transition medical follow-up,[5-7] reduced acute diabetes complications[5] and improved hemoglobin A1c levels.[6] While limited, available data suggest that the lack of effective transition from pediatric- to adult-focused care may contribute to fragmentation of young adult health care and increased risk for adverse outcomes.

RATIONALE FOR A YOUTH-REPORTED MEASURE OF TRANSITION PREPARATION

Because transition preparation must be tailored to adolescents' evolving self-management skills and level of independence, direct assessment of youth experiences within the health care system is an important means of quality measurement.

Adolescents are best able to judge how well their providers are meeting their needs. Notably, the association between patient-centered care and health outcomes has been shown to be stronger when patient-centeredness is measured by patient report rather than provider or researcher assessment.[8,9] Such assessment is likely to stimulate additional improvements in patient-centered processes and outcomes of care.

Research in adolescents has indicated that youth self-report is reliable in evaluation of health service delivery.[10,11] Because consensus recommendations for transition preparation identify 14-15 years as the ideal age to initiate the development of a patient-specific transition

plan,[12]querying patients at 16-17 years captures them at a time by which some transition preparation generally should have occurred. A review of 43 transition studies published from 1982-2003 found that the most frequently cited age range for ideal transition was between 16 and 22 years. Despite consensus recommendations, only few studies have reported initiation of transition planning at 15 years or younger.[13]

IDENTIFICATION OF KEY DOMAINS FOR TRANSITION PREPARATION

Although there are few existing measures of the quality of transition preparation, recent consensus statements recommend that health care providers prepare their patients by discussing realistic goals, creating a timeline, and developing a transition plan starting at age 14.[12] Explicit discussion of transfer to adult care is a key component of existing parent-reported measures. Other domains of transition preparation include development of self-management skills, appropriate adolescent autonomy, improved youth-provider communication, and skills for self-advocacy. Examples of self-management and self-advocacy skills include scheduling one's own medical appointments, obtaining medications and prescription refills, having one-on-one conversations with medical providers, being familiar with one's medical history, understanding health insurance coverage, and feeling empowered to manage one's own medical conditions. Many of these skills have been incorporated into transition readiness scales.[14-16] However, adolescent reports of receipt of counseling regarding these skills have not previously been included in measures of health care quality.

For more details regarding the association between HCT and other aspects of healthcare, see ***Evidence Table (Appendix M).***

References:

1. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians-American Society of Internal Medicine. A consensus statement on health care transitions for young adults with special health care needs. *Pediatrics*. 2002;110(6 Pt 2):1304-1306.
2. Fortuna RJ, Robbins BW, Halterman JS. Ambulatory care among young adults in the United States. *Ann Intern Med*. 2009;151(6):379-385.
3. Nakhla M, Daneman D, To T, Paradis G, Guttmann A. Transition to adult care for youths with diabetes mellitus: findings from a Universal Health Care System. *Pediatrics*. 2009;124(6):e1134-1141. doi:10.1542/peds.2009-0041.
4. Goodman DM, Mendez E, Throop C, Ogata ES. Adult survivors of pediatric illness: the impact on pediatric hospitals. *Pediatrics*. 2002;110(3):583-589.
5. Holmes-Walker DJ, Llewellyn AC, Farrell K. A transition care programme which improves diabetes control and reduces hospital admission rates in young adults with Type 1 diabetes aged 15-25 years. *Diabet Med J Br Diabet Assoc*. 2007;24(7):764-769. doi:10.1111/j.1464-5491.2007.02152.x.

6. Cadario F, Prodam F, Bellone S, Trada M, Binotti M, Trada M, Allochis G, Baldelli R, Esposito S, Bona G, Aimaretti G. Transition process of patients with type 1 diabetes (T1DM) from paediatric to the adult health care service: a hospital-based approach. *Clin Endocrinol (Oxf)*. 2009;71(3):346-350. doi:10.1111/j.1365-2265.2008.03467.x.
7. Van Walleghem N, Macdonald CA, Dean HJ. Evaluation of a systems navigator model for transition from pediatric to adult care for young adults with type 1 diabetes. *Diabetes Care*. 2008;31(8):1529-1530. doi:10.2337/dc07-2247.
8. Stewart M, Brown JB, Donner A, McWhinney IR, Oates J, Weston WW, Jordan J. The impact of patient-centered care on outcomes. *J Fam Pract*. 2000;49(9):796-804.
9. Theunissen NC, Vogels TG, Koopman HM, Verrips GH, Zwinderman KA, Verloove-Vanhorick SP, Wit JM. The proxy problem: child report versus parent report in health-related quality of life research. *Qual Life Res Int J Qual Life Asp Treat Care Rehabil*. 1998;7(5):387-397.
10. Klein JD, Graff CA, Santelli JS, Hedberg VA, Allan MJ, Elster AB. Developing quality measures for adolescent care: validity of adolescents' self-reported receipt of preventive services. *Health Serv Res*. 1999;34(1 Pt 2):391-404.
11. Santelli J, Klein J, Graff C, Allan M, Elster A. Reliability in adolescent reporting of clinician counseling, health care use, and health behaviors. *Med Care*. 2002;40(1):26-37.
12. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians, Transitions Clinical Report Authoring Group, Cooley WC, Sagerman PJ. Supporting the health care transition from adolescence to adulthood in the medical home. *Pediatrics*. 2011;128(1):182-200. doi:10.1542/peds.2011-0969.
13. Betz CL. Transition of adolescents with special health care needs: review and analysis of the literature. *Issues Compr Pediatr Nurs*. 2004;27(3):179-241. doi:10.1080/01460860490497903.
14. Sawicki GS, Lukens-Bull K, Yin X, Demars N, Huang I-C, Livingood W, Reiss J, Wood D. Measuring the transition readiness of youth with special healthcare needs: validation of the TRAQ--Transition Readiness Assessment Questionnaire. *J Pediatr Psychol*. 2011;36(2):160-171. doi:10.1093/jpepsy/jsp128.
15. Ferris ME, Harward DH, Bickford K, Layton JB, Ferris MT, Hogan SL, Gipson DS, McCoy LP, Hooper SR. A clinical tool to measure the components of health-care transition from pediatric care to adult care: the UNC TR(x)ANSITION scale. *Ren Fail*. 2012;34(6):744-753. doi:10.3109/0886022X.2012.678171.
16. Gilleland J, Amaral S, Mee L, Blount R. Getting ready to leave: transition readiness in adolescent kidney transplant recipients. *J Pediatr Psychol*. 2012;37(1):85-96. doi:10.1093/jpepsy/jsr049.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_ADAPT_Evidence_submission_form-635792190899030707.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

IMPORTANCE OF MEASURING THE QUALITY OF TRANSITION FROM PEDIATRIC TO ADULT-FOCUSED HEALTH CARE
Health care transition (HCT) has been defined as a planned, purposeful process in which adolescents and young adults move from pediatric-focused health care delivery to adult-focused delivery.[1] The goal of HCT is to maximize lifelong functioning and potential through the provision of uninterrupted, high-quality, developmentally-appropriate health care services.[1] The lack of effective transition from pediatric to adult-focused health care may contribute to fragmentation of health care and increased risk for adverse health outcomes. Those at highest risk during this period include youth with special health care needs (YSHCN).[2]

The process of HCT involves 3 key phases: 1) transition planning and preparation; 2) transfer of health care to an adult-focused model; and 3) intake to the adult-focused health system. There is broad consensus that preparation for HCT should start in adolescence and involve individualized planning and ongoing skills development.[3]

In 2002, a consensus statement from the American Academy of Pediatrics, the American Academy of Family Physicians, and the American College of Physicians envisioned the goal that by 2010 “all physicians who provide primary or subspecialty care to young people with special health care needs 1) understand the rationale for transition from child-oriented to adult-oriented health care; 2) have the knowledge and skills to facilitate that process; and 3) know if, how, and when transfer of care is indicated.”[1] For youth receiving care in pediatric-focused health care settings, preparation for HCT includes the acquisition of self-care skills and promotion of increased youth responsibility for chronic condition management. For many youth, transition preparation culminates in a transfer to a new health care setting. However, even for youth who do not change care settings (e.g., those in family medicine settings), the shift to adult-oriented health care still requires appropriate preparation. Because transition preparation is primarily a series of interactions with clinicians, obtaining reports directly from youth about their experience is critical to understanding current gaps in health care delivery for this population.

PREPARATION FOR HEALTH CARE TRANSITION: LACK OF STANDARDIZED QUALITY MEASUREMENT
In its 2011 Patient-Centered Medical Home Standards, the National Committee on Quality Assurance included a specific requirement to address care transitions in primary care.[4] The MCHB identified HCT services as a core outcome for the community-based services required for CSHCN under Title V and Healthy People 2000 and

reiterated this priority in the Healthy People 2010 and Healthy People 2020 goals.[1,5-6] However, systematic assessments of transition readiness are rarely incorporated as part of routine health care.[6] Measuring the quality of HCT preparation is intended to drive providers to adopt strategies that foster disease self-management among youth and reliably result in safe and effective transfer to adult care.[7]

DISPARITIES IN HEALTH CARE TRANSITION PREPARATION

Geographic, socioeconomic, racial and ethnic disparities have been documented in the receipt of HCT services.[8-10] In the 2005-2006 NS-CSHCN, fewer African-American and Latino respondents reported having discussed shifting their child's care to an adult-focused provider.[8] In the same survey, the proportion of respondents who met the core performance outcomes for successful transition increased significantly with increasing family income.[5] Additionally, the 2007 SATH revealed that low-income young adults had poorer access to health care than those with higher incomes.[8] Such disparities in transition preparation and access to care are likely to result in adverse health outcomes.

References:

1. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians-American Society of Internal Medicine. A consensus statement on health care transitions for young adults with special health care needs. Pediatrics. 2002;110(6 Pt 2):1304-1306.
2. Lotstein DS, McPherson M, Strickland B, Newacheck PW. Transition planning for youth with special health care needs: results from the National Survey of Children with Special Health Care Needs. Pediatrics. 2005;115(6):1562-1568. doi:10.1542/peds.2004-1262.
3. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians, Transitions Clinical Report Authoring Group, Cooley WC, Sagerman PJ. Supporting the health care transition from adolescence to adulthood in the medical home. Pediatrics. 2011;128(1):182-200. doi:10.1542/peds.2011-0969.
4. Standards for Patient-Centered Medical Home (PCMH), in Washington, DC: National Committee for Quality Assurance, Editor. 2011.
5. Lotstein DS, Ghandour R, Cash A, McGuire E, Strickland B, Newacheck P. Planning for health care transitions: results from the 2005-2006 National Survey of Children With Special Health Care Needs. Pediatrics. 2009;123(1):e145-152. doi:10.1542/peds.2008-1298.
6. McManus MA, Pollack LR, Cooley WC, McAllister JW, Lotstein D, Strickland B, Mann MY. Current status of transition preparation among youth with special needs in the United States. Pediatrics. 2013;131(6):1090-1097. doi:10.1542/peds.2012-3050.
7. Park MJ, Adams SH, Irwin CE. Health care services and the transition to young adulthood: challenges and opportunities. Acad Pediatr. 2011;11(2):115-122. doi:10.1016/j.acap.2010.11.010.
8. Lotstein DS, Kuo AA, Strickland B, Tait F. The transition to adult health care for youth with special health care needs: do racial and ethnic disparities exist? Pediatrics. 2010;126 Suppl 3:S129-136. doi:10.1542/peds.2010-1466F.
9. Richmond N, Tran T, Berry S. Receipt of transition services within a medical home: do racial and geographic disparities exist? Matern Child Health J. 2011;15(6):742-752. doi:10.1007/s10995-010-0635-2.
10. Kane DJ, Kasehagen L, Punyko J, Carle AC, Penziner A, Thorson S. What factors are associated with state performance on provision of transition services to CSHCN? Pediatrics. 2009;124 Suppl 4:S375-383. doi:10.1542/peds.2009-1255H.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

In 2013-2014, we conducted field testing in 3 populations: 2 health plans and a freestanding pediatric hospital. Although the test sites in our field testing varied widely in their geographic location and demographic characteristics, measure scores and responses to individual questions were similar across the 3 field tests.

For each health plan, surveys were sent to 1,500 families of youth with complex chronic disease and 1,500 families

of youth with noncomplex chronic disease in each health plan. There were 248 and 231 undeliverable surveys in each health plan sample, respectively. For Health Plan 1 recipients, surveys were sent in English with an option for the family to contact the health plan to receive a survey in Spanish. In Health Plan 2, both English and Spanish language surveys were sent to each recipient. We received 1,339 surveys (780 of 2,734 from Health Plan 1 [18 completed in Spanish] and 575 of 2,752 from Health Plan 2 [28 completed in Spanish]) for a final response rate of 28% and 21%, respectively.

For the clinical program field test, we emailed surveys to parents of 623 outpatients receiving care at Boston Children's Hospital aged 16-17 years identified as receiving care in 1 of 10 different clinical programs (Endocrinology, Gastroenterology, Hematology, Sickle Cell and Hemophilia, Immunology, Metabolism, Nephrology, Primary Care, Pulmonology, and Spina Bifida). The families were identified by clinicians or clinical coordinators of each of the participating clinics. A total of 293 of 617 of these surveys were returned (response rate 47%); 6 surveys were undeliverable.

OVERALL SCORES BY SITE
We provide below the performance measure scores by site for each of the ADAPT measures using data from the 2013-2014 national field test. We report mean and 95% confidence intervals for each measure.

Site X
Measure X
Line 1 – Mean (M); 95% confidence interval (CI)

Hospital 1 (n=293)
Counseling on Transition Self-Management:
M: 32; CI: 30, 35

Counseling on Prescription Medication
M: 61; CI: 59, 64

Transfer Planning
M: 5; CI: 3, 7

Health Plan 1 (n=780)
Counseling on Transition Self-Management:
M: 36; CI: 34, 38

Counseling on Prescription Medication
M: 57; CI: 55, 60

Transfer Planning
M: 4; CI: 3, 5

Health Plan 2 (n=575)
Counseling on Transition Self-Management:
M: 30; CI: 28, 33

Counseling on Prescription Medication
M: 58; CI: 54, 62

Transfer Planning
M: 3; CI: 2, 4

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

RACE ETHNICITY

To assess racial/ethnic disparities in provision of health care transition (HCT) preparation, we conducted field tests in diverse populations and assessed differences in ADAPT scores by race/ethnicity. Adolescent race/ethnicity is determined on the ADAPT survey using 2 questions based on those used by the Office of Minority Health: "Are you of Hispanic, Latino, or Spanish origin?" and "How would you describe your race?" (Questions 23 and 24).

Among respondents in the Health Plan 1 field test, 4.6% were Asian/Pacific Islander, 24.0% were Black, and 16.0% were Hispanic. For respondents in the Health Plan 2 field test, 5.7% were Asian/Pacific Islander, 18.0% were Black, and 59.0% were Hispanic. For the analyses of differences in ADAPT scores by race/ethnicity, we therefore categorized responses into the following groups: Asian/Pacific Islander, Black, Hispanic, White, and Other.

ADAPT scores by race/ethnicity for the 2 Health Plans are shown in Appendix J. Scores stratified by race/ethnicity for the Transition Self-Management measure were higher for most racial/ethnic groups in Health Plan 1 than Health Plan 2, although this difference was only statistically significant for Black patients. In the other 2 measure scores, no differences between health plans were observed. Within Health Plan 1, we observed higher Transition Self-Management Scores for Black patients compared to White patients, but no significant within-health-plan differences by race/ethnicity were observed in Health Plan 2. No within-health-plan differences were detected between White and Hispanic patients in either health plan. There were too few patients of Asian/Pacific Islander race/ethnicity in either health plan for comparison with White patients. It should be noted that this field test was not designed to provide statistical power to detect differences between racial/ethnic groups. If such comparisons are desired, we recommend a sample size of 300 respondents per group being compared. This would likely require oversampling of patients of less common race/ethnicity in a health plan. Given the range of scores in each of the 3 measures, a sample size of 300 respondents per group would provide 80% power to detect approximately a 10% difference in both the Counseling on Transition Self-Management measure and the Counseling on Prescription Medication measure and a 5% difference in the Transfer Planning measure.

SPECIAL HEALTH CARE NEEDS

The ADAPT survey is designed for adolescents with special health care needs, as defined by the presence of at least 1 chronic condition. However, experiences with HCT preparation may vary depending on the type of chronic condition. Therefore, we assessed differences in ADAPT scores based on patients´ type of chronic health condition as defined in 1 of 2 ways.

For the clinical program field test, we assigned patients´ type of condition according to the subspecialty of the clinical program (e.g., Endocrinology, Pulmonary) in which they received care. For the two health plan field tests, we determined the type of condition by applying the PMCA to claims data from the health plans. We evaluated variation in ADAPT scores associated with special health care needs in several ways. We compared scores among respondents with noncomplex chronic (NC-CD) diseases with those with complex chronic (C-CD) diseases and found no significant differences in any of the measure scores (Appendix J). However, as with the race/ethnicity analyses, statistical power was insufficient to detect differences in scores by this variable.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Affects large numbers, High resource use, Patient/societal consequences of poor quality

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

The healthcare system has shifted dramatically toward the delivery of patient-centered care. Patient-centeredness refers to the principle that care should be designed around patients' needs, preferences, circumstances, and well-being. It has been identified as a core aspect of healthcare quality that should be addressed as part of overall quality improvement strategies.[1-2]

Adolescents and young adults with chronic health care conditions are particularly vulnerable to adverse health outcomes related to prolonged interruptions in health care delivery. Such interruptions are particularly common during periods of HCT.[3] Young adults with diabetes who felt unprepared for transition had an increased likelihood of gaps in care greater than 6 months between pediatric and adult care than those who felt more prepared.[4] Furthermore, young adults use less ambulatory and preventive care.[5] For example, young adults with asthma were less likely to have a primary care visit, less likely to fill a short-acting beta-agonist prescription, and more likely to visit an emergency department compared with adolescents with asthma.[6] In 2009, individuals aged 18 to 26 had the lowest health care utilization rates of any age group, and a significant percentage delayed accessing health care due to cost.[7] Many young adults, particularly those with chronic disease and those with public health insurance, also have delayed HCT.[8]

Other data suggest that youth may be transitioning out of pediatric care without appropriate follow-up, skills, or knowledge needed to succeed in an adult-oriented system.[9-10] Measuring the quality of HCT preparation of youth with special health care needs (YSHCN) has great potential to motivate improvements by health care professionals and systems for the patients most likely to benefit. At the same time, approaches to improving HCT preparation for YSHCN could be applied more broadly to improve the transition process for all adolescents as they transition to adult-focused care delivery. Lack of preventive care and timely ambulatory services is associated with increased overall costs as health conditions progress and require higher levels of care.[11] Improving transition preparation for at-risk youth may well decrease costs associated with inappropriate or delayed health care utilization.

Nationally, there is a striking lack of attention to implementing recommendations for HCT outlined in consensus statements and little uniformity in approach even within health care systems. In the 2001 National Survey of Children with Special Health Care Needs (NS-CSHCN), a minority of parents reported having discussed transition with their child's physician,[3] and only 30% had a plan for addressing transition needs.[12] In the 2005-2006 NS-

CSHCN, this percentage remained below 50%.[13] Applying Maternal and Child Health Bureau (MCHB) transition services quality metrics to these parent-reported data revealed variable state-level performance, with an individual state´s performance predicted by the proportion of patients with a medical home and adequate health insurance.[14] Compared to the 2005-2006 survey, no significant improvement in rates of transition preparation was found in the 2009-2010 NS-CSHCN.[15-16] In the 2007 Survey of Adolescent Transition and Health (SATH), approximately half of patients aged 19 to 23 years reported receiving counseling around transition.[11] These findings all suggest considerable room for improvement in HCT preparation for YSHCN. However, this potential will be realized only with adequate measurement, benchmarking of performance, and concerted efforts to improve care. No national surveys have directly assessed transition preparation from the perspective of adolescents themselves. The ADAPT survey addresses this information gap.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Browne K, Roseman D, Shaller D, Edgman-Levitan S. Analysis & commentary. Measuring patient experience as a strategy for improving primary care. Health Aff Proj Hope. 2010;29(5):921-925. doi:10.1377/hlthaff.2010.0238.
2. Cosgrove DM, Fisher M, Gabow P, Gottlieb G, Halvorson GC, James BC, Kaplan GS, Perlin JB, Petzel R, Steele GD, Toussaint JS. Ten strategies to lower costs, improve quality, and engage patients: the view from leading health system CEOs. Health Aff Proj Hope. 2013;32(2):321-327. doi:10.1377/hlthaff.2012.1074.
3. Lotstein DS, McPherson M, Strickland B, Newacheck PW. Transition planning for youth with special health care needs: results from the National Survey of Children with Special Health Care Needs. Pediatrics. 2005;115(6):1562-1568. doi:10.1542/peds.2004-1262.
4. Garvey KC, Wolpert HA, Rhodes ET, Laffel LM, Kleinman K, Beste MG, Wolfsdorf JI, Finkelstein JA. Health care transition in patients with type 1 diabetes: young adult experiences and relationship to glycemic control. Diabetes Care. 2012;35(8):1716-1722. doi:10.2337/dc11-2434.
5. Fortuna RJ, Robbins BW, Halterman JS. Ambulatory care among young adults in the United States. Ann Intern Med. 2009;151(6):379-385.
6. Chua K-P, Schuster MA, McWilliams JM. Differences in health care access and utilization between adolescents and young adults with asthma. Pediatrics. 2013;131(5):892-901. doi:10.1542/peds.2012-2881.
7. Lau JS, Adams SH, Irwin CE. Young Adult Health Care Utilization and Expenditures Before the Implementation of the Affordable Care Act. J Adolesc Health. 2013;52(2):S44. doi:10.1016/j.jadohealth.2012.10.105.
8. Fortuna RJ, Halterman JS, Pulcino T, Robbins BW. Delayed transition of care: a national study of visits to pediatricians by young adults. Acad Pediatr. 2012;12(5):405-411. doi:10.1016/j.acap.2012.04.002.
9. Reiss J, Gibson R. Health care transition: destinations unknown. Pediatrics. 2002;110(6 Pt 2):1307-1314.
10. Rosen D. Between two worlds: bridging the cultures of child health and adult medicine. J Adolesc Health Off Publ Soc Adolesc Med. 1995;17(1):10-16. doi:10.1016/1054-139X(95)00077-6.
11. Sawicki GS, Whitworth R, Gunn L, Butterfield R, Lukens-Bull K, Wood D. Receipt of health care transition counseling in the national survey of adult transition and health. Pediatrics. 2011;128(3):e521-529. doi:10.1542/peds.2010-3017.
12. Scal P, Ireland M. Addressing transition to adult health care for adolescents with special health care needs. Pediatrics. 2005;115(6):1607-1612. doi:10.1542/peds.2004-0458.
13. Lotstein DS, Ghandour R, Cash A, McGuire E, Strickland B, Newacheck P. Planning for health care transitions: results from the 2005-2006 National Survey of Children With Special Health Care Needs. Pediatrics. 2009;123(1):e145-152. doi:10.1542/peds.2008-1298.
14. Kane DJ, Kasehagen L, Punyko J, Carle AC, Penziner A, Thorson S. What factors are associated with state performance on provision of transition services to CSHCN? Pediatrics. 2009;124 Suppl 4:S375-383. doi:10.1542/peds.2009-1255H.
15. McManus MA, Pollack LR, Cooley WC, McAllister JW, Lotstein D, Strickland B, Mann MY. Current status of transition preparation among youth with special needs in the United States. Pediatrics. 2013;131(6):1090-1097. doi:10.1542/peds.2012-3050.
16. Strickland B, McPherson M, Weissman G, van Dyck P, Huang ZJ, Newacheck P. Access to the medical home: results of the National Survey of Children with Special Health Care Needs. Pediatrics. 2004;113(5 Suppl):1485-1492.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related**

behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.
(Describe how and from whom their input was obtained.)

FOCUS GROUPS

Assessing patient perspectives about transition preparation and transfer process experiences is integral to the development of a valid, self-reported survey of the quality of transition preparation.[1] To understand the health care transition experiences of greatest salience to patients and families, we conducted a series of focus groups. Our objective was to identify the critical elements in the preparation for transition to independent self-care and the transfer to adult medical care for patients with a variety of chronic illnesses.

Focus groups were conducted with adolescent (age 16-18 years) and young adult (age 19-26 years) patients with 1 or more chronic health conditions, as well as parents/guardians of youth or young adults with chronic health conditions. Although the ADAPT survey is designed to be completed by youth aged 16-17 years as they prepare for transition, we conducted focus groups with young adults who had already transitioned to better understand the actual process of transfer to adult care and to ensure the relevance of the measure to all stages of the transition experience. In addition, to understand the role and perspective of caregivers in the transition process, we conducted focus groups with parents/guardians of both adolescents and young adults with chronic health conditions.

In total, we conducted 11 focus groups in Boston, Chicago, and Los Angeles: 3 with adolescents, 4 with young adults, and 4 with parents/guardians. One of the young adult groups and 2 of the parent/guardian groups consisted of participants whose primary language was Spanish, and the focus groups were conducted in this language. The focus groups included a diverse spectrum of patients with regard to gender, race, ethnicity, and type of chronic health condition.

During each focus group, a trained moderator facilitated discussion on the following domains: changing disease self-management responsibilities; readiness for transition; transition preparation; and health insurance during transition. In groups of post-transition young adults or parents/guardians of post-transition young adults, the moderator also asked about experiences of the transfer to adult health care.

Key findings that informed survey development included:
• Adolescents reported that they had thought little about transition to adult-focused care prior to focus group participation and infrequently discussed these issues with others.
• Very few adolescents perceived purposeful transition preparation on the part of pediatric health care providers.
• Adolescents frequently expressed ambivalence about taking charge of their own health, as well as frustration that their health care providers did not consistently involve them in discussions about their health.
• Post-transition young adults reported a near-complete lack of pediatric counseling regarding independent self-care or transfer to adult care.
• Young adults described feeling responsible for locating new adult providers with little support or guidance from pediatric health care providers.
• Both adolescents and young adults reported poor understanding regarding how health insurance works.
• Parents/guardians of adolescents and young adults were unsure of their roles relative to health care provider roles in counseling their children about disease self-care.
• Parents/guardians expressed great concern about gaps in care or inconsistent care during transition and the potential for related declines in their children's health.

COGNITIVE INTERVIEWS

We synthesized focus group findings with data from our extensive literature review and expert interviews to develop a draft survey.[1] We then conducted cognitive interviews to assess whether the intended respondents, 16- to 17-year-old adolescents with chronic health conditions, understood each of the draft survey questions as intended. Before the cognitive interviews, participants were asked to respond to the survey. The interview protocol contained candidate questions from the draft survey followed by pre-specified cognitive probes to evaluate the understandability of specific words and phrases and to clarify participant thought processes in answering the

questions. Participants were also given the opportunity to suggest alternative language for specific questions.

We performed 4 rounds of 26 total cognitive interviews of youth respondents in English and Spanish in Boston, Chicago, and Dallas. The goals of sequential rounds of interviews were to test versions of questions about transition and to make minor revisions to questions that were not uniformly understood. After 4 rounds, the results generally demonstrated that adolescents responded to most of the survey questions in the intended way. The English and Spanish versions of the survey elicited similar responses. Responses to many of the survey questions showed variation as expected based on the range of experiences of the participants.

To ensure that the survey will be useful and understandable to patients and their families, we have made iterative revisions based on feedback obtained throughout the survey development process.[1]

References:
1. Sawicki GS, Garvey KC, Toomey SL, Williams KA, Chen Y, Hargraves JL, Leblanc J, Schuster MA, Finkelstein JA. Development and Validation of the Adolescent Assessment of Preparation for Transition: A Novel Patient Experience Measure. J Adolesc Health. 2015;57(3):282-287. doi:10.1016/j.jadohealth.2015.06.004.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Prevention : Development/Wellness

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Patient and Family Engagement

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.childrenshospital.org/research-and-innovation/research/centers/center-of-excellence-for-pediatric-quality-measurement-cepqm/cepqm-measures/transition-from-child-focused-to-adult-focused-care

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ADAPT_Data_Dictionary.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last

endorsement date and explain the reasons.
Not applicable

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.*
The ADAPT survey consists of 26 questions assessing the quality of health care transition preparation for youth with chronic health conditions, based on youth report of whether specific recommended processes of care were received. The ADAPT survey generates measures for each of 3 domains: 1) Counseling on Transition Self-Management, 2) Counseling on Prescription Medication, and 3) Transfer Planning. ADAPT measure scores are calculated using the sum of the proportions of positive responses to between 3 and 5 individual items. Complete instructions for measure score calculations are provided in the Detailed Measure Specifications (Appendix A).

1) Counseling on Transition Self-Management:
The numerator is the sum of the proportions of positive responses to the five questions about counseling on transition self-management, among respondents with valid responses to all questions.

2) Counseling on prescription medication:
The numerator is the sum of the proportions of positive responses to the three questions about counseling on prescription medication, among respondents who indicate that they take prescription medication every day and with valid responses to all questions.

3) Transfer planning:
The numerator is the sum of the proportions of positive responses to the four questions about transfer planning, among respondents who report being treated by a pediatric provider and with valid responses to all questions.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The ADAPT survey can be sent to youth with a chronic health condition receiving care from a particular clinical program or health plan. The survey is based on recall of care over a 12-month period.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
ADAPT measure scores are calculated using the sum of the proportions of positive responses to between 3 and 5 individual items. Complete instructions for measure score calculations are provided in the Detailed Measure Specifications (Appendix A).

MEASURE 1. Counseling on Transition Self-Management:
For any individual respondent, the numerator is the number of positive responses to the five questions about counseling on transition self-management divided by five. For the group of respondents, the numerator is the sum of these proportions divided by the number of respondents with valid responses to all questions.

This measure is produced by combining responses to questions 4-8:
- Q4: In the last 12 months, did you talk with this provider without your parent or guardian in the room?
- Q5: In the last 12 months, did you and this provider talk about your being more in charge of your health?
- Q6: In the last 12 months, did you and this provider talk about your scheduling your own appointments with this

provider instead of your parent or guardian?
• Q7: In the last 12 months, how often did you schedule your own appointments with this provider?
• Q8: In the last 12 months, did you and this provider talk about how your health insurance might change as you get older?

MEASURE 2. Counseling on prescription medication:
For any individual respondent, the numerator is the number of positive responses to the three questions about counseling on prescription medication divided by three. For the group of respondents, the numerator is the sum of these proportions divided by the number of respondents who indicate that they take prescription medication every day and with valid responses to all questions.

The measure is produced by combining responses to questions 10, 12, and 13:
• Q10: In the last 12 months, how often did you and this provider talk about all of your prescription medicines at each visit?
• Q12: In the last 12 months, did you and this provider talk about remembering to take your medicines?
• Q13: In the last 12 months, did you and this provider talk about you refilling your own prescriptions instead of your parent or guardian?

MEASURE 3. Transfer planning:
For any individual respondent, the numerator is the number of positive responses to the four questions about transfer planning divided by four. For the group of respondents, the numerator is the sum of these proportions divided by the number of respondents who report being treated by a pediatric provider and with valid responses to all questions.

The measure is produced by combining responses to questions 15, 16, 17, and 18:
• Q15: In the last 12 months, did you and this provider talk about whether you may need to change to a new provider who treats mostly adults?
• Q16: In the last 12 months, did this provider ask if you had any questions or concerns about changing to a new provider who treats mostly adults?
• Q17: In the last 12 months, did you and this provider talk about a specific plan for changing to a new provider who treats mostly adults?
• Q18: Did this provider give you this plan in writing?

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The target population of the survey is 16- or 17-year-old adolescents with a chronic health condition who are either (a) receiving health care services in a clinical program or (b) enrolled in a health plan or similar defined population.

The denominator for each measure is the number of respondents with valid responses for all of the questions in the measure.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk : Individuals with multiple chronic conditions

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
SURVEY
The denominator for the survey is youth who meet the following criteria:
1. Either (a) receiving health care services in a particular clinical program or (b) enrolled in a health plan or similar defined population

2. Age 16 to 17 years old at the time of survey completion

3. At least 1 chronic health condition. In the case of a defined population (e.g., a health plan), tools such as the Pediatric Medical Complexity Algorithm (PMCA) can be used to identify eligible patients by chronic condition status.[1] The PMCA is a publicly available algorithm that uses International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) diagnosis codes in health plan claims to identify children with either complex chronic disease (C-CD) or noncomplex chronic disease (NC-CD).

4. At least 1 outpatient visit with a health care provider in the preceding 12 months

5. For health plan sampling, current enrollment at the time of the survey and enrollment over the preceding 12 months (allowing <45 day gaps during that period, if present)

MEASURE SCORES
A valid response for each question is that entered by the respondent or assigned according to the decision rules outlined in Appendix L.

For Measure 1, the denominator is the number of respondents with valid responses to all of the questions within the measure (Questions 4-8).

For Measure 2, the denominator is the number of respondents with responses of "Yes" to Question 11 and valid responses to all of the questions within the measure (Question 10, 12, 13).

For Measure 3, the denominator is the number of respondents with responses of "Yes," "Don't know," or left blank to Question 14 and valid responses to all of the questions within the measure (Question 15-18).

References:
1. Simon TD, Cawthon ML, Stanford S, Popalisky J, Lyons D, Woodcox P, Hood M, Chen AY, Mangione-Smith R, Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN) Medical Complexity Working Group. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. Pediatrics. 2014;133(6):e1647-1654. doi:10.1542/peds.2013-3875.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
SURVEY SAMPLE
Exclude patients in the following categories from the ADAPT survey sample frame:

1. "No-publicity" patients (i.e., those who requested that they not be contacted)

2. Court/law enforcement patients

3. Patients with a foreign home address

4. Patients who cannot be surveyed because of local, state, or federal regulations

SURVEY RESPONSE
Exclude survey respondents based on the following clinical and non-clinical criteria:

1. Undeliverable survey, i.e., the survey is returned by US Mail as undeliverable. "Undeliverable" should not be assumed merely because of non-response.

2. The survey is returned with clear indication that the patient does not meet eligibility criteria (e.g., ineligible age

or lack of a chronic health condition).

3. Patient unable to complete survey independently: This must be indicated by the appropriate checkbox in the cover letter or equivalent clear indication by the parent/guardian that the patient is unable to complete the survey independently (e.g., due to cognitive limitation).

4. Exclude all respondents who answered "None" to ADAPT question 3 ("In the last 12 months, how many times did you visit this provider?").

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Court/law enforcement patients (i.e., prisoners) are excluded from the sample frame because of the logistical difficulties of administering the survey in a timely manner and regulations governing surveys of this population.

Patients with a foreign home address are excluded because of the logistical difficulty and added expense of calling or mailing outside of the United States. (The US territories—American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and Virgin Islands—are not considered foreign addresses and are not excluded.)

Some state regulations place further restrictions on which patients may be contacted for surveys. It is the responsibility of the health plan, clinical program, or survey vendor to identify any applicable laws or regulations and to exclude those patients as required in the state in which the entity operates.

Note: Include patients in the sample frame unless there is positive evidence that they are ineligible or fall within an excluded category. If information is missing on any variable that affects survey eligibility when the sample frame is constructed, do not exclude the patient from the sample frame because of that variable.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Stratification is not required. However, users of the survey may choose to stratify scores. In a defined population (e.g., a health plan), potential variables for stratification could include type of chronic health condition or diagnosis.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Case-mix adjustment:
One of the methodological issues associated with making comparisons across populations is the need to adjust appropriately for case-mix differences. Case-mix refers to patient characteristics, such as demographic characteristics and health status, which may affect measures of outcomes or processes. Systematic effects of this sort create the potential for a population's scores to be higher or lower because of its characteristics, rather than because of the quality of care provided, making comparisons of unadjusted scores misleading. The basic goal of adjusting for case-mix is to estimate how different clinical programs or health plans would be rated if they all provided care to comparable groups of patients.

Case-mix adjustment using linear regression is used to adjust clinical program/health plan-level ADAPT measure scores based on patient characteristics, thus facilitating comparisons among clinical programs/health plans. We recommend adjusting for respondent age and self-reported health status.

The case-mix data are obtained from questions in the "About You" section of the survey: 1) Respondent age: ADAPT Q19, and 2) Self-reported health status: ADAPT Q21

Detailed instructions regarding how to use the case-mix adjustment model can be found in Case-Mix Adjustment Methodology (Appendix B).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
A case-mix adjustment model is available for each of the 11 ADAPT questions that are used in the 3 measure scores. In each model, the dependent variable is the response to a question and the independent variables are the two case-mix adjusters: respondent age and self-reported health status.

The methodology for these case-mix adjustment models was originally developed for the Child Hospital Consumer Assessment of Healthcare Provider and Systems (Child HCAHPS) Survey. A SAS Macro (e.g., CAHPS SAS macro, publically available at https://cahps.ahrq.gov/surveys-guidance/survey4.0-docs/2015_instructions_for_analyzing_data.pdf) may be used to generate unadjusted and adjusted measure scores for this survey. Detailed instructions regarding the case-mix adjustment model can be found in Appendix B.

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
There are 3 domain-level measures included in the ADAPT survey. The calculation of measure scores is described below.

1) Counseling on Transition Self-Management:
This measure is produced by combining responses to 5 questions:
• Q4: In the last 12 months, did you talk with this provider without your parent or guardian in the room?
• Q5: In the last 12 months, did you and this provider talk about your being more in charge of your health?
• Q6: In the last 12 months, did you and this provider talk about your scheduling your own appointments with this provider instead of your parent or guardian?
• Q7: In the last 12 months, how often did you schedule your own appointments with this provider?
• Q8: In the last 12 months, did you and this provider talk about how your health insurance might change as you get older?

The 5 questions are scored as indicated in Figure 1 in Appendix A.

Response options for questions 4-6 and 8 are "Yes" or "No":
• Assign a score of 0 for No
• Assign a score of 1 for Yes

Response options for question 7 are "Never," "Sometimes," "Usually," or "Always":
• Assign a score of 0 for Never
• Assign a score of 1 for Sometimes, Usually, or Always

Questions 6 and 7 are evaluated together as if they were a single question (Q67), the score of which is calculated as follows:
• Assign a score of 0 if Q6 = 0 AND Q7 = 0
• Assign a score of 1 if Q6 = 1 AND/OR Q7 = 1

The basic steps to calculate the measure score for a population are as follows:
• For each question, identify responses with non-missing values for that question
• For each respondent, calculate the proportion of responses with a score of 1 among all of the questions in the measure
• Calculate the numerator and denominator of the measure:
• Numerator = the sum of the proportions of positive responses among the questions in the measure for all respondents
• Denominator = the number of respondents with valid responses (i.e., non-missing values)

For each respondent, the proportion (P) of positive responses for the questions (Q) within the measure can be defined as follows:

P = (Q4 + Q5 + Q67 + Q8)/4
Measure score = (summation of values of P for N respondents/N)*100
Where N = the number of respondents with valid responses for P4, P5, P6, P7, and P8.

2) Counseling on prescription medication:

The measure is produced by combining responses to questions 10, 12, and 13:
• Q10: In the last 12 months, how often did you and this provider talk about all of your prescription medicines at each visit?
• Q12: In the last 12 months, did you and this provider talk about remembering to take your medicines?
• Q13: In the last 12 months, did you and this provider talk about you refilling your own prescriptions instead of your parent or guardian?

The 3 questions are scored as indicated in Figure 2 in Appendix A.

This measure score is calculated only for respondents who indicate on questions 9 ("in the last 12 months, did you take any prescription medicine?") and 11 ("in the last 12 months, were you prescribed any medicine to take every day for at least a month?") that they take prescription medication every day.

For each question, identify cases with non-missing values and for which the response for both question 9 and question 11 is "Yes":
• Respondents who do not report taking prescription medicine every day (responses of "No" to either questions 9 or 11) are not included in the population for which this measure is calculated

Response options for question 10 are "Never," "Sometimes," "Usually," or "Always"
• Assign a score of 0 for Never

• Assign a score of 1 for Sometimes, Usually, or Always

Response options for questions 12 and 13 are "Yes" or "No"
• Assign a score of 0 for No
• Assign a score of 1 for Yes

The basic steps to calculate the measure score for a population are as follows:
• For each question, identify responses with non-missing values for that question
• For each respondent, calculate the proportion of responses with a score of 1 among all of the questions in the measure
• Calculate the numerator and denominator of the measure:
• Numerator = the sum of the proportions of positive responses among the questions in the measure for all respondents
• Denominator = the number of respondents with valid responses (i.e., non-missing values)

For each respondent, the proportion (P) of positive responses for the questions (Q) within the measure can be defined as follows:

P = (Q10 + Q12 + Q13)/3
Measure score = (summation of values of P for N respondents/N)*100
Where N = the number of respondents with valid responses for P10, P12, and P13.

3) Transfer planning:

The measure is produced by combining responses to questions 15, 16, 17, and 18:
• Q15: In the last 12 months, did you and this provider talk about whether you may need to change to a new provider who treats mostly adults?
• Q16: In the last 12 months, did this provider ask if you had any questions or concerns about changing to a new provider who treats mostly adults?
• Q17: In the last 12 months, did you and this provider talk about a specific plan for changing to a new provider who treats mostly adults?
• Q18: Did this provider give you this plan in writing?

Only respondents who answer "Yes" or "Don't Know" to question 14 ("Does this provider treat mostly children and teens?") are included in the population for which this measure is calculated.
The 4 questions are scored as indicated in Figure 3 in Appendix A.
For each question, identify cases with non-missing values and for which the response for question 14 is "Yes," "Don't know," or left blank:
• Respondents who indicate the provider does not mostly treat children and teens (response of "No" to question 14) are not included in the population for which this measure is calculated

Response options for Questions 15-18 are "Yes" or "No." Valid responses for questions 16, 17, and 18 are provided by the respondent or assigned according to the decisions rules outlined in Appendix L.
• Assign a score of 0 for No
• Assign a score of 1 for Yes

The basic steps to calculate the measure score for a population are as follows:
• For each question, identify responses with non-missing values for that question
• For each respondent, calculate the proportion of responses with a score of 1 among all of the questions in the measure
• Calculate the numerator and denominator of the measure:
• Numerator = the sum of the proportions of positive responses among the questions in the measure for all

respondents
• Denominator = the number of respondents with valid responses (i.e. non-missing response OR assigned responses [see decision rules outlined in Appendix L])

For each respondent, the proportion (P) of positive responses for the questions (Q) within the measure can be defined as follows:

P = (Q15 + Q16 + Q17 + Q18)/4
Measure score = (summation of values of P for N respondents/N)*100
Where N = the number of respondents with valid responses for P15, P16, P17, and P18.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The sample for the ADAPT survey is drawn from pediatric patients ages 16-17 years old who have a chronic health condition and have seen a health care provider in the last 12 months. The measure is designed for completion by youth. The cover letter specifies that if a child is not able to understand the questions in the survey and answer them at all, the parent should not answer for the child. No proxy respondents are allowed.

SAMPLE FRAME CREATION
Clinical programs or health plans using the ADAPT survey are responsible for generating complete, accurate, and valid sample frame data files that contain all administrative information for each patient who meets the eligibility criteria. The minimum data elements for sample frame creation for the ADAPT survey are in Appendix E and the Data Dictionary.

The data elements that are most critical to the success of data collection are accurate and complete patient/member names, clinical program or health plan names, and home address.

De-duplication
Duplication of patients within the survey sample may occur if, for example, information for an eligible patient is received from multiple clinical programs within 1 hospital or practice setting. Perform de-duplication using the medical record number or health plan member identification number.

Sample size
The sample size goal for the survey should account for:
• The accuracy of patient/member home address
• The anticipated response rate based on prior surveys of the same or similar populations

SAMPLING PROCEDURE
For large practices, hospitals, or health plans, use Simple Random Sampling (SRS) to draw the desired final sample. To use SRS as the sampling method, randomly select the desired final sample size from all eligible patients. The chance that each patient will be selected is equal for all patients.

If using the PMCA[1] to identify chronic conditions, equally sized random samples should be drawn from the non-complex chronic disease (NC-CD) group and complex chronic disease (C-CD) group. The PMCA was used to identify children with chronic conditions in the health plan field test of this instrument.

Preparing sample files for survey administration
Once the sample has been selected, assign a unique survey identification number to each prospective respondent (sampled patient). This unique ID number should not be based on an existing identifier, such as a Social Security Number or a patient ID number. This number will be used only to track the respondents during data collection.

The sampling fraction of the total eligible population will vary depending on the overall size of the population. Some small clinical programs or health plans may not be able to obtain the minimum desired number of completed surveys. In such cases, sample all eligible patients or members in an attempt to obtain as many completed surveys as possible.

References:
1. Simon TD, Cawthon ML, Stanford S, Popalisky J, Lyons D, Woodcox P, Hood M, Chen AY, Mangione-Smith R, Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN) Medical Complexity Working Group. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. Pediatrics. 2014;133(6):e1647-1654. doi:10.1542/peds.2013-3875.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
The ADAPT survey can be used to assess the quality of transition preparation in a health plan or state Medicaid program, or as a tool for ongoing quality improvement in clinical programs. We have based our sample size recommendations on prior evaluations of widely used national patient experience surveys that have determined sample size requirements for adequate reliability.[1-3] For health plan or state Medicaid program comparisons, we recommend at least 300 completed surveys per health plan. By extension, we also recommend this sample size for comparisons of performance among large delivery systems (e.g., large multispecialty practices or hospitals with a number of outpatient programs for youth with chronic illness). Because response rates will vary among health plans and cannot be predicted with certainty, a conservative approach of aiming for slightly more than 300 completed surveys is recommended. The example in Appendix F shows the sample size calculation for a goal of 300 surveys for a health plan with a predicted response rate of 20 percent.

The ADAPT survey may also be used to assess performance for individual clinical programs. The number of responses for each administration will vary with the size of the available patient pool and the intended use. While further study is needed to determine the recommended sample size required for comparisons across programs, an individual program may use this measure over time to guide and assess improvement efforts. In general, the survey is not designed to measure or compare the performance of individual health care providers.

MAIL PROTOCOL
This section lists recommended steps for administering the survey by mail.
• Set up a toll-free number (or use an existing information line) to include in all correspondence with prospective respondents. Train staff members to respond to questions. Maintain a log of these calls and review them periodically for common issues that arise.
• Mail the survey addressed to the parent/guardian of the prospective respondents with a cover letter and a postage-paid envelope. The cover letter should include instructions for the adolescent patient to complete and return the survey. For examples, see English Mailed Survey Materials (Appendix G) and Spanish Mailed Survey Materials (Appendix H).

Tips for the cover letter:
> Personalize the letter with the name and address of the intended recipient (parent/guardian).
> Tailor the letter to include language that explains the purpose of the survey, the voluntary nature of participation, and the confidentiality of responses.
> Include language in the letter that asks the parent or guardian to give the survey to their adolescent child.
> Indicate that if the adolescent child is unable to complete the survey independently (e.g., due to developmental

delay), then the survey should not be completed. Include a check box in the cover letter for the parent/guardian to indicate that the identified child is unable to complete the survey, and instruct the parent or guardian to return the blank survey and cover letter for tracking.
> Note that non-participation will not affect the health care of either the parent/guardian or the adolescent child.
> Have the letter signed by a representative of the clinical program or health plan.
> Confirm that the reading level of the cover letter is appropriate for the population and meets all applicable regulatory requirements.

Tips for the outside envelope:
> Make the envelope look "official" but not bureaucratic or like "junk mail."
> Place a recognizable sponsor's name above the return address.
> Mark the envelopes "change of service requested" in order to receive information to update records for respondents who have moved and to increase the likelihood that the survey will reach the intended respondent.

• Maintain a database of returned surveys by unique survey identifier. Each prospective respondent in the response tracking system should be assigned a survey result code that indicates whether he or she completed and returned the survey, was ineligible to participate in the study, could not be located, or refused to participate.
• Send a second survey 3 weeks after the initial mailing. To avoid mailing another survey to those who have already responded, finish entry of returned surveys into the database before mailing second surveys. Include in the second mailing a slightly adapted reminder letter to those parents whose adolescent children have not responded to the first mailing and another postage-paid return envelope. Examples of the reminder letter can be found in the English Mailed Survey Materials (Appendix G) and Spanish Mailed Survey Materials (Appendix H).
• Close data collection 10 weeks from the first survey mailing.

References:
1. Lotstein DS, McPherson M, Strickland B, Newacheck PW. Transition planning for youth with special health care needs: results from the National Survey of Children with Special Health Care Needs. Pediatrics. 2005;115(6):1562-1568. doi:10.1542/peds.2004-1262
2. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians, Transitions Clinical Report Authoring Group, Cooley WC, Sagerman PJ. Supporting the health care transition from adolescence to adulthood in the medical home. Pediatrics. 2011;128(1):182-200. doi:10.1542/peds.2011-0969.
3. Chua K-P, Schuster MA, McWilliams JM. Differences in health care access and utilization between adolescents and young adults with asthma. Pediatrics. 2013;131(5):892-901. doi:10.1542/peds.2012-2881.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Decision Rules for Transfer Planning Measure Questions 16, 17, and 18
If Question 15 is "No", then code Questions 15, 16, 17, and 18 as "No."
If Question 15 is "Yes" or left blank, enter the value provided by the respondent for Questions 16, 17, and 18, except as follows:
• If Question 17 is "No" and Question 18 is left blank or not left blank, then code the value of Question 18 as "No"
• If Question 17 is left blank and Question 18 is not left blank, then code Question 17 as ".Missing" and enter the value provided by the respondent for Question 18.

Decision Rules for Screener and Dependent Questions
Decision rules for coding screener questions (Questions 3, 9, 11, 14, and 25; Does not apply to Question 15 or 17):
• Enter the value provided by the respondent. Do not impute a response based on the respondent's answers to the dependent questions.
• If a screener question is left blank, then code the value as ". Missing." Do no impute a response based on the respondent's answers to the dependent questions.
• In the situation where more than one option is marked for a screener question, see rules in the "Coding Ambiguous Responses" section.

Decision rules for coding dependent questions (Questions 4-14, and 26; Does not apply to Questions 15-18):
• If the marked screener question option requires the dependent question(s) to be answered, and the dependent question(s) is left blank, then code the value for the dependent questions(s) as ". Missing."
• If the marked screener question option requires the dependent question(s) to be answered, and the dependent question(s) is not left blank, then enter the value provided by the respondent for the dependent question(s).
• If the marked screener question option requires the dependent question(s) to be skipped, and the dependent question(s) is left blank, then code the value for the dependent questions(s) as ". Missing."
• If the marked screener question option requires the dependent question(s) to be skipped, and the dependent question(s) is not left blank, then code the value for the dependent questions as ". Missing."
• If the screener question is left blank and the dependent question(s) is left blank, then code the value for both the corresponding screener question and the dependent question(s) as ". Missing."
• If the screener question is left blank and the dependent question(s) is not left blank, then code the value for the corresponding screener question as ". Missing" and enter the value provided by the respondent for the dependent question(s).

As detailed in the scoring algorithm (Section S.18 above), any respondents with missing data to any question within a particular measure is not included in the population used to calculate the measure score.

S.23. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Patient Reported Data/Survey

S.24. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care Survey.

The ADAPT survey is available in English and Spanish. The recommended mode of administration is by mail. For a detailed explanation of survey administration modes, see S.21 – Survey/Patient Reported Data.

S.25. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility, Health Plan

S.27. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_ADAPT_Measure_Testing_form-635792195014185602.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care
Date of Submission: 9/30/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-**

PMs and composite performance measures, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score

include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: ADAPT survey	☒ other: ADAPT National Field Test Dataset

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
Not applicable

1.3. What are the dates of the data used in testing? 2013-2014

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.26*)**	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
We conducted national field testing of the ADAPT survey in three settings: a large freestanding pediatric hospital in Massachusetts, which we refer to in our submission as Hospital 1, and 2 health plans serving Medicaid enrollees, which we refer to as Health Plan 1 and Health Plan 2. Health Plan 1 is a Medicaid managed care health plan serving individuals across 2 regions in Pennsylvania, while Health Plan 2 is a pediatric-focused Medicaid health plan serving individuals in Texas.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
The target population for the ADAPT survey is 16- to 17-year-old adolescents with chronic health conditions. For the Hospital 1 field test, we mailed surveys to outpatients in this age group with a wide variety of chronic illnesses receiving care in 10 different clinical programs (Endocrinology, Gastroenterology, Hematology-Sickle Cell and Hemophilia, Immunology, Metabolism, Nephrology, Primary Care, Pulmonology, and Spina Bifida). We received a total of293 surveys.

For the Health Plan 1 and Health Plan 2 field tests, survey recipients were identified by analysis of health plan claims using the Pediatric Medical Complexity Algorithm (PMCA).[1] This publicly available algorithm uses International Classification of Diseases, Ninth Revision, Clinical Modification diagnosis codes in health plan claims to identify youth with either complex chronic disease (C-CD) or noncomplex chronic disease (NC-CD). The survey was fielded in both

English and Spanish. We received a total of 1,339 surveys (780 from Health Plan 1 and 575 from Health Plan 2).

Appendix K shows descriptive characteristics of the respondents included in our analysis. Female respondents outnumbered males in all three samples. Approximately 40%-45% of respondents in each sample were 16 years old, while the remaining respondents were 17 years old. The samples were diverse in race/ethnicity. Among the hospital respondents, 29% were insured by Medicaid, as were all respondents in the 2 health plan samples. Of note, all of the samples included individuals with a broad range of self-reported health status; 40% or more of each sample reported their overall health as only good, fair, or poor.

References:

1. Simon TD, Cawthon ML, Stanford S, et al, Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN) Medical Complexity Working Group. Pediatric medical complexity algorithm: A new method to stratify children by medical complexity. *Pediatrics* 2014; 133:e1647e54.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
Not applicable

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
The ADAPT survey includes the collection of respondent age, race/ethnicity, and education. Differences in ADAPT survey scores based on race/ethnicity can be seen in ***Appendix J***. Analyses according to age are presented as part of the case mix adjustment model. The survey does not include any additional assessment of respondent sociodemographic status (SDS).

2a2. RELIABILITY TESTING
Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Internal consistency reliability: Internal consistency reliability is a measure of the degree of consistency of responses to different questions intended to measure the same construct. Of the available statistical indicators for internal consistency, the ordinal reliability coefficient (ordinal alpha), which uses a polychoric correlation matrix, was determined to be most appropriate for questions with dichotomous responses. For questions with few response categories, the ordinal indicator more accurately estimates reliability compared to the more commonly used Cronbach's alpha.[1]

In order to summarize performance on processes associated with the quality of transition preparation, we developed measure scores that incorporate multiple individual survey questions (see ***Appendix A, Detailed Measure Specifications***) from three measures: (1) Counseling on Transition Self-Management, (2) Counseling on Prescription Medication, and (3) Transfer Planning. Each measure score was designed to measure a single underlying construct of transition preparation. The ordinal alpha provides reliability results for all measures. In general, internal consistency reliability of .70 or greater is desirable.

References:

1. Gadermann AM, Guhn M, Zumbo BD. Estimating Ordinal Reliability for Likert-Type and Ordinal Item Response Data: A Conceptual, Empirical, and Practical Guide. Pract Assess Res Eval. 2012;17(3).

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The ordinal alpha is provided for each of the measures in each of the three field test sites (Table 2a2.3.a). All measures in all sites had an internal consistency of .7-.8, with the exception of a single measure in 1 site.

Table 2a2.3.a: Internal consistency reliability for ADAPT survey measures by site

	Hospital 1	Health Plan 1	Health Plan 2
	Ordinal alpha		
Counseling on Transition Self-Management	0.79	0.70	0.78
Counseling on Prescription Medication	0.57	0.78	0.74
Transfer Planning	0.99	0.99	0.99

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

In general, our results indicate that internal consistency reliabilities for our measures are good to excellent. Furthermore, although the test sites in our field testing varied in their geographic location and demographic characteristics, measure scores and responses to individual questions were similar across the three field tests.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

In contrast with some types of quality measures, a "gold standard" does not exist for determining the criterion validity of patient-reported measures of quality.[1] However, to ensure the validity of the ADAPT survey results, we followed rigorous procedures representing best practices within the field to develop the survey. To ensure the content validity of measures of the transition experience from patients' perspectives, we used qualitative methods, including both focus groups and cognitive interviews, to inform development of the survey questions. We used quantitative methods, including confirmatory factor analysis, question-to-measure correlations, and measure-to-measure correlations, to evaluate the validity of the final survey.

Focus groups and cognitive interviews: We conducted focus groups and cognitive interviews early in the survey development process to ensure that the instrument covered topics of greatest importance to adolescent patients and their parents or guardians.[1] In total, we conducted 11 focus groups in Boston, Chicago, and Los Angeles: three with adolescents, four with young adults, and four with parents/guardians. One of the young adult groups and 2 of the parent/guardian groups consisted of participants whose primary language was Spanish, and the focus groups were conducted in this language. The focus groups included a diverse spectrum of patients with regard to gender, race, ethnicity, and type of chronic health condition. In addition, we performed four rounds of 26 total cognitive interviews of youth respondents in English and Spanish in Boston, Chicago, and Dallas. See Measure Submission Form 1c.5 for a description of methods used for focus groups and cognitive interviews

Factor analysis: Because the measures of the ADAPT survey and their associated questions were pre-defined, the validation of the measures is most appropriately performed through confirmatory factor analysis (CFA). In addition, since the questions (items) in these measures (factors) were designed with dichotomous responses, tetrachoric correlation coefficients were determined to be most appropriate for assessing the pairwise correlations among the measure questions.[2,3] CFA was performed only for the first 2 ADAPT measures because the sample sizes for the Transfer Planning measure were inadequate to conduct CFA. Mplus (Statistical Analysis With Latent Variables) software was used to conduct the CFA for each site.

References:

1. Sawicki GS, Garvey KC, Toomey SL, Williams KA, Chen Y, Hargraves JL, Leblanc J, Schuster MA, Finkelstein JA. Development and Validation of the Adolescent

Assessment of Preparation for Transition: A Novel Patient Experience Measure. J Adolesc Health. 2015;57(3):282-287. doi:10.1016/j.jadohealth.2015.06.004.
2. Muthén, L.K. and B.O. Muthén, Mplus User's Guide. Seventh Edition. 1998-2012, Los Angeles, CA: Muthén & Muthén.
3. Brown, T.A., Confirmatory Factor Analysis for Applied Research. 2006, New York: Guilliard Press.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Factor analysis: The standardized solutions for the CFA 2-factor models measuring independence of each measure are included in Table 2b2.3.a. In each site, the p-values of the loading factor estimates within each measure demonstrate that questions are strongly associated with their hypothesized construct. In addition, the association between the 2 constructs in all three sites is also significant.

The fit statistics for each of the three sites are presented in Table 2b2.3.b. In 2 sites, the p-value of the chi-square test of fit was <.05, indicating that the observed covariance matrix is statistically significantly different from the expected matrix predicted by the hypothesized model; however, the chi-square test is sensitive to sample size and therefore is not the only test of fit considered. In general, the other fit statistics are adequate across the three sites.

Table 2b2.3.a: Confirmatory factor analysis to evaluate the fit of a 2-factor model to the ADAPT data across samples

Hospital 1

Variable	Factor Loading Estimate	S.E.	Two-tailed T-test	P-value
Counseling on Transition Self-Management				
Q4	0.516	0.097	5.300	<.001
Q5	0.594	0.090	6.615	<.001
Q67	0.561	0.112	5.027	<.001
Q8	0.665	0.130	5.128	<.001
Counseling on Prescription Medication				
Q10	0.165	0.108	1.527	.127
Q12	0.463	0.112	4.14	<.001
Q13	0.826	0.16	5.163	<.001

Health Plan 1 Model

Variable	Factor Loading Estimate	S.E.	Two-tailed T-test	P-value
Counseling on Transition Self-Management				

Q4	0.332	0.075	4.442	<.001
Q5	0.480	0.076	6.306	<.001
Q67	0.694	0.093	7.489	<.001
Q8	0.551	0.114	4.809	<.001
Counseling on Prescription Medication				
Q10	0.600	0.080	7.503	<.001
Q12	0.673	0.084	7.968	<.001
Q13	0.576	0.089	6.471	<.001

Health Plan 2 Model

Variable	**Factor Loading Estimate**	**S.E.**	**Two-tailed T-test**	**P-value**
Counseling on Transition Self-Management				
Q4	0.527	0.092	5.702	<.001
Q5	0.753	0.1	7.515	<.001
Q67	0.447	0.105	4.269	<.001
Q8	0.469	0.113	4.152	<.001
Counseling on Prescription Medication				
Q10	0.594	0.11	5.414	<.001
Q12	0.643	0.11	5.851	<.001
Q13	0.408	0.119	3.428	.001

Table 2b2.3.b: Goodness of fit measures for CFA

	Hospital 1	**Health Plan 1**	**Health Plan 2**
Chi-square test of fit p-value	0.013	<0.001	0.244
Root mean squared error of approximation RMSEA (90% CI)	0.064 (0.028, 0.098)	0.081 (0.061, 0.103)	0.026 (0, 0.062)
Comparative Fit Index (CFI)	0.892	0.792	0.974
Tucker Lewis Index (TLI)	0.826	0.664	0.958

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e.*, *what do the results mean and what are the norms for the test conducted?*)

Focus groups and cognitive interviews: The focus groups and cognitive interviews generally

confirmed the understandability of the questions' intended meaning, question construction, survey administration process, and skip patterns.[1] The analysis of the cognitive interview data resulted in simplification and/or clarification of some survey questions, refinement of skip patterns, and deletion of questions that were not clear to respondents and deemed to be less essential to assessing transition preparation than originally hypothesized. Analysis of the national field test results led to additional small revisions in survey wording. These minor changes were then tested in an additional round of cognitive interviews in Boston with six 16- to 17-year-old adolescents with chronic health conditions. These interviews confirmed the understandability of each question in the final ADAPT survey and that no additional changes were needed.

Factor analysis: The goal of the CFA was to test the construct validity of the survey using a 2-factor structure for including (1) Counseling on Transition Self-Management (4 questions – 2 levels) and (2) Counseling on Prescription Medication (3 questions – 2 levels). Results from these analyses supported the hypothesis that the individual questions within each of the 2 measures are associated with one another. CFA results were similar across the three sites, providing further confirmation that questions grouped together on conceptual grounds are also empirically related.

References:

1. Sawicki GS, Garvey KC, Toomey SL, Williams KA, Chen Y, Hargraves JL, Leblanc J, Schuster MA, Finkelstein JA. Development and Validation of the Adolescent Assessment of Preparation for Transition: A Novel Patient Experience Measure. J Adolesc Health. 2015;57(3):282-287. doi:10.1016/j.jadohealth.2015.06.004.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 2 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

When comparing clinical programs or health plans, it may be appropriate to adjust for case-mix differences. Case-mix refers to patient characteristics, such as demographic characteristics and health status, that are not under the control of the clinical program/health plan and may affect scores on performance measures.[1] Systematic effects of this sort create the potential for a clinical program's/health plan's unadjusted score to be higher or lower because of characteristics of its patient population rather than the quality of care it provides. Comparisons of unadjusted scores may therefore be misleading. The basic goal of adjusting for case-mix is to estimate how different clinical programs/health plans would score if they all provided care to the same mix of patients.

To evaluate potential variables for case-mix adjustment of ADAPT scores, we started with the case-mix variables used for the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS (age, health status, education, and language preference). We evaluated age, self-reported health status, gender, and education using data obtained from ADAPT survey questions and evaluated health condition type (Complex Chronic vs. Non-Complex Chronic) as determined using the Pediatric Medical Complexity Algorithm (PMCA)[2] with administrative data from the health plan test sites. We did not evaluate language because only 46 total Spanish surveys were completed in the field test.

We assessed a series of multivariate linear regression models predicting various outcomes. These models controlled for clinical programs/health plans to isolate which characteristics affect care within clinical programs/health plans rather than simply being more concentrated in certain clinical programs / health plans. Scores for each of 11 ADAPT questions (4, 5, 6&7, 8, 10,

12, 13, 15, 16, 17, 18) were modeled as the dependent variable in a model with each of the core set of adjusters (Complex Chronic vs. Non-Complex Chronic health condition derived from the PMCA and ADAPT questions Q21 [self-reported health status], Q19 [age], Q20 [gender], Q22 [education]) as independent variables (Table 2b4.4a below). A distribution of the strength of association for each adjuster with each outcome was compiled. Adjusters that had stronger associations with a greater number of outcomes were interpreted as having a more substantial impact on patient experience.

Next, to evaluate the variation of the core set of demographic adjusters (age, gender) among adolescents in a broader population, we assessed data from the Medicaid Analytic Extract (MAX) 2008 person file. These data are available to researchers from the Centers for Medicare and Medicaid Services. Deidentified outpatient claims data from 9 states (Arizona, Indiana, Kansas, Kentucky, Missouri, New Jersey, New Mexico, Virginia, Wisconsin) were used. We examined variation across counties in the proportion of subjects who were 16 versus 17 years old and the proportion of subjects who were male versus female, using the federal information processing standard (FIPS) code, which indicates the eligible person's county of residence, as the county code. In this dataset, we found no variation across counties in age 16 versus 17 or in gender. We did find variation by county based on medical complexity.

References:

1. O'Malley AJ, Zaslavsky AM, Elliott MN, Zaborski L, Cleary PD. Case-Mix Adjustment of the CAHPS(R) Hospital Survey. *Health Serv Res*. 2005;40(6 Pt 2):2162-2181. doi:10.1111/j.1475-6773.2005.00470.x.
2. Simon TD, Cawthon ML, Stanford S, et al, Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN) Medical Complexity Working Group. Pediatric medical complexity algorithm: A new method to stratify children by medical complexity. *Pediatrics* 2014; 133:e1647e54.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Potential Adjusters: Complex Chronic vs. Non-Complex Chronic Health Condition (only evaluated for Health Plans 1 and 2). The rest include Health Plan 1, 2 and Hospital 1: Self-reported health status [Q21], age [Q19], gender [Q20], education [Q22] (evaluated for Health Plans 1 and 2 and Hospital 1). Language was not included because there only 18 Spanish surveys of 780 patients from Health Plan 1 and only 28 of 575 surveys from Health Plan 2 were administered in Spanish. Age and education were highly correlated (0.68; 95% CI 0.64-0.72) based on a polychoric correlation, and age was considered a more appropriate adjuster for the ADAPT scores.

Table 2b4.4a. ADAPT Strength of association for individual questions*

	Complex Chronic	Self-Reported Health Status [Q21]	Age (16 versus 17) [Q19]	Male Gender [Q20]	Education [Q22]

p<.001			1		5
.001≤p<.01		1			0
.01≤p<.05		3	4	2	1
p≥.05	11	7	6	9	5

* The number of 11 ADAPT question models with the p-value range for the associated independent adjuster.

Results for candidate adjusters (e.g. gender, education) that were considered but not retained in the final case-mix model were are included in the table as examples of adjusters that were considered but rejected. P-values are included for the overall (Type III) association of the adjuster with each of outcomes in bivariate models that included each of the selected current case-mix adjusters.

Results from the evaluation methods were combined to determine the final set of case-mix adjusters. Based on this analysis, we recommend use of the following two categorical variables in the ADAPT case-mix adjustment model: self-reported health status and age.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
Age was included in our CMA (along with self-reported health status). See above (2b4.4a) for analyses and interpretation.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
R-squared values were calculated to assess the fit of the final case-mix model for each of the 11 ADAPT questions.

Case-mix adjusted models using our selected covariates of health status and age were created for each of the 11 ADAPT items. The R-squared values associated with these models were used to assess the fit of the case-mix model.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Adjusted R-squared values were determined for models containing each of the 11 ADAPT questions as independent variables and self-reported health status and age as independent variables (case-mix adjusters). The median adjusted R-squared was 0.0034 (range -0.0011 to 0.0145; 25th percentile 0.0022, 75th percentile 0.0101). The adjusted R-squared increases when a new independent variable is included only if the new variable improves the R-squared more than would be expected by chance. Therefore, it is useful to also consider the R-squared

(unadjusted) in our case. The median R-squared (unadjusted) was 0.0057 (range 0.0002 to 0.0158; 25th percentile 0.0036, 75th percentile 0.0115).

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable

2b4.9. Results of Risk Stratification Analysis:
Not applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
In many analyses, the goal is to explain as much of the variance as possible, in which case a high R-squared is desired. In this case, the value of the R-squared represents the extent to which case-mix adjustment affected measure scores. For example, if the case-mix adjusters had no effect (e.g., age was not predictive of measure scores), then the R-squared value would be zero. Overall, case-mix adjustments had only minimal effects on clinical program/health plan ADAPT measure scores.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
Differences in population-level scores were compared using t-tests and f-tests based on the case-mix adjustment model estimates.

Statistically significant differences in performance are assessed for the case-mix adjusted scores for measures by examining whether the three sites were different. A t-test of means was used for comparing between the 2 health plan sites. An f-test of means was used for comparing across the three sites. A level of alpha error of $p < .05$ was set as the criterion for significance.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how*

was meaningful difference defined)

Table 2b5.2: Case-Mix Adjusted Measure Scores (CMADS) across samples

	Hospital 1 (n=293)		Health Plan 1 (n=780)		Health Plan 2 (n=575)		*P* 3 sites	*P* Health Plan 1 versus 2
	n	CMADS	n	CMADS	n	CMADS		
Counseling on Transition Self-Management	266	32 (30, 35)	707	36 (34, 38)	489	30 (28, 33)	0.028	0.024
Counseling on Prescription Medication	237	61 (59, 64)	426	57 (55, 60)	209	58 (54, 62)	0.267	0.075
Transfer Planning	266	5 (3, 7)	704	4 (3, 5)	489	3 (2, 4)	0.225	0.158

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e.*, *what do the results mean in terms of statistical and meaningful differences?*)

Overall, population level scores in all three measures of the ADAPT survey were low in all three populations included in the national ADAPT field test. These low scores are consistent with other national findings on measures of transition readiness as reported by parents of adolescents or young adults.[1-4] The small differences seen in the Counseling on Transition Self-Management scores are not clinically meaningful given the overall low performance across all three sites. As health systems develop efforts to improve transition planning for their adolescent populations, we anticipate ADAPT scores will improve to variable degrees based on the extent to which these quality improvement efforts succeed. This will allow for identification of clinically meaningful differences in performance across entities using ADAPT as an adolescent-reported measure of transition preparation.

References:

1. American Academy of Pediatrics, American Academy of Family Physicians, American College of Physicians-American Society of Internal Medicine. A consensus statement on health care transitions for young adults with special health care needs. *Pediatrics*. 2002;110(6 Pt 2):1304-1306.
2. Fortuna RJ, Robbins BW, Halterman JS. Ambulatory care among young adults in the United States. *Ann Intern Med*. 2009;151(6):379-385.
3. Nakhla M, Daneman D, To T, Paradis G, Guttmann A. Transition to adult care for youths with diabetes mellitus: findings from a Universal Health Care System. *Pediatrics*. 2009;124(6):e1134-1141. doi:10.1542/peds.2009-0041.
4. Goodman DM, Mendez E, Throop C, Ogata ES. Adult survivors of pediatric illness: the impact on pediatric hospitals. *Pediatrics*. 2002;110(3):583-589.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e.*, *what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

ADAPT optimizes the quality of responses by using several screening questions to direct respondents through survey skip patterns. The screening questions result in a high percentage of appropriately missing data due to appropriate skips, as some questions are not relevant to respondents based on their care plan or their previous interactions with the provider. Survey item screeners have been found to reduce noise by ensuring that respondents who are not "qualified" to answer a question are screened out instead of providing unreliable responses.[1]

We calculated the number of surveys with truly missing responses (i.e., missing for reasons other than being left blank appropriately because of screener items) for Hospital 1 and for Health Plans 1 & 2 for each question related to the 3 domains: 1) Counseling on Transition Self-Management, 2) Counseling on Prescription Medication, and 3) Transfer Planning (Table 2b7.2.a).

In addition, we performed a clinical program- and health plan-level analysis comparing respondents and non-respondents demographic based on characteristics and medical complexity. ***Appendix K*** includes detailed information on ADAPT Survey field test respondent characteristics.

References:

1. Rodriguez HP, Glahn T von, Li A, Rogers WH, Safran DG. The effect of item screeners on the quality of patient survey data: a randomized experiment of ambulatory care experience measures. *The patient*. 2009;2(2):135-141. doi:10.2165/01312067-200902020-00009.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Table 2b7.2a: Missing ADAPT survey item responses by site

	Hospital 1 n (%)	Health Plan 1 n (%)	Health Plan 2 n (%)
	n=293	n=780	n=575
Counseling on Self-Management			
In the last 12 months, did you talk with this provider without your parent or guardian in the room?	1 (0.3)	7 (0.9)	7 (1.2)
In the last 12 months, did you and this provider talk about you being more in charge of your health?	1 (0.3)	9 (1.2)	4 (0.7)
In the last 12 months, did you and this provider talk about you scheduling your own appointments with this provider instead of your parent or guardian?	1 (0.3)	6 (0.8)	4 (0.7)
In the last 12 months, how often did you schedule your own appointments with this provider?	0 (0.0)	12 (1.5)	4 (0.7)
In the last 12 months, did you and this provider talk about how your health insurance might change as you get older?	1 (0.3)	12 (1.5)	8 (1.4)
Counseling on Prescription Medications			
In the last 12 months, did you and this provider talk about all of your prescription medicines at each visit?	0 (0.0)	14 (1.8)	7 (1.2)

In the last 12 months, did you and this provider talk about remembering to take your medicines?	1 (0.3)	15 (1.9)	8 (1.4)
In the last 12 months, did you and this provider talk about you refilling your own prescriptions instead of your parent or guardian?	0 (0.0)	16 (2.0)	7 (1.2)
Transfer Planning			
In the last 12 months, did you and this provider talk about whether you may need to change to a new provider who treats mostly adults?	2 (0.7)	20 (2.6)	11 (1.9)
In the last 12 months, did this provider ask if you had any questions or concerns about changing to a new provider who treats mostly adults?	1 (0.3)	15 (1.9)	10 (1.7)
In the last 12 months, did you and this provider talk about a specific plan for changing to a new provider who treats mostly adults?	2 (0.7)	16 (2.0)	10 (1.7)
Did this provider give you this plan in writing?	1 (0.3)	16 (2.0)	11 (1.9)

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias? (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Respondents and non-respondents were generally similar in all three samples. Compared to non-respondents, there was a higher proportion of 17 year-old adolescents in the Health Plan 1 respondent sample only ($p<.05$). There were lower proportions of black patients in the respondent samples compared to non-respondents in the Hospital 1 sample (5% vs. 12%) and Health Plan 1 (24% vs. 35%) (both $p<.01$), but the proportion of Hispanic adolescents among respondents and non-respondents was similar in all three sites.

There was a higher response rate among females in the hospital sample (52% vs. 44%, $p = .02$) but not in the two health plans. There was a higher response rate among 17-18 year olds only in Health Plan 1 (28% vs. 24%, $p = .04$) but not in the other two sites. There were significant differences in response by race/ethnicity only in the hospital site driven by a lower response rate among blacks. There was no statistically significant difference in response rate based on chronic condition complexity category. Overall, therefore, there does not appear to be a systematic bias in response based on these demographic factors.

For all three sites, all questions had less than 3% of cases as truly missing, which suggests that question-level results are unlikely to be biased by systematic missing data due to question non-response. The mean percentage missing was 1.3% and ranged from 0.67% for scheduling own appointments to 2.00% for discussing whether there is a need to change to a new provider who treats mostly adults.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Other
If other: Collected via survey completed by youth

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**
The ADAPT survey is administered by mail. The rationale for not using electronic sources (e.g., administration by email) is that mail and telephone administration are the best ways to obtain representative samples of patients based on the contact information (mailing address and telephone number) that is most often available for sampling and data collection.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
As previously described (see Section 1c.5), we have used an iterative process to ensure that the survey is understandable to patients and their families.[1] Additionally, the survey can be completed easily within a short time period (approximately 10 minutes or fewer) and therefore is minimally burdensome to respondents. Through

our field testing, we learned that it is feasible for a health plan to pull the data necessary to develop a survey sample frame and that it is feasible to obtain survey responses from adolescents.

References:
1. Sawicki GS, Garvey KC, Toomey SL, Williams KA, Chen Y, Hargraves JL, Leblanc J, Schuster MA, Finkelstein JA. Development and Validation of the Adolescent Assessment of Preparation for Transition: A Novel Patient Experience Measure. J Adolesc Health. 2015;57(3):282-287. doi:10.1016/j.jadohealth.2015.06.004.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
The ADAPT survey is available to users free of charge.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Not applicable

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

The ADAPT survey is a newly developed measure for which we only recently completed national field testing. We are not aware of any restrictions on access to performance results or impediments to implementation that would prevent ADAPT from being used in public reporting or other accountability applications.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
We intend that the measures be available for public use. For ease of implementation, we have prepared Detailed Measure Specifications (Appendix A). In addition, we have given a series of conference presentations and invited webinars on the development, testing, and use of ADAPT.

The ADAPT survey measure is not currently used for public reporting. Endorsement will facilitate use of ADAPT by public and private payers, provider organizations, and consumer groups that require NQF endorsement of quality measures and will help support the integration of the survey into other quality measure sets. We anticipate that the ADAPT survey could be included as a supplement to collection of other patient experience measures. ADAPT results will be useful to everyone with a need for information on the quality of care for adolescents with chronic conditions, including patients, parents, hospitals, health plans, insurers, and policy makers. The survey-based measures will identify areas for quality improvement for outpatient settings and could be used to evaluate performance against benchmarks.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not applicable

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No unintended negative consequences were identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0005 : CAHPS Clinician & Group Surveys (CG-CAHPS)-Adult, Child

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
ADAPT was developed with similar principles to CG CAHPS. If administered following a health care visit for an adolescent, the CG CAHPS survey is intended to be completed by parents of an adolescent as opposed to the adolescents themselves. However, both surveys target the outpatient care setting experience. The ADAPT survey complements the CG CAHPS survey well and has the potential to be administered concurrently, with both surveys mailed to the patient residence so that parents can complete the CG CAHPS survey and adolescents can complete the ADAPT survey.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** NQF_ADAPT_Appendix-635792196694015138.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Center of Excellence for Pediatric Quality Measurement
Co.2 **Point of Contact:** Mark, Schuster, MD, PhD, cepqm@childrens.harvard.edu, 617-355-5859-
Co.3 **Measure Developer if different from Measure Steward:** Center of Excellence for Pediatric Quality Measurement
Co.4 **Point of Contact:** Mark, Schuster, MD, PhD, cepqm@childrens.harvard.edu, 617-355-5859-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The core team from Boston Children's Hospital and University of Massachusetts Center for Survey Research:
Yuefan Chen, MS
Carol A. Cosenza, MSW
Jonathan A. Finkelstein, MD, MPH (Measure Co-Lead)
Alison A. Galbraith, MD, MPH
Katharine C. Garvey, MD, MPH
Shannon C. Hardy, BA
J. Lee Hargraves, PhD
Chelsea K. Johnson, BA
Jessica L. LeBlanc, BA
Lindsey L. Mahoney, BS
Mari M. Nakamura, MD, MPH
Jessica A. Quinn, MS
Gregory S. Sawicki MD, MPH (Measure Co-Lead)
Mark A. Schuster, MD, PhD (CEPQM Principal Investigator and Director)
Shanna Shulman, PhD
Cassandra J. Thomson, AB
Sara L. Toomey, MD, MPhil, MPH, MSc (CEPQM Managing Director)
Kathryn A. Williams, MS

Key collaborators from AmeriHealth Caritas Pennsylvania:
Wanzhen Gao, MD, PhD
Thomas James III, MD
Susan Tan-Torres, MD, MPH

Key collaborators from Texas Children's Health Plan:
Angelo P. Giardino, MD, PhD
Jean L. Raphael, MD, MPH

Xuan G. Tran, MHA Christopher C. Williams, MS, MBA Staff of the Center of Excellence for Pediatric Quality Measurement (CEPQM) at Boston Children's Hospital Members of CEPQM's Scientific Advisory Board and National Stakeholder Panel Members of the Massachusetts Child Health Quality Coalition Members of the Boston Children's Hospital Teen Advisory Council
Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** Ad.3 **Month and Year of most recent revision:** Ad.4 **What is your frequency for review/update of this measure?** Ad.5 **When is the next scheduled review/update for this measure?**
Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

APPENDIX

Measuring the Preparation for Transition from Pediatric-Focused to Adult-Focused Health Care:

The Adolescent Assessment of Preparation for Transition (ADAPT) Survey

NQF # 2789

Appendix

Center of Excellence for Pediatric Quality Measurement

Division of General Pediatrics

Boston Children's Hospital

September 2015

Core team:

Yuefan Chen, MSc

Jonathan A. Finkelstein, MD, MPH (Measure Co-Lead)

Katharine C. Garvey, MD, MPH

J. Lee Hargraves, PhD

Gregory S. Sawicki, MD, MPH (Measure Co-Lead)

Mark A. Schuster, MD, PhD (Principal Investigator and Director)

Sara L. Toomey, MD, MPhil, MPH, MSc (Managing Director)

Kathryn A. Williams, MStat

Funding: Support for this work was provided by the U.S. Department of Health and Human Services Agency for Healthcare Research and Quality and Centers for Medicare & Medicaid Services, CHIPRA Pediatric Quality Measures Program Centers of Excellence under grant number U18 HS 020513. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Table of Contents

Acknowledgments

We would like to thank the following people who participated in the development of the ADAPT Survey:

Additional team members from Boston Children's Hospital and the Center for Survey Research, University of Massachusetts, Boston:

Carol A. Cosenza, MSW

Shannon C. Hardy, BA

Isabel Janmey, BA

Chelsea K. Johnson, BA

Jessica L. LeBlanc, BA

Lindsey L. Mahoney, BS

Mari M. Nakamura, MD, MPH

Jessica A. Quinn, MS

Shanna Shulman, PhD

Cassandra J. Thomson, AB

Partners from AmeriHealth Caritas Pennsylvania and Texas Children's Health Plan

Members of Boston Children's Hospital Transition Measure Advisory Committee

Staff of the Center of Excellence for Pediatric Quality Measurement (CEPQM) at Boston Children's Hospital

Members of CEPQM's Scientific Advisory Board and National Stakeholder Panel

Members of the Massachusetts Child Health Quality Coalition

Members of the Boston Children's Hospital Teen Advisory Council

We thank the participants in our focus groups, cognitive interviews, and field tests and all others who contributed to the development and testing of the ADAPT survey.

Appendix A. Detailed Measure Specifications

ADolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care

Introduction

Overview of development of ADAPT survey

Generation of sample frame

- Eligibility guidelines
- Identification of youth with chronic health conditions
- Exclusions
- Sample creation
- De-duplication
- Sample size
- Sampling procedure
- Preparing sample for survey administration

Data collection protocol

- Mail protocol

Calculation of response rate

- Criteria for completeness
- Numerator
- Denominator

Data cleaning protocols

Calculation of measure scores

Counseling on Transition Self-Management

Counseling on Prescription Medication

Transfer Planning

Introduction

This document explains how to administer, analyze, and calculate scores from the ADolescent Assessment of Preparation for Transition (ADAPT) survey in a sample derived from either (a) a primary care or specialty practice in a hospital or community setting (hereafter referred to as a clinical program) or (b) a defined population of covered individuals (e.g., health plan, accountable care organization). A version of the survey was developed for each of these types of samples. These versions differ only in how the patient's physician or other health care provider is identified. In addition, it is possible for a health care institution (e.g., hospital or multispecialty group practice) to field the ADAPT survey in a number, or all of its relevant clinical programs. The clinical program version should be used if a particular clinician of interest is known (generally the patient's "main provider" for his or her chronic illness) or the health plan version should be used if claims or billing data are available.

Instructions and recommendations are provided in the following sections:

- Overview of development of the ADAPT survey
- Generation of a sample frame
- Data collection protocols
- Response rate calculation and data cleaning
- Calculation of measure measure scores

Overview of development of the ADAPT survey

The ADAPT survey is a validated, youth-reported measure of the quality of health care transition (HCT) preparation. The survey is designed to be completed by 16- and 17-year-old patients receiving care in a pediatric-focused health system. It was designed and validated for use among youth with chronic health conditions. Its purpose is to measure the quality of transition preparation based on youth reports of whether specific, recommended aspects of care were received. Three measure scores summarize responses in key domains of HCT preparation:

1. Counseling on Transition Self-Management
2. Counseling on Prescription Medication
3. Transfer Planning

Development of the ADAPT survey included an extensive review of the HCT literature; expert interviews; parent, adolescent, and young adult focus groups in 3 large US cities; cognitive interviews in 3 cities; 3 field tests (1 with youth cared for in specialty clinics at a freestanding pediatric hospital and 2 with health plans serving Medicaid enrollees); and analysis for validity, reliability, and measure development.

To properly identify the treating health care provider, the first question of the clinical program version and the health plan version of the survey differ. The complete ADAPT surveys are

available in ***Appendix C (Health Plan and Clinical Program versions, English)*** and ***Appendix D* (*Health Plan and Clinical Program versions, Spanish).***

Generation of a sample frame

Eligibility

The ADAPT survey is intended to be completed by youth either (a) receiving health care services in a clinical program or (b) enrolled in a health plan or similar defined population. Eligibility for participation is based on the following criteria:

- Age 16 to 17 years old at the time of survey completion
- At least 1 chronic health condition
- At least 1 outpatient visit with a health care provider in the preceding 12 months
- For health plan sampling, current enrollment at the time of the survey and enrollment over the preceding 12 months (allowing for ≤45 day gaps during that period)

Identification of youth with chronic health conditions

For a clinical program, patient registries, electronic health records, or patient panels can be used to determine eligibility for the survey based on the goals for quality measurement. For example, a group practice might choose to survey patients receiving longitudinal care from a specific group of subspecialty providers. The approach to selection of the sample varies depending on the size of the patient population and the data available for identification.

For a health plan or other entity with access to administrative claims data, identification of patients for the ADAPT survey can be accomplished by applying the Pediatric Medical Complexity Algorithm (PMCA) to claims data. Use of this standard approach will identify a valid sample that can be compared across health plans or other entities. The PMCA is a recently developed, publicly available algorithm that identifies children with complex chronic disease in claims or hospital discharge data with good sensitivity and specificity.[1] The PMCA was developed as part of the Pediatric Quality Measures Program to classify levels of medical complexity for children with special health care needs. The PMCA assigns body system flags, based on International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes, to enable counts of affected body systems and subsequent assignment to 1 of the 2 chronic disease categories: 1) a noncomplex chronic disease (NC-CD) is defined as a non-progressive and non-malignant chronic condition in **only 1** body system; and 2) a complex chronic disease (C-CD) is defined as a chronic condition that is progressive or malignant or in which **more than 1** body system is involved. As detailed below, a stratified random sample of patients identified by the PMCA was used in the validation studies of the ADAPT survey.

Exclusions

Patients who meet the eligibility criteria outlined above should generally be included in the ADAPT survey sample. However, the following categories of otherwise eligible patients should be excluded from the sample frame:

- Patients who request that they not be contacted
- Court/law enforcement involved patients (i.e., prisoners); this category does not include those residing in halfway houses
- Patients with a foreign home address (the US territories – American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and Virgin Islands – are not considered foreign addresses and therefore are not excluded)
- Patients who cannot be surveyed because of local, state, or federal regulations

Note: Include patients in the sample frame unless there is positive evidence that they are ineligible or fall within an excluded category. If information is missing on any variable that affects survey eligibility when the sample frame is constructed, do not exclude the patient from the sample frame because of that variable.

Sample creation

Clinical programs or health plans utilizing the ADAPT survey are responsible for generating complete, accurate, and valid sample frame data files that contain all administrative information for each patient who meets the eligibility criteria. The minimum data elements for sample frame creation for the ADAPT survey are in ***Appendix E***.

The data elements that are most critical to the success of data collection are accurate and complete patient/member names, clinical program or health plan names, and home address.

De-duplication

Duplication of patients within the survey sample may occur if, for example, information for an eligible patient is received from multiple clinical programs within one hospital or practice setting. Perform de-duplication using the medical record number or health plan member identification number.

Sample size

The sample size goal for the survey should account for:

- The accuracy of patient/member home address
- The anticipated response rate based on prior surveys of the same or similar populations

The ADAPT survey can be used to assess the quality of transition preparation in a health plan or state Medicaid program, or as a tool for ongoing quality improvement in clinical programs. We have based our sample size recommendations on prior evaluations of widely used national patient experience surveys that have determined sample size requirements for adequate

reliability.[2, 3, 4] For health plan or state Medicaid program comparisons, we recommend at least 300 completed surveys per health plan. By extension, we also recommend this sample size for comparisons of performance among large delivery systems (e.g., large multispecialty practices or hospitals with a number of outpatient programs for youth with chronic illness). This estimate may be further refined in the future, as data are collected from a larger number of health plans than was possible in the current field test. Because response rates will vary among health plans and cannot be predicted with certainty, a conservative approach of aiming for slightly more than 300 completed surveys is recommended. The example in ***Appendix F*** shows the sample size calculation for a goal of 300 surveys for a health plan with a predicted response rate of 20 percent.

The ADAPT survey may also be used to assess performance for individual clinical programs. The number of responses for each administration will vary with the size of the available patient pool and the intended use. While further study is needed to determine the recommended sample size required for comparisons across programs, an individual program may use this measure over time to guide and assess improvement efforts. In general, the survey is not designed to measure or compare the performance of individual health care providers.

Sampling procedure

For large practices, hospitals, or health plans, use Simple Random Sampling (SRS) to draw the desired final sample. To use SRS as the sampling method, randomly select the desired final sample size from all eligible patients. The chance that each patient will be selected is equal for all patients. For smaller populations of interest (e.g., a single clinical program), it may be necessary to select all of the treated patients to receive the survey in order to achieve the desired sample size.

In the case of a defined population (e.g., a health plan), use the PMCA algorithm to identify eligible patients, then draw equally sized random samples from the identified non-complex chronic disease (NC-CD) group and complex chronic disease (C-CD) group.

Since the survey is mailed to parent(s)/guardian(s) of identified patients, sampling populations of adolescents whose sole chronic condition is a mental health condition may pose an unacceptable risk of a breach of confidentiality. In the validation studies of the ADAPT survey, youth in the NC-CD group with only a mental health condition were excluded due to privacy concerns. Youth in the C-CD group were included if they had a mental health condition concurrent with a health condition affecting another body system. In addition, the sampling procedure should ensure that no more than 20% of the patients in the NC-CD sample have a condition affecting any one body system.

Preparing sample files for survey administration

Once the sample has been selected, assign a unique survey identification number to each prospective respondent (sampled patient). This unique ID number should **not** be based on an existing identifier, such as a Social Security Number or a patient ID number. This number will be used **only** to track the respondents during data collection.

The sampling fraction of the total eligible population will vary depending on the overall size of the population. Some small clinical programs or health plans may not be able to obtain the minimum desired number of completed surveys. In such cases, sample **all** eligible patients or members in an attempt to obtain as many completed surveys as possible.

Data collection protocols

Mail protocol

This section lists recommended steps for administering the survey by mail.

- **Set up a toll-free number (or use an existing information line) to include in all correspondence with prospective respondents.** Train staff members to respond to questions. Maintain a log of these calls and review them periodically for common issues that arise.

- **Mail the survey addressed to the parent/guardian of the prospective respondents with a cover letter and a postage-paid envelope.** The cover letter should include instructions for the adolescent patient to complete and return the survey. For examples, see ***English mailed survey materials (Appendix G)*** and ***Spanish mailed survey materials (Appendix H)*.**

    - **Tips for the cover letter:**
        - Personalize the letter with the name and address of the intended recipient (parent/guardian).
        - Tailor the letter to include language that explains the purpose of the survey, the voluntary nature of participation, and the confidentiality of responses.
        - Include language in the letter that asks the parent or guardian to give the survey to their adolescent child.
        - Indicate that if the adolescent child is unable to complete the survey independently (e.g., due to developmental delay), then the survey should not be completed. Include a check box in the cover letter for the parent/guardian to indicate that the identified child is unable to complete the survey, and instruct the parent or guardian to return the blank survey and cover letter for tracking.
        - Note that non-participation will not affect the health care of either the parent/guardian or the adolescent child.
        - Have the letter signed by a representative of the clinical program or health plan.
        - Confirm that the reading level of the cover letter is appropriate for the population and meets all applicable regulatory requirements.

    - **Tips for the outside envelope:**
        - Make the envelope look “official” but not bureaucratic or like “junk mail.”

  - Place a recognizable sponsor's name above the return address.
  - Mark the envelopes "change of service requested" in order to receive information to update records for respondents who have moved and to increase the likelihood that the survey will reach the intended respondent.

- **Maintain a database of returned surveys by unique survey identifier**. Each prospective respondent in the response tracking system should be assigned a survey result code that indicates whether he or she completed and returned the survey, was ineligible to participate in the study, could not be located, or refused to participate.

- **Send a second survey 3 weeks after the initial mailing**. To avoid mailing another survey to those who have already responded, finish entry of returned surveys into the database before mailing second surveys. Include in the second mailing a slightly adapted reminder letter to those parents whose adolescent children have not responded to the first mailing and another postage-paid return envelope. Examples of the reminder letter can be found in the ***mailed survey materials, English (Appendix G)*** and ***mailed survey materials, Spanish (Appendix H)***.

- **Close data collection 10 weeks from the first survey mailing.**

Calculation of the response rate

The response rate is the total number of completed surveys divided by the total number of surveys mailed, excluding from the denominator those that are either undeliverable or are returned with the indication that the patient does not meet eligibility criteria or is unable to complete the survey independently.

Numerator

- *Completed surveys:* A survey should be considered *complete* if it has responses for greater than 50% of questions 4-8, or if a respondent answers "None" to question 3.

Denominator

- *Completed surveys plus non-responses:* Non-responses include all surveys mailed but not returned, except for the following exclusions:
  - *Undeliverable:* The survey was returned by US Mail as undeliverable. "Undeliverable" should not be assumed merely because of non-response.
  - *Patient ineligible:* The survey was returned with clear indication that the patient does not meet eligibility criteria (e.g., ineligible age or lack of a chronic health condition).
  - *Patient unable to complete survey independently:* This must be indicated by the appropriate checkbox in the cover letter or equivalent clear indication by the parent/guardian that the patient is unable to complete the survey independently (e.g., due to cognitive limitation).

Data cleaning protocols

Basic data cleaning procedures that include identifying out-of-range values, replacing numeric missing values with codes for “missing,” and checking for high missing rates for individual items are recommended prior to analysis of survey responses. In addition, “forward cleaning” of items that could be legitimately skipped also is recommended: if a question was supposed to be skipped because of the response to a screening question but was not, then replace the dependent response with the value “missing”. The value of a screening response should not be changed because a response was present for a question that should have been legitimately skipped. For a more detailed description of the data cleaning approach, see ***Decision Rules and Coding Guidelines (Appendix L).***

Calculation of measure scores

There are 3 domain-level measures included in the ADAPT survey. The calculation of measure scores is described below.

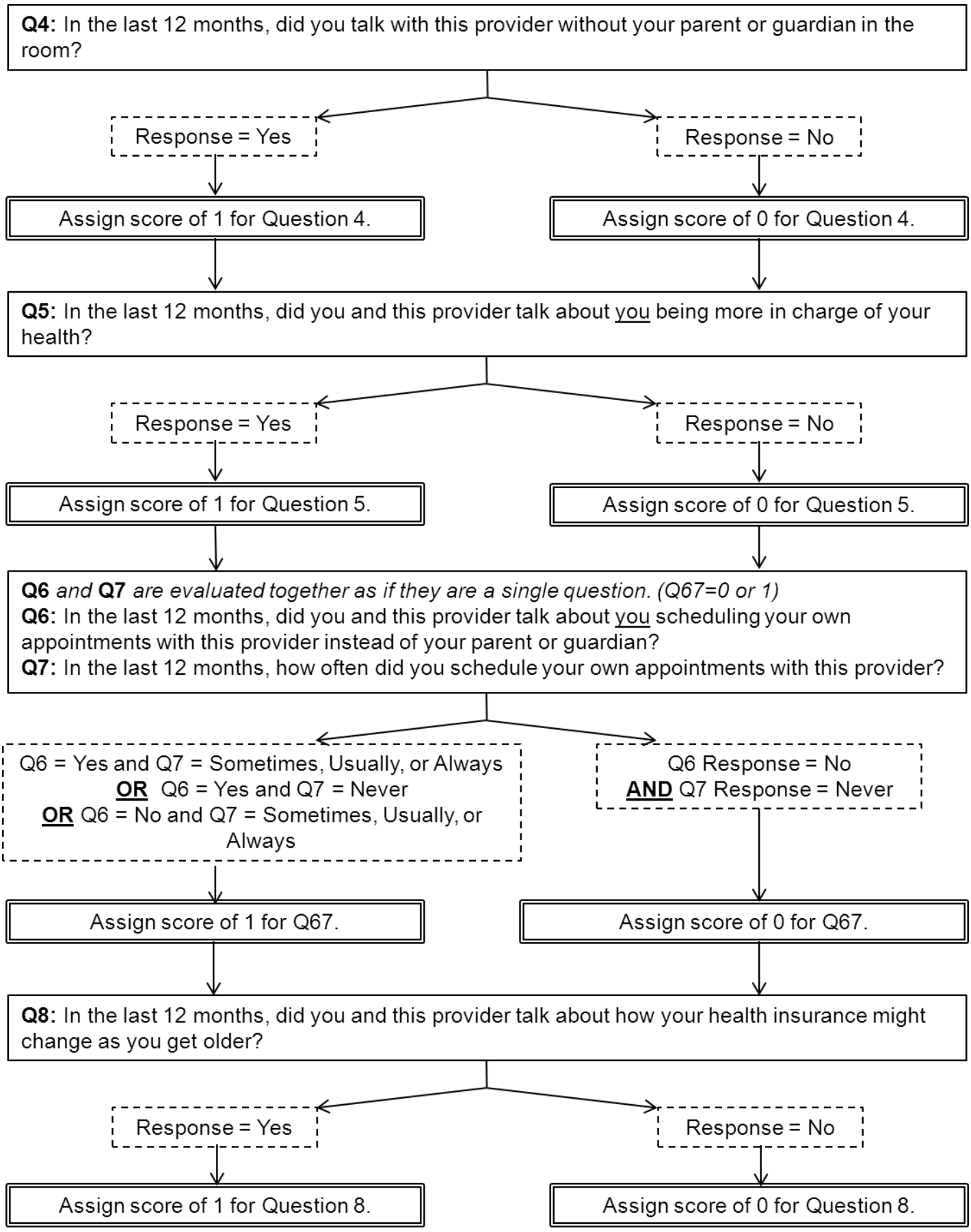
1) Counseling on Transition Self-Management:

This measure is produced by combining responses to 5 questions:

- Q4: In the last 12 months, did you talk with this provider without your parent or guardian in the room?
- Q5: In the last 12 months, did you and this provider talk about your being more in charge of your health?
- Q6: In the last 12 months, did you and this provider talk about your scheduling your own appointments with this provider instead of your parent or guardian?
- Q7: In the last 12 months, how often did you schedule your own appointments with this provider?
- Q8: In the last 12 months, did you and this provider talk about how your health insurance might change as you get older?

The 5 questions are scored as indicated in Figure 1 below.

Figure 1. Flow Diagram for Measure 1



Response options for questions 4-6 and 8 are “Yes” or “No”:

- Assign a score of 0 for No
- Assign a score of 1 for Yes

Response options for question 7 are “Never,” “Sometimes,” “Usually,” or “Always”:

- Assign a score of 0 for Never
- Assign a score of 1 for Sometimes, Usually, or Always

Questions 6 and 7 are evaluated together as if they were a single question (Q67), the score of which is calculated as follows:

- Assign a score of 0 if Q6 = 0 AND Q7 = 0
- Assign a score of 1 if Q6 = 1 AND/OR Q7 = 1

The basic steps to calculate the measure score for a population are as follows:

- For each question, identify responses with non-missing values for that question
- For each respondent, calculate the proportion of responses with a score of 1 among all of the questions in the measure
- Calculate the numerator and denominator of the measure
    - Numerator = the sum of the proportions of positive responses among the questions in the measure for all respondents
    - Denominator = the number of respondents with valid responses (i.e., non-missing values)

For each respondent, the proportion (P) of positive responses for the questions (Q) within the measure can be defined as follows:

P = (Q4 + Q5 + Q67 + Q8)/4

Measure score = (summation of values of P for N respondents/N)*100

Where N = the number of respondents with valid responses for P4, P5, P6, P7, and P8.

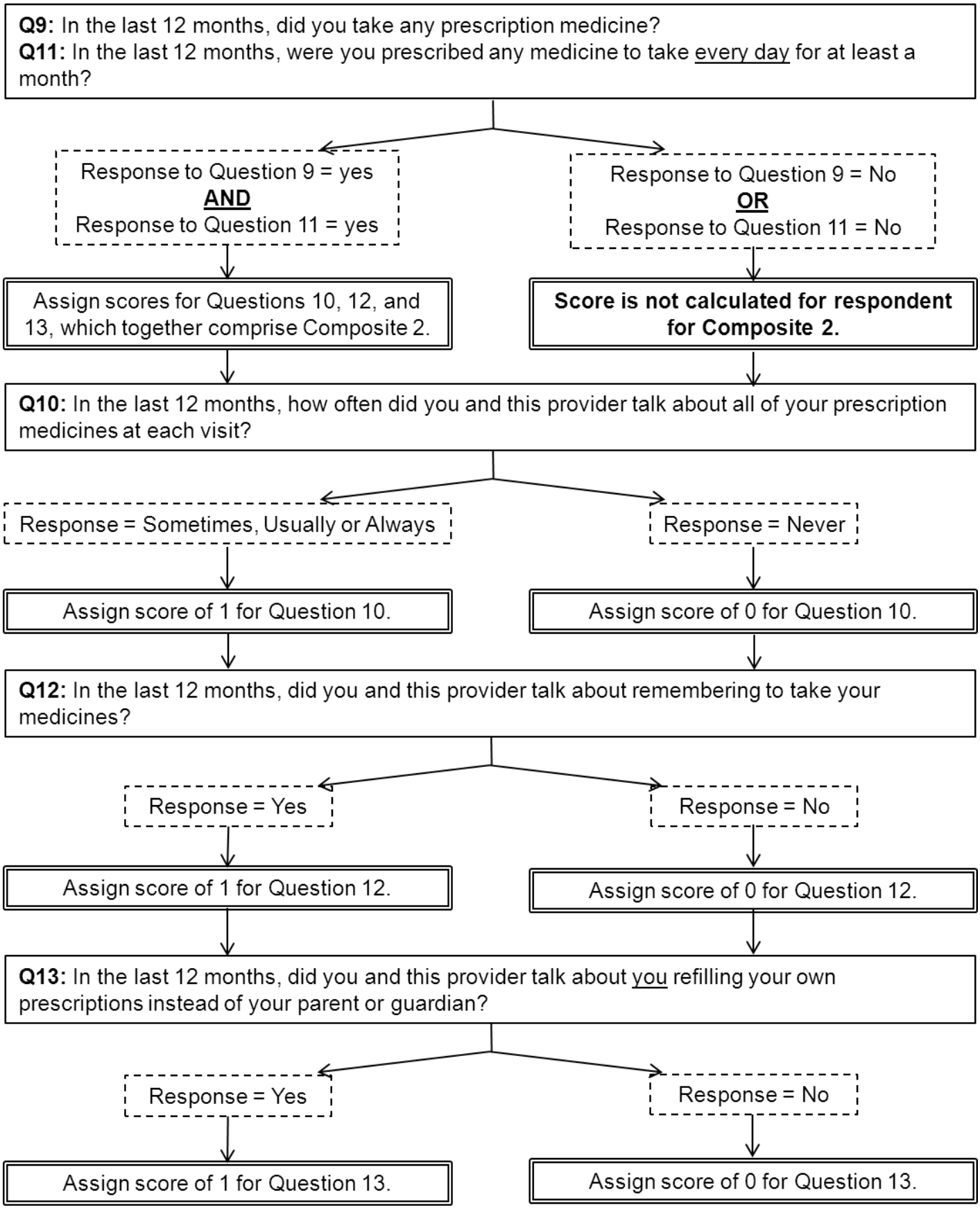
2) Counseling on Prescription Medication:

The measure is produced by combining responses to questions 10, 12, and 13:

- Q10: In the last 12 months, how often did you and this provider talk about all of your prescription medicines at each visit?
- Q12: In the last 12 months, did you and this provider talk about remembering to take your medicines?
- Q13: In the last 12 months, did you and this provider talk about you refilling your own prescriptions instead of your parent or guardian?

The 3 questions are scored as indicated in Figure 2 below.

Figure 2. Flow Diagram for Measure 2



This measure score is calculated only for respondents who indicate on questions 9 (“in the last 12 months, did you take any prescription medicine?”) and 11 (“in the last 12 months, were you prescribed any medicine to take every day for at least a month?”) that they take prescription medication every day.

For each question, identify cases with non-missing values and for which the response for both question 9 and question 11 is “Yes”:

- Respondents who do not report taking prescription medicine every day (responses of “No” to either questions 9 or 11) are not included in the population for which this measure is calculated

Response options for question 10 are “Never,” “Sometimes,” “Usually,” or “Always”

- Assign a score of 0 for Never
- Assign a score of 1 for Sometimes, Usually, or Always

Response options for questions 12 and 13 are “Yes” or “No”

- Assign a score of 0 for No
- Assign a score of 1 for Yes

The basic steps to calculate the measure score for a population are as follows:

- For each question, identify responses with non-missing values for that question
- For each respondent, calculate the proportion of responses with a score of 1 among all of the questions in the measure
- Calculate the numerator and denominator of the measure
    - Numerator = the sum of the proportions of positive responses among the questions in the measure for all respondents
    - Denominator = the number of respondents with valid responses (i.e., non-missing values)

For each respondent, the proportion (P) of positive responses for the questions (Q) within the measure can be defined as follows:

P = (Q10 + Q12 + Q13)/3

Measure score = (summation of values of P for N respondents/N)*100

Where N = the number of respondents with valid responses for P10, P12, and P13.

3) Transfer Planning:

The measure is produced by combining responses to questions 15, 16, 17, and 18:

- Q15: In the last 12 months, did you and this provider talk about whether you may need to change to a new provider who treats mostly adults?

- Q16: In the last 12 months, did this provider ask if you had any questions or concerns about changing to a new provider who treats mostly adults?
- Q17: In the last 12 months, did you and this provider talk about a specific plan for changing to a new provider who treats mostly adults?
- Q18: Did this provider give you this plan in writing?

Only respondents for which the response to question 14 (“Does this provider treat mostly children and teens?”) is “Yes,” “Don’t Know,” or left blank are included in the population for which this measure is calculated.

The 4 questions are scored as indicated in Figure 3 below.

For each question, identify cases with non-missing values and for which the response for question 14 is “Yes” or “Don’t know”:

- Respondents who indicate the provider does not mostly treat children and teens (response of “No” to question 14) are not included in the population for which this measure is calculated

Response options for Questions 15-18 are “Yes” or “No”

- Assign a score of 0 for No
- Assign a score of 1 for Yes

The basic steps to calculate the measure score for a population are as follows:

- For each question, identify responses with non-missing values for that question
- For each respondent, calculate the proportion of responses with a score of 1 among all of the questions in the measure
- Calculate the numerator and denominator of the measure
    - Numerator = the sum of the proportions of positive responses among the questions in the measure for all respondents
    - Denominator = the number of respondents with valid responses (i.e. non-missing response OR assigned responses [see decision rules outlined in ***Appendix L***])

For each respondent, the proportion (P) of positive responses for the questions (Q) within the measure can be defined as follows:

P = (Q15 + Q16 + Q17 + Q18)/4

Measure score = (summation of values of P for N respondents/N)*100

Where N = the number of respondents with valid responses for P15, P16, P17, and P18.

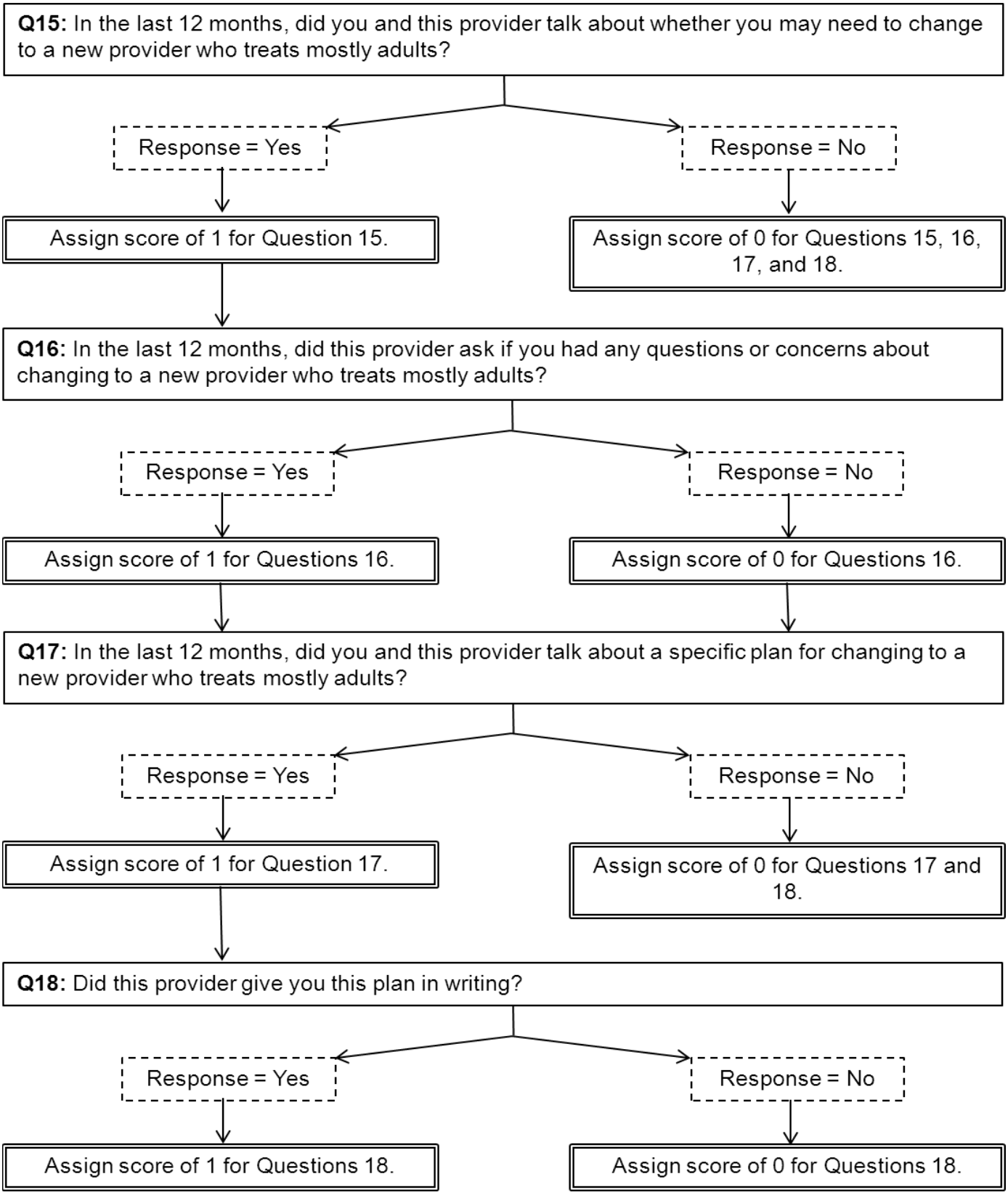


Figure 3. Flow Diagram for Measure 3

References

1. Simon TD, Cawthon ML, Stanford S, Popalisky J, Lyons D, Woodcox P, Hood M, Chen AY, Mangione-Smith R, Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN) Medical Complexity Working Group. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. Pediatrics. 2014;133(6):e1647-1654. doi:10.1542/peds.2013-3875.

2. Fielding the CAHPS® Clinician & Group Surveys. US Department of Health and Human Services, Agency for Healthcare Research and Quality. https://cahps.ahrq.gov/surveys-guidance/survey4.0-docs/1033_CG_Fielding_the_Survey.pdf. Updated 9/1/2011. Accessed September 24, 2015.

3. Solomon LS, Hays RD, Zaslavsky AM, Ding L, Cleary PD. Psychometric Properties of a Group-Level Consumer Assessment of Health Plans Study (CAHPS) Instrument. *Medical Care.* 2005;43(1):53-60

4. Safran DG, Karp M, Coltin K, Chang H, Li A, Ogren J, Rogers WH. Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. J Gen Intern Med. 2006;21(1):13-21. doi:10.1111/j.1525-1497.2005.00311.x.

Appendix B. Case-Mix Adjustment Methodology

One of the methodological issues associated with making comparisons across clinical programs/health plans is the need to adjust appropriately for case-mix differences. Case mix refers to patient characteristics that are not under the control of the clinical programs/health plans that may affect measures of outcomes or processes, such as demographic characteristics and health status. Systematic effects of this sort create the potential for clinical programs/health plans ratings to be higher or lower because of the characteristics of their patient population, rather than because of the quality of care they provide, making comparisons of unadjusted scores misleading. The basic goal of adjusting for case mix is to estimate how different clinical programs/health plans would be rated if they all provided care to comparable groups of patients.

The case-mix adjustment will use a regression methodology also referred to as covariance adjustment. As an example of how this will work, let y_{ipj} represent the response to question *i* of respondent *j* from clinical program/health plan *p* (after recoding, if any, has been performed). The model for adjustment of a single item *i* is of the form:

$$y_{ipj} = \beta_i' x_{ipj} + \mu_{ip} + \varepsilon_{ipj}$$

where β_i is a regression coefficient vector, x_{ipj} is a covariate vector consisting of two adjuster, μ_{ip} is an intercept parameter for clinical program/health plan *p*, and ε_{ipj} is the error term. The estimates are given by the following equation:

$$\left(\hat{\beta}' \hat{\mu}'\right)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y_i}$$

where $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots \mu_{ip})'$ is the vector of intercepts, *y*, is the vector of responses and the covariate matrix is

$$\mathbf{X} = (\mathbf{X}_a\, u_1\, u_2 \ldots u_p)$$

where the columns of $\mathbf{X}_a$ are the vectors of values of each of the adjuster covariates, and u_p is a vector of indicators for being included in a clinical program/health plan *p*, *p* = 1, 2,…*P*, with entries equal to 1 for respondents in clinical program/health plan *p* and 0 for others.

Finally, the estimated intercepts are shifted by a constant amount to force their mean to equal the mean of the unadjusted clinical program/health plan means $\bar{y}_{ip}$ (to make it easier to compare adjusted and unadjusted means), giving adjusted clinical program/health plan means

$$\hat{a}_{ip} = \hat{\mu}_{ip} + (1/P)\sum_p y_{ip} - (1/P)\sum_p \hat{\mu}_{ip}$$

For single-item responses, these adjusted means are reported. For measure scores, the adjusted clinical program/health plan means are combined using the mean of the adjusted clinical program/health plan means for all the relevant items:

$$\hat{a}_p = \sum_i \hat{a}_{ip}$$

ADAPT survey: Health Plan **YOUR EXPERIENCES GETTING HEALTH CARE**

INSTRUCTIONS

Answer all the questions by checking the box next to your answer.

You are sometimes told to skip over some questions in this survey. When this happens you will see an arrow with a note that tells you what question to answer next, like this:

☑ Yes → **If Yes, go to #1**

☐ No

Your participation is voluntary. You may choose to answer the survey or not. If you choose not to, this will not affect the health care you get.

What to do when you're done. Please return the completed survey in the postage paid envelope.

1. Your main provider is the doctor or other health care provider who is in charge of the care for your health condition. If you have more than one health condition, please think about the condition that concerns you the most.

Which of the providers named below is your main provider?

☐ PROVIDER NAME 1 HERE

☐ PROVIDER NAME 2 HERE

☐ PROVIDER NAME 3 HERE

☐ None of these are my main provider, my main provider is_____
(please print)

The questions in this survey will refer to the provider chosen in question 1 as "this provider." Please think of that provider as you answer the survey.

2. How long have you been going to this provider?

☐ Less than 6 months
☐ At least 6 months but less than 1 year
☐ At least 1 year but less than 3 years
☐ At least 3 years but less than 5 years
☐ 5 years or more

3. In the last 12 months, how many times did you visit this provider?

☐ None → **If None, go to #19**
☐ 1 time
☐ 2
☐ 3
☐ 4
☐ 5 to 9
☐ 10 or more times

4. In the last 12 months, did you talk with this provider without your parent or guardian in the room?

☐ Yes
☐ No

5. In the last 12 months, did you and this provider talk about you being more in charge of your health?

☐ Yes
☐ No

6. In the last 12 months, did you and this provider talk about you scheduling your own appointments with this provider instead of your parent or guardian?

- ☐ Yes
- ☐ No

7. In the last 12 months, how often did you schedule your own appointments with this provider?

- ☐ Never
- ☐ Sometimes
- ☐ Usually
- ☐ Always

8. In the last 12 months, did you and this provider talk about how your health insurance might change as you get older?

- ☐ Yes
- ☐ No

Your Prescription Medicines

9. In the last 12 months, did you take any prescription medicine?

- ☐ Yes → **If Yes, go to #10**
- ☐ No → **If No, go to #14**

10. In the last 12 months, how often did you and this provider talk about all of your prescription medicines at each visit?

- ☐ Never
- ☐ Sometimes
- ☐ Usually
- ☐ Always

11. In the last 12 months, were you prescribed any medicine to take <u>every day</u> for at least a month?

- ☐ Yes → **If Yes, go to #12**
- ☐ No → **If No, go to #14**

12. In the last 12 months, did you and this provider talk about remembering to take your medicines?

- ☐ Yes
- ☐ No

13. In the last 12 months, did you and this provider talk about <u>you</u> refilling your own prescriptions instead of your parent or guardian?

- ☐ Yes
- ☐ No

Your Provider

14. Does this provider treat mostly children and teens?

- ☐ Yes → **If Yes, go to #15**
- ☐ No → **If No, go to #19**
- ☐ Don't Know → **If Don't Know, go to #15**

15. In the last 12 months, did you and this provider talk about whether you may need to change to a new provider who treats mostly adults?

- ☐ Yes → **If Yes, go to #16**
- ☐ No → **If No, go to #19**

16. In the last 12 months, did this provider ask if you had any questions or concerns about changing to a new provider who treats mostly adults?

- ☐ Yes
- ☐ No

17. In the last 12 months, did you and this provider talk about a specific plan for changing to a new provider who treats mostly adults?

- ☐ Yes → **If Yes, go to #18**
- ☐ No → **If No, go to #19**

18. Did this provider give you this plan in writing?

- ☐ Yes
- ☐ No

About You

19. How old are you?

- ☐ 15
- ☐ 16
- ☐ 17
- ☐ 18

20. Are you male or female?

- ☐ Male
- ☐ Female

21. In general, how would you rate your overall health?

- ☐ Excellent
- ☐ Very good
- ☐ Good
- ☐ Fair
- ☐ Poor

22. What is the highest grade or level of school that you have completed?

- ☐ 8th grade or less
- ☐ 9th grade
- ☐ 10th grade
- ☐ 11th grade
- ☐ 12th grade, high school graduate or GED
- ☐ Some college

23. Are you of Hispanic, Latino, or Spanish origin? Mark one or more.

- ☐ No, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican American, Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin

24. How would you describe your race? Mark one or more.

- ☐ White
- ☐ Black or African American
- ☐ Asian
- ☐ Native Hawaiian or other Pacific Islander
- ☐ American Indian or Alaska Native

25. Did someone help you complete this survey?

- ☐ Yes→ **If Yes, go to #26**
- ☐ No → Thank you. Please return the completed survey in the postage-paid envelope.

26. How did that person help you? Mark one or more.

- ☐ Read the questions to me
- ☐ Wrote down the answers I gave
- ☐ Answered the questions for me
- ☐ Translated the questions into my language
- ☐ Helped in some other way:_____

_____

Please print

Thank you.

Please return the survey in the postage-paid envelope.

ADAPT survey: Clinical Program **YOUR EXPERIENCES GETTING HEALTH CARE**

INSTRUCTIONS

Answer all the questions by checking the box next to your answer.

You are sometimes told to skip over some questions in this survey. When this happens you will see an arrow with a note that tells you what question to answer next, like this:

☑ Yes → **If Yes, go to #1**

☐ No

Your participation is voluntary. You may choose to answer the survey or not. If you choose not to, this will not affect the health care you get.

What to do when you're done. Please return the completed survey in the postage paid envelope.

1. Your main provider is the doctor or other health care provider who is in charge of the care for your health condition. If you have more than one health condition, please think about the condition that concerns you the most.

Is the provider named below your main provider?

Name of provider label goes here

☐ Yes
☐ No, my main provider is _____
(please print)

The questions in this survey will refer to the provider chosen in question 1 as "this provider." Please think of that provider as you answer the survey.

2. How long have you been going to this provider?

☐ Less than 6 months
☐ At least 6 months but less than 1 year
☐ At least 1 year but less than 3 years
☐ At least 3 years but less than 5 years
☐ 5 years or more

3. In the last 12 months, how many times did you visit this provider?

☐ None → **If None, go to #19**
☐ 1 time
☐ 2
☐ 3
☐ 4
☐ 5 to 9
☐ 10 or more times

4. In the last 12 months, did you talk with this provider without your parent or guardian in the room?

☐ Yes
☐ No

5. In the last 12 months, did you and this provider talk about you being more in charge of your health?

☐ Yes
☐ No

6. In the last 12 months, did you and this provider talk about you scheduling your own appointments with this provider instead of your parent or guardian?

☐ Yes
☐ No

7. **In the last 12 months, how often did you schedule your own appointments with this provider?**
   - ☐ Never
   - ☐ Sometimes
   - ☐ Usually
   - ☐ Always

8. **In the last 12 months, did you and this provider talk about how your health insurance might change as you get older?**
   - ☐ Yes
   - ☐ No

Your Prescription Medicines

9. **In the last 12 months, did you take any prescription medicine?**
   - ☐ Yes → **If Yes, go to #10**
   - ☐ No → **If No, go to #14**

10. **In the last 12 months, how often did you and this provider talk about all of your prescription medicines at each visit?**
    - ☐ Never
    - ☐ Sometimes
    - ☐ Usually
    - ☐ Always

11. **In the last 12 months, were you prescribed any medicine to take <u>every day</u> for at least a month?**
    - ☐ Yes → **If Yes, go to #12**
    - ☐ No → **If No, go to #14**

12. **In the last 12 months, did you and this provider talk about remembering to take your medicines?**
    - ☐ Yes
    - ☐ No

13. **In the last 12 months, did you and this provider talk about <u>you</u> refilling your own prescriptions instead of your parent or guardian?**
    - ☐ Yes
    - ☐ No

Your Provider

14. **Does this provider treat mostly children and teens?**
    - ☐ Yes → **If Yes, go to #15**
    - ☐ No → **If No, go to #19**
    - ☐ Don't Know → **If Don't Know, go to #15**

15. **In the last 12 months, did you and this provider talk about whether you may need to change to a new provider who treats mostly adults?**
    - ☐ Yes → **If Yes, go to #16**
    - ☐ No → **If No, go to #19**

16. **In the last 12 months, did this provider ask if you had any questions or concerns about changing to a new provider who treats mostly adults?**
    - ☐ Yes
    - ☐ No

17. **In the last 12 months, did you and this provider talk about a specific plan for changing to a new provider who treats mostly adults?**
    - ☐ Yes → **If Yes, go to #18**
    - ☐ No → **If No, go to #19**

18. **Did this provider give you this plan in writing?**
    - ☐ Yes
    - ☐ No

About You

19. **How old are you?**

- ☐ 15
- ☐ 16
- ☐ 17
- ☐ 18

20. Are you male or female?

- ☐ Male
- ☐ Female

21. In general, how would you rate your overall health?

- ☐ Excellent
- ☐ Very good
- ☐ Good
- ☐ Fair
- ☐ Poor

22. What is the highest grade or level of school that you have completed?

- ☐ 8th grade or less
- ☐ 9th grade
- ☐ 10th grade
- ☐ 11th grade
- ☐ 12th grade, high school graduate or GED
- ☐ Some college

23. Are you of Hispanic, Latino, or Spanish origin? Mark one or more.

- ☐ No, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican American, Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin

24. How would you describe your race? Mark one or more.

- ☐ White
- ☐ Black or African American
- ☐ Asian
- ☐ Native Hawaiian or other Pacific Islander
- ☐ American Indian or Alaska Native

25. Did someone help you complete this survey?

- ☐ Yes→ **If Yes, go to #26**
- ☐ No → Thank you. Please return the completed survey in the postage-paid envelope.

26. How did that person help you? Mark one or more.

- ☐ Read the questions to me
- ☐ Wrote down the answers I gave
- ☐ Answered the questions for me
- ☐ Translated the questions into my language
- ☐ Helped in some other way:_____

_____

Please print

Thank you.

Please return the survey in the postage-paid envelope.

ADAPT survey: Health Plan **TUS EXPERIENCIAS CON LA ATENCIÓN MÉDICA**

INSTRUCCIONES

Contesta todas las preguntas marcando el cuadrito junto a la respuesta que desees escoger.

A veces en la encuesta te dicen que saltes algunas preguntas. Cuando esto suceda, verás una flecha con una nota que dice cuál pregunta debes contestar a continuación, como se muestra abajo:

☑ Sí → **Si contestas Sí, pasa a la pregunta 1**

☐ No

Tu participación es voluntaria. Puedes decidir si vas a contestar la encuesta o no. Aunque decidas no contestarla, la atención médica que recibes no se verá afectada.

Qué hacer cuando termines de contestarla. Por favor envíanos la encuesta completada en el sobre con porte o franqueo pagado.

1. Tu proveedor principal es el doctor u otro profesional médico que está a cargo de la atención médica por tu problema de salud. Si tienes más de un problema de salud, por favor piensa en el problema que más te preocupa.

¿Cuál de los proveedores mencionados a continuación es tu proveedor principal?

☐ PROVIDER NAME 1 HERE

☐ PROVIDER NAME 2 HERE

☐ PROVIDER NAME 3 HERE

☐ Ninguno de estos es mi proveedor principal; mi proveedor principal es _____ *(en letra de imprenta o de molde)*

Cuando las preguntas en esta encuesta dicen "este proveedor" se están refiriendo al proveedor que elegiste en la pregunta 1. Por favor piensa en ese proveedor cuando contestes la encuesta.

2. ¿Cuánto tiempo hace que estás yendo a este proveedor?

☐ Menos de 6 meses
☐ Al menos 6 meses pero menos de 1 año
☐ Al menos 1 año pero menos de 3 años
☐ Al menos 3 años pero menos de 5 años
☐ 5 años o más

3. En los últimos 12 meses, ¿cuántas veces visitaste a este proveedor?

☐ Ninguna → **Si contestas Ninguna, pasa a la pregunta 19**
☐ 1 vez
☐ 2
☐ 3
☐ 4
☐ 5 a 9
☐ 10 o más veces

4. En los últimos 12 meses, ¿hablaste con este proveedor a solas, sin que uno de tus padres o tutores estuviera en el consultorio?

☐ Sí
☐ No

5. En los últimos 12 meses, ¿habló contigo este proveedor acerca de que tú estuvieras más a cargo de tu salud?

☐ Sí
☐ No

6. **En los últimos 12 meses, ¿habló contigo este proveedor acerca de que <u>tú</u> hicieras tus propias citas con este proveedor en vez de que las hicieran tus padres o tutores?**
   - ☐ Sí
   - ☐ No

7. **En los últimos 12 meses, ¿qué tan seguido hiciste tú mismo(a) tus citas con este proveedor?**
   - ☐ Nunca
   - ☐ Algunas veces
   - ☐ Generalmente
   - ☐ Siempre

8. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de que tal vez tengas que cambiar de seguro de salud cuando seas mayor?**
   - ☐ Sí
   - ☐ No

Tus medicinas recetadas

9. **En los últimos 12 meses, ¿tomaste alguna medicina recetada?**
   - ☐ Sí → **Si contestas Sí, pasa a la pregunta 10**
   - ☐ No → **Si contestas No, pasa a la pregunta 14**

10. **En los últimos 12 meses, ¿qué tan seguido habló este proveedor contigo en cada visita acerca de todas tus medicinas recetadas?**
    - ☐ Nunca
    - ☐ Algunas veces
    - ☐ Generalmente
    - ☐ Siempre

11. **En los últimos 12 meses, ¿te recetaron alguna medicina para tomar <u>todos los días</u> por al menos un mes?**
    - ☐ Sí → **Si contestas Sí, pasa a la pregunta 12**
    - ☐ No → **Si contestas No, pasa a la pregunta 14**

12. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de acordarte de tomar tus medicinas?**
    - ☐ Sí
    - ☐ No

13. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de que <u>tú</u> hagas surtir tus medicinas recetadas en vez de tus padres o tutores?**
    - ☐ Sí
    - ☐ No

Tu proveedor

14. **¿Este proveedor trata principalmente a niños y adolescentes?**
    - ☐ Sí → **Si contestas Sí, pasa a la pregunta 15**
    - ☐ No → **Si contestas No, pasa a la pregunta 19**
    - ☐ No sé → **Si contestas No sé, pasa a la pregunta 15**

15. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de si podrías necesitar cambiarte a un proveedor nuevo que trate principalmente a adultos?**
    - ☐ Sí → **Si contestas Sí, pasa a la pregunta16**
    - ☐ No → **Si contestas No, pasa a la pregunta 19**

16. En los últimos 12 meses, ¿te preguntó este proveedor si tenías alguna pregunta o inquietud acerca de cambiarte a un proveedor nuevo que trate principalmente a adultos?

- ☐ Sí
- ☐ No

17. En los últimos 12 meses, ¿habló este proveedor contigo acerca de un plan específico para cambiarte a un proveedor nuevo que atiende principalmente a adultos?

- ☐ Sí → **Si contestas Sí, pasa a la pregunta 18**
- ☐ No → **Si contestas No, pasa a la pregunta 19**

18. ¿Te dio este proveedor el plan por escrito?

- ☐ Sí
- ☐ No

Acerca de ti

19. ¿Cuántos años tienes?

- ☐ 15
- ☐ 16
- ☐ 17
- ☐ 18

20. ¿Eres hombre o mujer?

- ☐ Hombre
- ☐ Mujer

21. En general, ¿cómo calificarías toda tu salud?

- ☐ Excelente
- ☐ Muy buena
- ☐ Buena
- ☐ Regular
- ☐ Mala

22. ¿Cuál es el grado o nivel escolar más alto que has completado?

- ☐ 8 años de escuela o menos
- ☐ 9 años de escuela
- ☐ 10 años de escuela
- ☐ 11 años de escuela
- ☐ 12 años de escuela, graduado de *high school*, diploma de *high school*, preparatoria, o su equivalente (o GED)
- ☐ Algunos cursos de *college* o universidad

23. ¿Eres de origen hispano, latino o español?
Marca todas las opciones que correspondan.

- ☐ No, ni de origen hispano, ni latino, ni español
- ☐ Sí, de origen mexicano, mexicano-americano, chicano
- ☐ Sí, de origen puertorriqueño
- ☐ Sí, de origen cubano
- ☐ Sí, de otro origen hispano, latino o español

24. ¿Cómo describirías tu raza?
Marca todas las opciones que correspondan.

- ☐ Blanca
- ☐ Negra o afroamericana
- ☐ Asiática
- ☐ Nativa de Hawái o de otras islas del Pacífico
- ☐ Indígena americana o nativa de Alaska

25. ¿Te ayudó alguien a contestar esta encuesta?

- ☐ Sí → **Si contestaste Sí, pasa a la pregunta 26**
- ☐ No → Gracias. Por favor, devuelve esta encuesta en el sobre con el porte o franqueo pagado.

26. ¿Cómo te ayudó esta persona?
Marca todas las opciones que correspondan.

- ☐ Me leyó las preguntas.
- ☐ Anotó las respuestas que le di.
- ☐ Contestó las preguntas por mí.
- ☐ Tradujo las preguntas a mi idioma.
- ☐ Me ayudó de otra forma:_____

_____

Escribe de qué forma te ayudó

Muchas Gracias.
Por favor envíanos la encuesta en el sobre con porte o franqueo pagado.

INSTRUCCIONES

Contesta todas las preguntas marcando el cuadrito junto a la respuesta que desees escoger.

A veces en la encuesta te dicen que saltes algunas preguntas. Cuando esto suceda, verás una flecha con una nota que dice cuál pregunta debes contestar a continuación, como se muestra abajo:

☑ Sí → **Si contestas Sí, pasa a la pregunta 1**

☐ No

Tu participación es voluntaria. Puedes decidir si vas a contestar la encuesta o no. Aunque decidas no contestarla, la atención médica que recibes no se verá afectada.

Qué hacer cuando termines de contestarla. Por favor envíanos la encuesta completada en el sobre con porte o franqueo pagado.

ADAPT survey: Clinical Program **TUS EXPERIENCIAS CON LA ATENCIÓN MÉDICA**

1. Tu proveedor <u>principal</u> es el doctor u otro profesional médico que está a cargo de la atención médica por tu problema de salud. Si tienes más de un problema de salud, por favor piensa en el problema que más te preocupa.

   ¿El proveedor que aparece a continuación es tu proveedor principal?

   Name of provider label goes here

   ☐ Sí
   ☐ No, mi proveedor principal es_____

   (en letra de imprenta o de molde)

Cuando las preguntas en esta encuesta dicen "este proveedor" se están refiriendo al proveedor que elegiste en la pregunta 1. Por favor piensa en ese proveedor cuando contestes la encuesta.

2. ¿Cuánto tiempo hace que estás yendo a este proveedor?

   ☐ Menos de 6 meses
   ☐ Al menos 6 meses pero menos de 1 año
   ☐ Al menos 1 año pero menos de 3 años
   ☐ Al menos 3 años pero menos de 5 años
   ☐ 5 años o más

3. **En los últimos 12 meses, ¿cuántas veces visitaste a este proveedor?**

   ☐ Ninguna → **Si contestas Ninguna, pasa a la pregunta 19**
   ☐ 1 vez
   ☐ 2
   ☐ 3
   ☐ 4
   ☐ 5 a 9
   ☐ 10 o más veces

4. **En los últimos 12 meses, ¿hablaste con este proveedor a solas, sin que uno de tus padres o tutores estuviera en el consultorio?**

   ☐ Sí
   ☐ No

5. **En los últimos 12 meses, ¿habló contigo este proveedor acerca de que tú estuvieras más a cargo de tu salud?**
   - ☐ Sí
   - ☐ No

6. **En los últimos 12 meses, ¿habló contigo este proveedor acerca de que tú hicieras tus propias citas con este proveedor en vez de que las hicieran tus padres o tutores?**
   - ☐ Sí
   - ☐ No

7. **En los últimos 12 meses, ¿qué tan seguido hiciste tú mismo(a) tus citas con este proveedor?**
   - ☐ Nunca
   - ☐ Algunas veces
   - ☐ Generalmente
   - ☐ Siempre

8. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de que tal vez tengas que cambiar de seguro de salud cuando seas mayor?**
   - ☐ Sí
   - ☐ No

Tus medicinas recetadas

9. **En los últimos 12 meses, ¿tomaste alguna medicina recetada?**
   - ☐ Sí → **Si contestas Sí, pasa a la pregunta 10**
   - ☐ No → **Si contestas No, pasa a la pregunta 14**

10. **En los últimos 12 meses, ¿qué tan seguido habló este proveedor contigo en cada visita acerca de todas tus medicinas recetadas?**
    - ☐ Nunca
    - ☐ Algunas veces
    - ☐ Generalmente
    - ☐ Siempre

11. **En los últimos 12 meses, ¿te recetaron alguna medicina para tomar todos los días por al menos un mes?**
    - ☐ Sí → **Si contestas Sí, pasa a la pregunta 12**
    - ☐ No → **Si contestas No, pasa a la pregunta 14**

12. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de acordarte de tomar tus medicinas?**
    - ☐ Sí
    - ☐ No

13. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de que tú hagas surtir tus medicinas recetadas en vez de tus padres o tutores?**
    - ☐ Sí
    - ☐ No

Tu proveedor

14. **¿Este proveedor trata principalmente a niños y adolescentes?**
    - ☐ Sí → **Si contestas Sí, pasa a la pregunta 15**
    - ☐ No → **Si contestas No, pasa a la pregunta 19**
    - ☐ No sé → **Si contestas No sé, pasa a la pregunta 15**

15. **En los últimos 12 meses, ¿habló este proveedor contigo acerca de si podrías necesitar cambiarte a un proveedor nuevo que trate principalmente a adultos?**
    - ☐ Sí → **Si contestas Sí, pasa a la pregunta16**
    - ☐ No → **Si contestas No, pasa a la**

pregunta 19

16. En los últimos 12 meses, ¿te preguntó este proveedor si tenías alguna pregunta o inquietud acerca de cambiarte a un proveedor nuevo que trate principalmente a adultos?

- ☐ Sí
- ☐ No

17. En los últimos 12 meses, ¿habló este proveedor contigo acerca de un plan específico para cambiarte a un proveedor nuevo que atiende principalmente a adultos?

- ☐ Sí → **Si contestas Sí, pasa a la pregunta 18**
- ☐ No → **Si contestas No, pasa a la pregunta 19**

18. ¿Te dio este proveedor el plan por escrito?

- ☐ Sí
- ☐ No

Acerca de ti

19. ¿Cuántos años tienes?

- ☐ 15
- ☐ 16
- ☐ 17
- ☐ 18

20. ¿Eres hombre o mujer?

- ☐ Hombre
- ☐ Mujer

21. En general, ¿cómo calificarías toda tu salud?

- ☐ Excelente
- ☐ Muy buena
- ☐ Buena
- ☐ Regular
- ☐ Mala

22. ¿Cuál es el grado o nivel escolar más alto que has completado?

- ☐ 8 años de escuela o menos
- ☐ 9 años de escuela
- ☐ 10 años de escuela
- ☐ 11 años de escuela
- ☐ 12 años de escuela, graduado de *high school*, diploma de *high school*, preparatoria, o su equivalente (o GED)
- ☐ Algunos cursos de *college* o universidad

23. ¿Eres de origen hispano, latino o español?
Marca todas las opciones que correspondan.

- ☐ No, ni de origen hispano, ni latino, ni español
- ☐ Sí, de origen mexicano, mexicano-americano, chicano
- ☐ Sí, de origen puertorriqueño
- ☐ Sí, de origen cubano
- ☐ Sí, de otro origen hispano, latino o español

24. ¿Cómo describirías tu raza?
Marca todas las opciones que correspondan.

- ☐ Blanca
- ☐ Negra o afroamericana
- ☐ Asiática
- ☐ Nativa de Hawái o de otras islas del Pacífico
- ☐ Indígena americana o nativa de Alaska

25. ¿Te ayudó alguien a contestar esta encuesta?

- ☐ Sí → **Si contestaste Sí, pasa a la pregunta 26**
- ☐ No → Gracias. Por favor, devuelve esta encuesta en el sobre con el porte o franqueo pagado.

26. ¿Cómo te ayudó esta persona?
Marca todas las opciones que correspondan.

- ☐ Me leyó las preguntas.
- ☐ Anotó las respuestas que le di.
- ☐ Contestó las preguntas por mí.
- ☐ Tradujo las preguntas a mi idioma.
- ☐ Me ayudó de otra forma:_____

_____

Escribe de qué forma te ayudó

Muchas Gracias.
Por favor envíanos la encuesta en el sobre con porte o franqueo pagado.

Appendix E: Sample Frame Elements for Administration of the ADAPT Survey.

Sample frame elements for administration of the ADAPT Survey

Name of clinical program / health plan
State of participating clinical program / health plan
Patient or Member ID
Name of patient
Gender of patient
Date of birth
Home address
Type of chronic condition
Names of the health care providers who have most frequently been the billing or treating provider for an encounter with the patient in the past 12 months (up to 3 names are selected for inclusion in the survey)
Length (in months) of continuous enrollment (ignoring gaps ≤45 days)*

Appendix F: Sample Size

Calculation of Estimated Sample Size Needed for a Health Plan

Goal	300 completed surveys
Predicted response rate	20 percent (=.2)
Minimum total sample size	(300/.2)=1500 per health system

Appendix G: English Mailed Survey Materials

Cover letter for initial mailing

Parent or Guardian of [name of child]

Address

City, State, Zip

Dear Parent or Guardian of [name of child]:

You received this survey because your child is 16 or 17 years old and has seen a health care provider in the last 12 months. We would like you to give your child the attached survey to fill out.

This survey is voluntary. If you allow your child to answer the survey, please give it to them to complete. This survey should take 10 minutes or less. If possible, we would like your child to answer the survey on their own. However, it is ok if they need some help from you, for example, to read the questions or to write down the answers for them.

<u>If your child is not able to understand the questions in this survey and answer them at all, please do not answer for them.</u> Please check the box below and return this letter and the survey without completing it. Please do not answer for them.

- ☐ **My child is not able to answer the survey.**

Your child may choose not to answer this survey. This will not affect their medical care in any way.

The information that your child provides will be kept completely private and confidential. Answers will not be matched with your child's name. Their individual answers will never be seen by their provider or anyone else involved with their care. When your child has completed the survey, please mail it back in the envelope that came with it. No postage is needed.

If you have any questions about this survey, please call XX XXX at XXX-XXX.

Sincerely,

XXX XXXX MD

Cover letter for second mailing

Parent or Guardian of [name of child]

Address

City, State, Zip

Dear Parent or Guardian of [name of child]:

About [number of weeks/days] ago, we sent you a survey to give to your child. You received this survey because your child is 16 or 17 years old and has seen a health care provider in the last 12 months. We would like you to give your child the attached survey to fill out. If your child has already returned the survey to us, please accept our thanks and ignore this letter.

This survey is voluntary. If you allow your child to answer the survey, please give it to them to complete. This survey should take 10 minutes or less. If possible, we would like your child to answer the survey on their own. However, it is ok if they need some help from you, for example, to read the questions or to write down the answers for them.

If your child is not able to understand the questions in this survey and answer them at all, please do not answer for them. Please check the box below and return this letter and the survey without completing it. Please do not answer for them.

- ☐ **My child is not able to answer the survey.**

Your child may choose not to answer this survey. This will not affect their medical care in any way.

The information that your child provides will be kept completely private and confidential. Answers will not be matched with your child's name. Their individual answers will never be seen by their provider or anyone else involved with their care. When your child has completed the survey, please mail it back in the envelope that came with it. No postage is needed.

If you have any questions about this survey, please call XX XXX at XXX-XXX.

Sincerely,

XXX XXXX MD

Appendix H: Spanish Mailed Survey Materials

Cover letter for initial mailing

Padre/Madre o Tutor/Guardián Legal de [name of child]

Dirección

Ciudad, Estado, Código Postal

Estimado padre, madre, o tutor/guardián legal de [name of child]:

Usted recibió esta encuesta porque su hijo(a) tiene 16 o 17 años y ha visitado a un proveedor de atención médica en los últimos 12 meses. Nos gustaría que le entregue a su hijo(a) la encuesta que viene incluida para que la complete.

Esta encuesta es voluntaria. Si usted le da permiso a su hijo(a) para que conteste la encuesta, por favor entréguesela para que la complete. Completar la encuesta deberá tomar unos 10 minutos o menos. Si es posible, nos gustaría que su hijo(a) sea quien conteste las preguntas por su cuenta. Sin embargo, si él/ella necesita algo de su ayuda, por ejemplo, que usted le lea las preguntas o le escriba sus respuestas, usted puede hacerlo.

<u>Si su hijo(a) no puede entender las preguntas de esta encuesta y no puede contestarlas en absoluto, por favor no las conteste usted en nombre de él/ella.</u> Por favor marque el cuadrito que está a continuación y devuelva esta carta y la encuesta sin completar.

- ☐ **Mi hijo(a) no puede contestar la encuesta.**

Su hijo(a) puede decidir no contestar esta encuesta. Esa decisión no tendrá ningún efecto en absoluto en su atención médica.

La información que su hijo(a) proporcione se mantendrá de manera totalmente privada y confidencial. Las respuestas no serán asociadas con el nombre de su hijo(a). Sus respuestas individuales nunca serán vistas por su proveedor de atención médica o por alguien más que esté involucrado con su atención médica. Cuando su hijo(a) haya completado la encuesta, por favor envíela por correo en el sobre que le enviamos con la encuesta. . No hace falta poner sellos postales.

Si tiene preguntas acerca de esta encuesta, por favor llame a XX XXX al teléfono XXX-XXXX. Este es un teléfono gratuito.

Le saluda atentamente,

xxxxxxxxx, MD

Cover letter for second mailing

Padre/Madre o Tutor/Guardián Legal de [name of child]

Dirección

Ciudad, Estado, Código Postal

Estimado padre, madre, o tutor/guardián legal de [name of child]:

Aproximadamente [number of weeks/days] (semanas/días) atrás le enviamos una encuesta para su hijo(a).Usted recibió esta encuesta porque su hijo(a) tiene 16 o 17 años y ha visitado a un proveedor de atención médica en los últimos 12 meses. Nos gustaría que le entregue a su hijo(a) la encuesta que viene incluida para que la complete. Si su hijo(a) ya nos ha enviado la encuesta, le estamos muy agradecidos y usted puede ignorar esta carta.

Esta encuesta es voluntaria. Si usted le da permiso a su hijo(a) para que conteste la encuesta, por favor entréguesela para que la complete. Completar la encuesta deberá tomar unos 10 minutos o menos. Si es posible, nos gustaría que su hijo(a) sea quien conteste las preguntas por su cuenta. Sin embargo, si él/ella necesita algo de ayuda, por ejemplo, que usted le lea las preguntas o escriba sus respuestas, usted puede hacerlo.

<u>Si su hijo(a) no puede entender las preguntas de esta encuesta y no puede contestarlas del todo, por favor no las conteste usted en nombre de él/ella.</u> Por favor marque el cuadrito que está a continuación y devuelva esta carta y la encuesta sin completar.

☐ **Mi hijo(a) no puede contestar la encuesta.**

Su hijo(a) puede decidir no contestar la encuesta. Esa decisión no tendrá ningún efecto en absoluto en su atención médica.

La información que su hijo(a) proporcione se mantendrá de manera totalmente privada y confidencial. Las respuestas no serán asociadas con el nombre de su hijo(a). Sus respuestas individuales nunca serán vistas por otro proveedor de atención médica o por alguien más que esté involucrado en su atención médica. Cuando su hijo(a) haya completado la encuesta, por favor envíela por correo en el sobre que le enviamos junto con la encuesta. No hace falta poner sellos postales.

Si tiene preguntas acerca de esta encuesta, por favor llame a XX XXX al teléfono XXX-XXXX. Este es un número de teléfono gratuito.

Le saluda atentamente,

xxxxxxxxx, MD

Appendix I: Overall Scores by Site

ADAPT Unadjusted Measure Scores (UCS) and Case-Mix Adjusted Measure Scores (CMACS) - Overall Scores by Site

Measure	**Hospital 1** (n=293)			**Health Plan 1 (HP 1)** (n=780)			**Health Plan 2 (HP 2)** (n=575)			***P*** **3 sites**	***P*** **HP 1 v HP 2**
	n	UCS (mean)	CMACS*	n	UCS (mean)	CMACS*	n	UCS (mean)	CMACS*		
Counseling on Transition Self-Management	266	32	32 (30,35)	707	36	36 (34,38)	489	30	30 (28,33)	0.028	0.024
Counseling on Prescription Medication	237	61	61 (59,64)	426	57	57 (55,60)	209	58	58 (54,62)	0.267	0.075
Transfer Planning	266	5	5 (3, 7)	704	4	4 (3, 5)	489	3	3 (2, 4)	0.225	0.158

* mean (95% confidence interval)

Appendix J: Disparities Data

ADAPT Survey Scores Based on Race/Ethnicity

Hospital 1	**Counseling on Transition Self-Management** n=292			**Counseling on Prescription Medication** n=258			**Transfer Planning** n=292		
	n	UCS	CMACS*	n	UCS	CMACS*	n	UCS	CMACS*
Asian/Pacific Islander	6	29.2	28.9 (13, 45)	5	60	58.1 (44, 72)	6	0	0
Black	13	33	33.3 (21, 46)	10	59.3	59.9 (44, 76)	13	0	0
Hispanic	18	47.2	47.3 (34, 60)	15	75.6	74.8 (65, 85)	18	12.9	13 (1.4, 25)
White	220	31.1	31.1 (28, 34)	199	61.5	60.8 (58, 64)	220	4.2	4.2 (2.2, 6.1)
Other	9	36.1	36.2 (17, 56)	8	58.3	56.7 (41, 72)	9	16.7	15.8 (0, 37)
P-value	0.128			0.147			0.134		
Health Plan 1	n=757			n=455			n=753		
Asian/Pacific Islander	26	31.1	32.5 (21, 44)	11	53.6	51.9 (30, 73)	26	0	0
Black	195	43.5	43.8 (40, 48)	106	64.2	64.3 (59, 69)	194	3.5	3.5 (1.3, 5.7)
Hispanic	80	35.8	35.8 (30, 42)	39	61.5	61.9 (53, 71)	80	4.1	4 (0.7, 7.3)
White	379	32.2	32.3 (30, 35)	253	53.8	54.2 (51, 58)	377	3.8	3.8 (2.4, 5.3)
Other	27	35.2	35.4 (27, 44)	17	51	49.6 (37, 62)	27	5.6	5.1 (0, 11.5)
P-value	<0.001			0.023			0.69		
Health Plan 2	n=564			n=237			n=561		
Asian/Pacific Islander	29	34.9	33.8 (24, 44)	9	44.4	48 (22, 74)	29	0	0
Black	81	32.2	31.6 (26, 37)	44	62.9	62.1 (55, 69)	81	3.4	3.1 (0, 6.4)
Hispanic	281	31	30.9 (28, 34)	102	57.8	58.2 (53, 63)	281	2.7	2.6 (1.3, 4)
White	77	23.4	23.3 (17, 30)	43	55.8	55.7 (48, 64)	77	2.9	2.9 (0, 6.1)
Other	9	27.8	27.5 (16, 39)	4	66.7	69.7 (68, 71)	9	0	1.4 (0, 5.4)
P-value	0.148			0.506			0.738		
Overall	n=1613			n=950			n=1606		
Asian/Pacific Islander	61	32.6	32.6 (26, 39)	25	51.6	51.6 (38, 65)	61	0	0
Black	289	39.9	39.9 (37, 43)	160	63.5	63.5 (59, 68)	288	3.3	3.3 (1.5, 5.1)

Hispanic	379	32.8	32.8 (30, 35)	156	60.5	60.5 (56, 65)	379	3.4	3.4 (2.1, 4.8)
White	676	30.9	30.9 (29, 33)	495	56.9	56.9 (55, 59)	674	3.8	3.8 (2.7, 4.9)
Other	45	33.9	33.9 (27, 40)	29	54.9	54.9 (46, 63)	45	6.7	6.7 (0.8, 12.5)
P-value	<0.001			0.04			0.179		

* mean (95% confidence interval)

ADAPT Survey Scores Based on Chronic Disease (Hospital 1 not available)

Health Plan 1	**Complex Chronic[a]** n=286			**Non-Complex Chronic[a]** n=394			***P-value***
	n	UCS	CMACS*	n	UCS	CMACS*	
Counseling on Transition Self-Management	345	35.5	35.5 (33, 38)	362	36.1	36.4 (34, 39)	0.631
Counseling on Prescription Medication	243	58	58 (55, 61)	183	55.6	55.8 (51, 60)	0.44
Transfer Planning	344	4	4 (2.4, 5.6)	360	3.3	3.4 (2, 4.9)	0.614
Health Plan 2	n=285			n=288			
Counseling on Transition Self-Management	246	32.4	32.4 (28.9, 35.8)	243	28.6	28.3 (25, 31)	0.087
Counseling on Prescription Medication	114	55.8	55.8 (51, 61)	95	61.1	60.8 (56, 66)	0.174
Transfer Planning	245	2.6	2.6 (1.1, 4)	244	2.8	2.7 (1, 4.3)	0.926
Overall	n=671			n=682			
Counseling on Transition Self-Management	591	34.2	34.2 (32, 36)	605	33.1	33.1 (31, 35)	0.466
Counseling on Prescription Medication	357	57.3	57.3 (54, 60)	278	57.5	57.5 (54, 61)	0.943
Transfer Planning	589	3.4	3.4 (2.3, 4.5)	604	3.1	3.1 (2, 4.2)	0.713

[a] Derived from the Pediatric Medical Complexity Algorithm

* mean (95% confidence interval)

Appendix K: ADAPT Survey Field Test Respondent Characteristics

ADAPT Survey Respondent Characteristics

	Hospital 1	Health Plan 1	Health Plan 2
N	293	780	575
Variable	**n (%)**	**n (%)**	**n (%)**
Sex			
Female	157(53.6)	397(51.1)	323 (56.4)
Male	136 (46.4)	380(48.9)	250 (43.6)
Age (years)*			
16	124(42.3)	350(45.1)	229 (40.1)
17	169(57.7)	426(54.9)	342 (59.9)
Race/Ethnicity			
Hispanic or Latino	19 (6.5)	119 (16.0)	331 (59.0)
American Indian or Alaskan Native	0 (0.0)	2 (0.3)	2 (0.4)
Asian/Pacific Islander	7 (2.4)	34 (4.6)	32 (5.7)
Black, Non-Hispanic	14 (4.8)	178 (24.0)	101(18.0)
White, Non-Hispanic	244 (83.3)	386 (52.0)	87 (15.5)
Other	1 (0.3)	0 (0.0)	0 (0.0)
Multiple	8 (2.7)	23 (3.1)	8 (1.5)
Education			
9th grade or less	21 (7.2)	112 (14.6)	57 (9.9)
10th grade	114 (39.0)	284 (37.1)	188 (32.7)
11th grade	136(46.6)	299 (39.0)	222 (38.6)
12th grade or some college	21 (7.2)	71 (9.3)	108 (18.8)
Health Insurance**			
Private	207 (70.6)	0	0
Public	86 (29.4)	780 (100.0)	575 (100.0)
Health Condition Category***			
Complex Chronic	N/A	386 (49.5)	285 (49.7)
Non-Complex Chronic		394 (50.5)	288 (50.3)

*A few were completed by 15 year olds (n=7 in health plan #2) or 18 year olds (n=24 in Hospital 1, n=44 in health plan #1, and n=63 in health plan #2)

** All health plan respondents are enrolled in Medicaid

*** Participants from Hospital 1 were identified as receiving specialty care for a chronic health condition. The sample from the Health Plans was identified using the Pediatric Medical Complexity Algorithm,[15] which results in 2 distinct categories of health condition.

Appendix L: Decision Rules and Coding Guidelines

To ensure accurate collection of all survey data, quality control procedures should be developed, implemented, and documented for all survey administration activities. The ADAPT survey decision rules and coding guidelines were developed to capture appropriate information for data submission. They provide guidance for addressing situations in which survey responses are ambiguous, missing or incorrectly provided. Adherence to the following decision rules and coding guidelines should ensure valid and consistent coding of such instances.

Multiple returned surveys from the same respondent

If health plans, practices, hospitals, or survey vendors administer the ADAPT survey using a multiple-wave mail protocol, it is possible for a respondent to receive, complete and return multiple surveys. When multiple surveys from the same respondent are received, code the first returned completed survey.

Coding ambiguous responses

A common problem in mailed surveys is ambiguity of responses on returned surveys. To ensure uniformity in data coding, strictly apply the following guidelines. When scanning or key-entering paper-based surveys, use the following decision rules for resolving common ambiguous situations:

- If a value is missing, then code the value as ". Missing." A response should not be imputed; in other words, do not try to determine what the respondent would have responded for the missing value based on answers to other questions. Except
- If a mark falls between two response options but is obviously closer to one than the other, then select the choice to which the mark is closest.
- If a mark falls equidistant between two response options, then code the value as ". Missing".
- If more than one response option is marked for Questions 4, 5, 6, 8, 12, 13, 16, 18, and 20 (i.e., both yes and no are marked, code the value as ". Missing").
- If more than one response option is marked for Questions 2, 3, 7, 10, 19, 21, and 22, code the option that represents the highest level of value to this question, e.g., higher level of school completed or more frequent visits.
- There are 7 screener questions in this survey (Questions 3, 9, 11, 14, 15, 17, and 25). When more than one response option is marked for any of the screener questions, the decision of which option to code depends on how the associated dependent question(s) is answered:
    - If the associated dependent question(s) is answered, code the option of the screener question that allows the dependent question(s) to be answered.
    - If the associated dependent question(s) is not answered, code the option of the screener questions that allows the dependent question(s) to be skipped.
    - Exception: for Question 14, "Does this provider treat mostly children and teens?", if more than one option is marked, code "Don't know".

In instances in which multiple options are marked **but** the respondent's intent is clear, code the respondent's **clearly identified** intended response.

For question 23 “Are you of Hispanic, Latino, or Spanish origin? Mark one or more,” and question 24 “How would you describe your race? Mark one or more,” enter responses for all of the categories that the respondent has selected.

Skip patterns

Several questions in the ADAPT survey can and should be skipped by certain respondents. These questions form skip patterns. Seven questions in the ADAPT survey serve as screener questions (Questions 3, 9, 11, 14, 15, 17, and 25) that determine whether the associated dependent questions should be answered. The following decision rules are provided to assist in coding responses to skip pattern questions.

Decision Rules for Transfer Planning Measure Questions 16, 17, and 18

If Question 15 is “No”, then code Questions 15, 16, 17, and 18 as ”No.”

If Question 15 is “Yes” or left blank, enter the value provided by the respondent for Questions 16, 17, and 18, except as follows:

- If Question 17 is “No” and Question 18 is left blank or **not** left blank, then code the value of Question 18 as “No”
- If Question 17 is left blank and Question 18 is **not** left blank, then code Question 17 as “.Missing” and enter the value provided by the respondent for Question 18.

Decision Rules for Screener and Dependent Questions

Decision rules for coding **screener questions** (Questions 3, 9, 11, 14, and 25; Does not apply to Question 15 or 17):

- Enter the value provided by the respondent. Do not impute a response based on the respondent’s answers to the dependent questions.
- If a screener question is left blank, then code the value as “. Missing.” Do no impute a response based on the respondent’s answers to the dependent questions.
- In the situation where more than one option is marked for a screener question, see rules in the “Coding Ambiguous Responses” section.

Decision rules for coding **dependent questions** (Questions 4-14, and 26; Does not apply to Questions 15-18):

- If the marked screener question option requires the dependent question(s) to be answered, and the dependent question(s) is left blank, then code the value for the dependent questions(s) as “. Missing.”
- If the marked screener question option requires the dependent question(s) to be answered, and the dependent question(s) is **not** left blank, then enter the value provided by the respondent for the dependent question(s).
- If the marked screener question option requires the dependent question(s) to be skipped, and the dependent question(s) is left blank, then code the value for the dependent questions(s) as “. Missing.”
- If the marked screener question option requires the dependent question(s) to be skipped, and the dependent question(s) is **not** left blank, then code the value for the dependent questions as “. Missing.”

- If the screener question is left blank and the dependent question(s) is left blank, then code the value for both the corresponding screener question and the dependent question(s) as ". Missing."
- If the screener question is left blank and the dependent question(s) is **not** left blank, then code the value for the corresponding screener question as ". Missing" and enter the value provided by the respondent for the dependent question(s).

Recoding and collapsing variables

In instances in which some variables need to be recoded or collapsed for analysis or reporting, the following rules can be used.

Collapsing **Race and Ethnicity** from Question 23 (Are you of Hispanic, Latino, or Spanish origin? Mark one or more) and Question 24 (How would you describe your race? Mark one or more):

- If Question 23 is marked "Yes", including "Yes, Mexican, Mexican American, Chicano", "Yes, Puerto Rican," "Yes, Cuban," or "Yes, another Hispanic, Latino, or Spanish origin", code the respondent as "Hispanic" regardless of what race(s) is marked.
- If Question 23 is marked "No, not of Hispanic, Latino, or Spanish origin" and only one option of Question 24 is marked, code the respondent as their marked race, for example "White Non-Hispanic", "Black Non-Hispanic", "American Indian or Alaska Native Non-Hispanic."
- If Question 23 is marked "No, not of Hispanic, Latino, or Spanish origin" and multiple races are marked for Question 23, code the respondent as "Multi-Racial."

Recoding **Help Received to Complete this Survey** from Question 26 (How did that person help you? Mark one or more):

- If only one option is marked for Question 26, code the recoded variable as their marked level of help, for example "Read the questions only", "Wrote the answers only", "Helped in some other way only."
- If multiple options are marked for Question 26, code the recoded variable as "Helped in multiple ways."

Appendix M. Evidence Table

Literature assessing patient experience of health care transition preparation across various health conditions		
Multiple Complex Conditions		
Study Design	**Description**	**Reference**
Cross-sectional Survey-based; young adult report	This study examined factors associated with receipt of health care transition (HCT) counseling services as reported by young adults. The 2007 Survey of Adult Transition and Health (SATH) was used to explore self-reported receipt of transition support. Of 1865 SATH respondents, 55% reported that their health care providers had discussed how their needs would change with age, 53% reported that providers had discussed how to obtain health insurance as an adult, and 62% reported having participated in a transition plan. Only 24% reported receiving all three transition counseling services. Provider-youth communication was associated with increased health care transition guidance.	Sawicki GS, Whitworth R, Gunn L, Butterfield R, Lukens-Bull K, Wood D. Receipt of health care transition counseling in the national survey of adult transition and health. *Pediatrics.* 2011;128(3):e521-529.
Cross-sectional Survey-based; parent report	This study analyzed data from 17,114 parent respondents to the 2009–2010 National Survey of Children with Special Health Care Needs, which was fielded to assess transition preparation among youth with special health care needs (YSHCN). With only 40% of YSHCN receiving elements of recommended transition care , researchers concluded that most YSHCN are not receiving needed transition preparation. Although most providers are encouraging YSHCN to assume responsibility for their own health, far fewer are discussing transfer to an adult provider and insurance continuity.	McManus M, Pollack LR, Cooley WC, et al. Current Status of Transition Preparation Among Youth with Special Needs in the United States. *Pediatrics.* 2013;131(6):1090-1097.

Cross-sectional Qualitative; young adult, caregiver, and provider report	To examine transition experiences, including facilitating practices and obstacles, this study analyzed focus groups and interviews of 143 young adults with disabilities and special health care needs, their family members, and their health care providers. Findings showed that pediatric and adult-oriented providers represent different medical subcultures. Young adults' and family members' lack of preparation for successful participation in the adult health care system contributes to problems with health care transition was common.	Reiss JG, Gibson RW, Walker LR. Health care transition: youth, family, and provider perspectives. *Pediatrics.* 2005;115(1):112-120.
Cross-sectional Survey-based; parent report	This study aimed to provide a baseline measure of the proportion of US children who meet the six Maternal and Child Health Bureau's (MCHB) core outcomes for youth with special health care needs (YSHCN). One of these core outcomes is that youth will receive the services necessary to make transitions to adult life, including adult health care, work, and independence. Results from the 2001 National Survey of CSHCN (n=38,866) and the 2001 National Health Interview Survey (n=13,579) showed that only 6% of youth aged 13 and older are receiving the services needed to successfully transition. These services include support and communication about changing to an adult provider and changing health care needs with age.	McPherson M, Weissman G, Strickland BB, van Dyck PC, Blumberg SJ, Newacheck PW. Implementing community-based systems of services for children and youths with special health care needs: how well are we doing? *Pediatrics.* 2004;113(5 Suppl):1538-1544.

Autism		
Cross-sectional Survey-based; parent report	This study examined the receipt of HCT services in youth with autism spectrum disorder (ASD). The 2005–2006 National Survey of Children with Special Health Care Needs (NSCSHCN) was used to examine receipt of HCT services for youth with ASD. Only 14% of youth with ASD had a discussion with their pediatrician about transitioning to an adult provider, less than a quarter had a discussion about health insurance retention, and just under half had a discussion of adult health care needs or were encouraged to take on appropriate responsibility.	Cheak-Zamora NC, Yang X, Farmer JE, Clark M. Disparities in transition in planning for youth with autism spectrum disorder. *Pediatrics*. 2013. 131(3):447-454.
Cancer Survivorship		
Cross-sectional Qualitative interview and focus group based; youth, young adult and parent report	This study sought to determine barriers or facilitators to transition from pediatric to adult-centered survivorship care as perceived by Latino adolescent and young adults (AYA) cancer survivors and their parents. Twenty-seven Latino AYA (aged $\geq$15 years) completed interviews, and 21 Latino parents participated in focus groups. Both AYA survivors and parents identified two major facilitative factors for survivorship care: involvement of the nuclear family in care at the adult setting and inclusion of symptom communication in late effects discussions. Barriers included perceived stigma of a cancer history and emotional trauma related to discussions about the childhood cancer experience.	Casillas J, Kahn KL, Doose M, et al. Transitioning childhood cancer survivors to adult-centered health care: insights from parents, adolescent, and young adult survivors. *Psychooncology*. 2010; 19(9):982-990.

Cross-sectional Survey-based; adult report	This study assessed knowledge of adult survivors of childhood cancer about their primary cancer diagnosis and therapies. A telephone survey of 635 survivors was conducted. Overall, 72% accurately reported their diagnosis with precision and 19% were accurate but not precise, compared with medical record documentation. History of receiving a written medical summary, attending a long-term follow-up clinic, and anxiety about late effects were not associated with greater knowledge. Investigators noted that these knowledge deficits could impair survivors' ability to seek and receive appropriate follow-up care.	Kadan-Lottick NS, Robison LL, Gurney JG, et al. Childhood cancer survivors' knowledge about their past diagnosis and treatment: Childhood Cancer Survivor Study. *JAMA.* 2002; 287(14):1832-1839.
Congenital Heart Disease		
Cross-sectional Qualitative interview based; youth and parent report	This qualitative study explored how 50 youth and 28 parents affected by congenital heart disease (CHD) and cystic fibrosis (CF) negotiate constructions of 'normal developmental time' in both anticipating and dealing with the transition from adolescence to adulthood. Concerns related to deteriorating health and occupational restrictions in the future were paramount for youth with CHD and CF. For young women, the loss of 'normal' gendered roles was also a concern. Attending to youth's temporal anxieties and future concerns may enhance the transition process for youth with CHD and CF.	Moola FJ, Norman ME. 'Down the rabbit hole': enhancing the transition process for youth with cystic fibrosis and congenital heart disease by re-imagining the future and time. *Child Care Health Dev.* 2011; 37(6): 841-851.

Cross-sectional Survey-based; young adult report	This study aimed to determine the percent of young adults with congenital heart defects (CHDs) who successfully transferred from pediatric to adult care and examine correlates of successful transfer. Two hundred thirty-four patients aged 19-21 completed the measure. In the total cohort, 47% had transferred to adult care. More than one quarter of the patients reported having had no cardiac appointments since 18 years. Successful transfer was associated with more pediatric cardiovascular surgeries, older age at last visit to the pediatric hospital, recommended follow-up at a CACH center, patient beliefs that adult CHD care should be at a CACH center, and attending cardiac appointments without parents or siblings.	Reid GJ, Irvine MJ, McCrindle BW, et al. Prevalence and correlates of successful transfer from pediatric to adult health care among a cohort of young adults with complex congenital heart defects. *Pediatrics*. 2004; 113(3 Pt 1):e197-205.
Cystic Fibrosis		
Cross-sectional Survey-based; adult report	This study surveyed adult CF patients on their concerns regarding a transition program. A survey was sent to members of the International Association of Cystic Fibrosis Adults (IACFA), n = 334. The majority of patients (81%) received care from a CF center. Nearly one-fourth of patients seen at a CF center continued to receive care from a pediatrician (mean age 30 years). Those patients seen in an adult program described criteria for their transfer to the adult program, but no findings suggested a standard transition program. The patients reported their level of concern about transfer as minimal, far less than what CF physicians had perceived, which may impede successful transition of patients into an adult program.	Anderson DL, Flume PA, Hardy KK, Gray S. Transition programs in cystic fibrosis centers: perceptions of patients. *Pediatr. Pulmonol*. May 2002;33(5):327-331.

Cross-sectional Qualitative interview and survey-based; youth and young adult report	The purpose of this study was to investigate how adolescents and adults with CF view preventative counseling and their transition to adult-centered care within a children's hospital. Thirty-two patients ≥16 years old diagnosed with CF were recruited from a pediatric tertiary care setting to undergo interviews and complete a self-administered questionnaire on preventive counseling by health care providers and transition issues. Participants felt that 13-16 years of age was the best time for them to begin spending time alone with their main doctor. Less than half of the participants recalled receiving preventive counseling during the previous 12 months. Qualitative data emphasized the importance of independence in making decisions in health care and establishing relationships with providers, and many patients did not desire to transfer care to an adult hospital. Participants identified adult-focused services such as inpatient rooms, discussion groups, work options, and social service support that would enhance care.	Zack J, Jacobs CP, Keenan PM, Harney K, Woods ER, Colin AA, Emans SJ. Perspectives of patients with cystic fibrosis on preventive counseling and transition to adult care. *Pediatr. Pulmonol*. 2003; 36(5): 376-383.
Cross-sectional Interview and survey-based; youth and parent report	To aid in the development of CF specific transition guidelines, a pre-transition questionnaire and post-transition interview were used to assess the concerns and expectations of 60 CF patients and their parents as they underwent transition from pediatric to adult care. The two most important concerns identified by patients prior to transition were potential exposure to infection and having to leave their previous caregivers. Introduction to the adult CF team prior to transition was associated with lower levels of concern in all areas. Parents' most significant concern was the ability of their child to care for their disease independently.	Boyle MP, Farukhi Z, Nosky ML. Strategies for improving transition to adult cystic fibrosis care, based on patient and parent views. *Pediatr. Pulmonol*. 2001; 32(6): 428-436.

Cross-sectional Survey-based; young adult and adult report	This study examined participation in health behaviors, health locus of control, and negotiation of developmental tasks of adulthood with 75 patients with CF, aged 18–42 years old. Participants completed the Multidimensional Health Locus of Control Scale. Results indicate a number of behaviors for which respondents had not yet assumed responsibility, such as managing medical insurance and nutrition. Respondents were compliant with their medical regimen currently than when first assuming responsibility for their health as adolescents.	Hamlett KW, Murphy M, Hayes R, Doershuk C. Health independence and developmenal tasks of adulthood in cystic fibrosis. *Rehabilitation Psychology*. 1996; 41(2): 149-160.
Cross-sectional Survey-based; parent report	This study described the development and psychometric properties of a survey tool designed to evaluate children's level of independence in their CF treatment, as this may have a direct effect on their involvement in treatment and adherence. The Self-Care Independence Scale (SCIS) was completed by parents of 76 CF patients (ages 4-17 years). Results support the SCIS as a sound measure of CF self-care independence. The SCIS may be used as a screening tool for adolescents preparing to transition to adult CF centers care.	Patton SR, Graham JL, Varlotta L, Holsclaw D Jr. Measuring self-care independence in children with cystic fibrosis: the Self-Care Independence Scale (SCIS). *Pediatr. Pulmonol*. 2003; 36(2): 123-130.
Cross-sectional Survey-based; youth and young adult report	A survey of adolescents and young adults with CF attending an adult CF center was conducted to evaluate a transition program as a means of transferring care from pediatric to adult setting. Forty patients completed a self-administered questionnaire. Most thought that the transition program made the change from pediatric to adult care easier. Of the 40 patients, 17 (42%) recommended that other patients go through the transition program. Twenty-six patients (65%) preferred the adult program. These findings suggest that adolescents with CF should be encouraged to transfer to an adult CF center once they have reached an	Nasr SZ, Campbell C, Howatt W. Transition program from pediatric to adult care for cystic fibrosis patients. *J. Adolesc. Health.*1992;13(8): 682-685.

	agreed-upon age.	

Type 1 Diabetes		
Cross-sectional Survey-based; young adult report	This study examined characteristics of the transition from pediatric to adult care in emerging adults with type 1 diabetes and evaluated associations between transition characteristics and glycemic control. A survey was developed and mailed to 484 diabetic adults aged 22-30 years, receiving adult diabetes care at a single center. Current A1C data were obtained from the medical record. Response rate was 53% (n = 258). The mean transition age was 19.5 ± 2.9 years, and 34% reported a gap >6 months in establishing adult care. Common reasons for transition included feeling too old (44%), pediatric provider suggestion (41%), and college (33%). Less than half received an adult provider recommendation and <15% reported having a transition preparation visit or receiving written transition materials. Respondents who felt mostly or completely prepared for transition had lower likelihood of a gap between pediatric and adult care. There was no independent association of preparation with post-transition A1C.	Garvey KC, Wolpert HA, Rhodes ET, et al. Health care transition in patients with type 1 diabetes: young adult experiences and relationship to glycemic control. *Diabetes Care*. 2012; 35(8):1716-1722
Cross-sectional Survey-based; young adult report	Youth with type 1 diabetes are at risk for poor glycemic control as they age into adulthood. The aim of this study was to describe sociodemographic and clinical correlates of poor glycemic control associated with the transfer of care from pediatric to adult diabetes providers. One hundred eighty-five adolescents ≥18 with recently diagnosed type 1 diabetes participated. At a follow-up visit, 57% had transitioned to adult diabetes providers. The estimated median age of transition of care was 20.1 years. Older age, lower baseline glycosylated hemoglobin, and less	Lotstein D, Seid M, Klingensmith G, et al. Transition from pediatric to adult care for youth diagnosed with type 1 diabetes in adolescence. *Pediatrics.* 2013;131(4):1062-1070.

	parental education were associated with increased odds of transition. The odds of poor glycemic control at follow-up were 2.5 times higher for participants who transitioned to adult care compared with those who remained in pediatric care.	
Cross-sectional Survey and qualitative interview-based; adolescent, young adult and parents report	Participants completed questionnaires and responded to open-ended questions regarding self-management, self-efficacy, expectations and experiences with pediatric and adult care providers across the transition process. At a mean age of 16.1 years, most pre-transition adolescents had not yet discussed transferring care with their parents or doctors. Although many post-transition young adults reported positive interactions, several described challenges locating or establishing a relationship with an adult diabetes care provider. Qualitative themes emerged related to the anticipated timing of transfer, early preparation for transition, the desire for developmentally appropriate interactions with providers, the maintenance of family and social support, and strategies for coordinating care between pediatric and adult providers.	Hilliard ME, Perlus JG, Clark LM, et al. Perspectives From Before and After the Pediatric to Adult Care Transition: A Mixed-Methods Study in Type 1 Diabetes. *Diabetes Care.* 2014; 37(2):346-354.
Human Immunodeficiency Virus (HIV)		

Cross-sectional Qualitative interview based; youth, young adult, and parent report	Semi-structured interviews were conducted with 40 perinatally infected adolescents (mean age 17 years), currently receiving care in a pediatric infectious disease clinic and 17 guardians, about their expectations related transition. Many adolescents reported that they did not know what to expect out of the transition. Others looked forward to increased responsibility and control, while some expressed concerns over leaving their current providers. Most guardians viewed the transition to adult care as a tool to facilitate maturity. Several indicated they had not discussed transition with their child and were waiting for their child to initiate the discussion. The results indicate a need for improved communication between youth and providers to enhance transition success.	Fair CD, Sullivan K, Dizney R, Stackpole A. "It's like losing a part of my family": transition expectations of adolescents living with perinatally acquired HIV and their guardians. *AIDS Patient Care STDS.* 2012; 26(7):423-429.

Inflammatory Bowel Disease		
Cross-sectional Survey-based; youth report	This study sought to determine whether adolescents with inflammatory bowel disease (IBD) have developed key skills of self-management prior to the age at which many transfer to adult care. Adolescents aged 16 to 18 years old in the Boston Children's Hospital IBD database (43 total) responded to a mailed survey assessing knowledge and confidence of their own health information and behaviors. Respondents could name medication and dose with confidence but had very poor knowledge of important side effects. Most patients deferred responsibility mostly or completely to parents for scheduling appointments, requesting refills, or contacting provider between visits.	Fishman LN, Barendse RM, Hait E, Burdick C, Arnold J. Self-management of older adolescents with inflammatory bowel disease: a pilot study of behavior and knowledge as prelude to transition. *Clin. Pediatr. (Phila).* 2010; 49(12):1129-1133.

Cross-sectional Survey-based; youth report	This study evaluated the knowledge of 78 adolescents ages 14-18 with inflammatory bowel disease (IBD) and 64 of their parents. Patients and their parents completed the MyHealth Passport for IBD and responses were evaluated for accuracy using medical records. Patients and parents were equally likely to answer questions correctly regarding disease characteristics and treatment, but not health services resources. Most patients accurately identified IBD classification and listed medications. Neither patients nor parents accurately identified disease location or previous investigation results. Parents were more likely to name insurance provider and pharmacy location. Future educational interventions should target areas of weakness in adolescent knowledge.	Benchimol EI, Walters TD, Kaufman M, et al. Assessment of knowledge in adolescents with inflammatory bowel disease using a novel transition tool. *Inflamm. Bowel Dis.* 2011;17(5):1131-1137.
Kidney Disease and Transplant		
Cross-sectional Survey-based; youth, young adult, and parent report	The goal of this study was to develop a measure of transition readiness for adolescent kidney transplant recipients. The Readiness for Transition Questionnaire (RTQ-teen; RTQ-parent) was created to assess overall transition readiness, adolescent health care behavior, and familial involvement in health care. Participants were 48 adolescent kidney transplant recipients, ages 15-21 years and 32 of their caregivers. Adolescents completed the RTQ-teen and self-reported measures of adherence. Parents completed the RTQ-parent. The RTQ showed good internal consistency, inter rater reliability, and demonstrated construct validity. Increased adolescent responsibility and decreased parental involvement predicted higher transition readiness. Greater adolescent adherence factors predicted greater transition readiness.	Gilleland J, Amaral S, Mee L, Blount R. Getting ready to leave: transition readiness in adolescent kidney transplant recipients. *J Pediatr Psychol*. 2012; 37(1):85-96.

Cross-sectional Survey-based; youth and young adult report	The purpose of this study was to describe and compare mastery of health care management in adolescent (aged 14-17 years) and young adult (age ≥ 18 years) recipients of a liver transplant expected to transfer from pediatric to adult care settings. Fifty-two liver transplant recipients completed the Developmentally Based Skills Checklist, which asks how often patients independently engage in specific health care management skills. Overall, young adult patients reported greater health care management than adolescents. However, less than half of the young adults surveyed reported consistently managing their liver disease independently, making their own appointments, and understanding insurance issues.	Annunziato RA, Parkar S, Dugan CA, et al. Brief report: Deficits in health care management skills among adolescent and young adult liver transplant recipients transitioning to adult care settings. *J. Pediatr. Psychol.* 2011; 36(2):155-159.
Cross-sectional Survey-based; youth, young adult and parent report	This study aimed to determine adolescent and young adult liver transplant recipient (LTR) and parent perceptions about the transition process. Participants included 46 LTR (mean age 16.6 years) and 31 parents. Recipients and parents reported moderate concern about transition. LTR ≥16 yr reported greater health care responsibility and increased thought, interest, and knowledge about transition. LTR perceive having more independence than their parents are report.	Fredericks EM, Dore-Stites D, Lopez MJ, et al. Transition of pediatric liver transplant recipients to adult care: patient and parent perspectives. *Pediatr. Transplant.* 2011;15(4): 414-424.
Rheumatologic Disease		

Cross-sectional Survey-based; parent report	This study used data from the 2005-2006 National Survey of Children with Special Health Care Needs (NS-CSHCN) to determine the proportion of adolescents with arthritis who receive health care transition (HCT) services and compared the rates with those reported for adolescents with other special health care needs. Parents of youth with arthritis (n = 1,052), diabetes (n = 389), and other special health care needs (n = 18,189) responded. Many adolescents with arthritis are being encouraged to assume self-care responsibilities (74.8%); fewer discussed changing health needs in adulthood (52.1%), acquiring insurance (22.5%), or transferring care to a provider who sees adults (19.0%). These results are similar to youth with other special health care needs, but behind youth with diabetes.	Scal P, Horvath K, Garwick A. Preparing for adulthood: health care transition counseling for youth with arthritis. *Arthritis and Rheumatism*. 2009;61(1):52-57.
Cross-sectional Qualitative focus groups and interviews; youth report	This study sought to explore the self-management needs of adolescents with juvenile idiopathic arthritis and the acceptability of a Web-based self-management program. A convenience sample of 36 adolescents who varied in age, gender, disease onset subtype, and disease severity were recruited from 4 Canadian tertiary care pediatric centers. Individual (n=25) and 3 focus-group (n=11) interviews were conducted Adolescents articulated how they developed effective self-management strategies through the process of "letting go" from others who had managed their illness (health care professionals, parents) and "gaining control" over managing their illness on their own. The 2 strategies that assisted in this process were gaining knowledge and skills to manage the disease and experiencing understanding through social support.	Stinson JN, Toomey PC, Stevens BJ, Kagan S, Duffy CM, Huber A, Malleson P, McGrath PJ, Yeung RS, Feldman BM. Asking the experts: exploring the self-management needs of adolescents with arthritis. *Arthritis Rheum.* 2008; 59(1):65-72.

Sickle Cell Disease		
Cross-sectional Survey-based; young adult report	The goal of this study is to assess transition readiness of patients with sickle cell disease (SCD) in a transition program and to evaluate a SCD-specific assessment tool that measures 5 knowledge skill sets and 3 psychological assessments. Of the 47 patients between the ages of 18 and 22, 33 completed the assessment. The majority of patients reported good medical knowledge of SCD. There were knowledge gaps in the areas of independent living skills. A majority of patients reported being worried that SCD would prevent them from doing things in their life; however, few said they were worried or anxious about transitioning to adult care.	Sobota A, Akinlonu A, Champigny M, et al. Self-reported Transition Readiness Among Young Adults With Sickle Cell Disease. *J. Pediatr. Hematol. Oncol.* 2014. doi:10.1097/ MPH.0000000000000110.
Cross-sectional Survey-based; youth and young adult report	This study sought to assess adolescent SCD patients' preparation for transition and identify variables that predict patient readiness. Seventy adolescent patients (14 to 20 years) receiving care at a pediatric SCD center completed a survey about the transition from pediatric to adult care. Mean readiness scores were low, with greatest deficiencies in prior thought, knowledge, anticipated difficulty, and interest regarding transition. Younger age was associated with less knowledge and interest; disease severity was associated with lower interest but greater anticipated difficulty. Adolescents with SCD demonstrate poor preparation for transition to adult-oriented care.	McPherson M, Thaniel L, Minniti CP. Transition of patients with sickle cell disease from pediatric to adult care: Assessing patient readiness. *Pediatric blood & cancer.* 2009;52(7):838-841.
Cross-sectional Interview-based; youth and young adult report	This study identified concerns and expectations of pediatric SCD patients as they begin to transition to adult care, as well as what program priorities they perceive would facilitate a smooth transition. Data were collected by means of interviews. The sample included 172	Telfair J, Ehiri JE, Loosier PS, Baskin ML. Transition to adult care for adolescents with sickle cell disease: results of a national survey. *International journal of adolescent medicine and health.* 2004;16(1):47-64.

	adolescents with SCD aged 14 years and older still in pediatric care within community-based and medical center SCD programs. The top concerns of adolescents were: lack of information relating to their transition; fear of leaving their familiar health care provider, fear that adult providers might not understand their needs; belief that an SCD transition program was needed; information provision about adult care programs; ways to meet adult providers; and ways to help health care providers understand their needs.	
Cross-sectional Survey-based; youth, young adult and parent report	This study determined the issues, concerns, and expectations of adolescents and young adults with SCD and primary caretakers with regard to transfer to adult care. Results revealed that adolescents and young adults with sickle cell disease were concerned about how they would pay for medical care and how they would be treated by adult providers. Caretakers were concerned about their teens leaving pediatric care and assuming responsibility for care. All three groups reported mixed emotions about leaving pediatric care.	Telfair J, Myers J, Drezner S. Transfer as a component of the transition of adolescents with sickle cell disease to adult care: adolescent, adult, and parent perspectives. *J. Adolesc. Health.* 1994; 15(7):558-565.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2797
Measure Title: Transcranial Doppler Ultrasonography Screening Among Children with Sickle Cell Anemia
Measure Steward: Q-METRIC – University of Michigan
Brief Description of Measure: The percentage of children ages 2 through 15 years old with sickle cell anemia (Hemoglobin SS) who received at least one transcranial Doppler (TCD) screening within a year.
Developer Rationale: Children with sickle cell anemia (HbSS) have over three hundred times the stroke risk than children with normal hemoglobin (Verduzco and Nathan, 2009). Without intervention, approximately 11% of children with sickle cell anemia will have a stroke by age 20 (Verduzco and Nathan, 2009; Ohene-Frempong et al., 1998). Transcranial Doppler (TCD) ultrasonography measures the blood velocities within the cerebral vessels (Adams et al., 1997; Adams et al., 1992). Children over the age of 2 with a time-average mean maximum blood flow velocity of 200cm/sec or greater as measured by TCD ultrasonography have been shown to have 27 times the risk of stroke than children with velocities less than 200cm/sec. This corresponds to a 40% risk of stroke among those with high velocities within 3 years (Adams et al., 1997). Initiation of chronic blood transfusions reduces the risk of stroke by 92% among children at highest risk of stroke as identified through TCD screening (Adams et al., 1997; Adams et al., 1992). TCD screening is a reasonable method to assess stroke risk among children with sickle cell anemia, as it is safe, non-invasive and low cost (Markus, 2000). Although other predictors of stroke have been examined, such as hematocrit levels and white blood cell count, TCD velocities have been shown to be the only independent predictor of stroke (Adams et al., 1992). Given the importance of TCD screening to stroke prevention among children with sickle cell anemia, the National Heart, Lung, and Blood Institute (NHLBI) recommends each child with sickle cell anemia receive one TCD screen per year from ages 2 to 16 years (National Heart, Lung, and Blood Institute, 2014). Although the benefits of TCD screening among children with sickle cell anemia have been known since the late nineties, prior studies indicate that TCD screening rates are low. However, these reports are limited in their generalizability, as they are often focused on a single healthcare provider or registry. This measure establishes a claims-based method for identifying receipt of TCD screening among larger and broader populations of children with sickle cell anemia. The measure specifications are reflective of the guidelines from the NHLBI, and the performance scores calculated through this measure will identify areas in need of improvement in receipt of TCD screening among children with sickle cell anemia.

Citations:

Adams RJ, McKie VC, Carl EM, et al. Long-term stroke risk in children with sickle cell disease screened with transcranial Doppler. Ann Neurol. Nov 1997;42(5):699-704.

Adams R, McKie V, Nichols F, et al. The use of transcranial ultrasonography to predict stroke in sickle cell disease. N Engl J Med. Feb 27 1992;326(9):605-610.

Markus HS. Transcranial Doppler ultrasound. Br Med Bull. 2000;56(2):378-388.

National Heart Lung and Blood Institute. Evidence Based Management of Sickle Cell Disease. 2014; http://www.nhlbi.nih.gov/health-pro/guidelines/sickle-cell-disease-guidelines/sickle-cell-disease-report.pdf. Accessed 11/11, 2014.

Ohene-Frempong K, Weiner SJ, Sleeper LA, et al. Cerebrovascular accidents in sickle cell disease: rates and risk factors. Blood. Jan 1 1998;91(1):288-294.

Verduzco LA, Nathan DG. Sickle cell disease and stroke. Blood. Dec 10 2009;114(25):5117-5125.

Numerator Statement: The numerator is the number of children ages 2 through 15 years old with sickle cell anemia who received at least one TCD screening within the measurement year.

Denominator Statement: The denominator is the number of children ages 2 through 15 years with sickle cell anemia within the measurement year.
Denominator Exclusions: There are no denominator exclusions.

Measure Type: Process
Data Source: Administrative claims
Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measures evaluation and voting processes. The preliminary analysis will help to guide the Standing Committee evaluation of each measure by summarizing the measure developer submission, guide measure evaluation discussion, and identify topic areas for additional input. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- The evidence for this process measure is based on clinical practice guidelines for management of Sickle Cell Disease from the National Heart, Lung, and Blood Insititute. Dated 2014, this is a strong recommendation with moderate quality evidence (the second highest ranking in this grading system). The recommendation is: "In children with SCA, screen annually with TCD according to methods employed in the STOP studies, beginning at age 2 and continuing until at least age 16."
- The guideline is based on two RCTs and 50 observational studies enrolling more than 11,000 patients.
- Receipt of TCD screening does not directly impact the risk of stroke among children with sickle cell anemia, however, the screening allows identification of children at high risk and prompts the initation of primary stroke prevention (blood transfusions).
- Studies reported between 2% and 33% abnormal TCD screening results within their study populations; this large range may be attributable to differing study population inclusion criteria. All studies investigating the relationship between blood flow velocity as detected by TCD screening and stroke risk show that children with high blood flow velocities in the cerebral vessels are at a significantly increased risk of stroke.
- All studies that assessed stroke rates pre- and post-TCD screening recommendations found a significantly decreased rate of first stroke among children with sickle cell anemia post-TCD recommendations when compared with the pre-TCD recommendation time period.

Question for the Committee:

- *The specifications focus on children ages 2 to 15; the guidelines recomend "children...beginning age 2 and continuing until at least age 16." Does the Committee wish to clarify why the specifications differ?*

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- Children with sickle cell anemia have more than 300 times the stroke risk of children with normal hemoglobin.

Without intervention, approximately 11% of children with sickle cell disease will have a stroke by age 20. Initiation of chronic blood transfusions reduces the risk of children at highest risk of stroke by 40%. TCD velocities, as measured by TCD ultrasonography, has been shown to be the only independent predictor of stroke.

- TCD screening rates from 2010 among children enrolled in Medicaid range from 28.5% (IL) to 50.7% (SC). Based on the Medicaid data provided, there are no gender disparities, and since the data is state Medicaid based, disparities were not identified by insurance or socioeconomic status. Younger children were more likely to be screened.

Questions for the Committee:

- *Are there any data from commercial populations? Any data more recent than 2010 that would inform whether a gap still exists?*
- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities—i.e., between Medicaid and commercial plans—that exist in this area of healthcare?*
- *Should this measure be indicated as disparities sensitive?*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b)

1a. Evidence.

- Clarify denominator (may be ok depending on how you interpret documentation). Is measurement really the key outcome or prevention of strokes? This is indirect...
- Systemic review for this process measure is supported by a clinical practice guideline recommendation (NHLBI-Evidence based Management of Sickle Cell Disease). Recommend that the developer clarify why the age specification differed from the practice guideline.
- There is strong evidence that risk of stroke can be predicted in children with SS Disease by using TCD. This measure is only looking at the process outcome of annual screening. In and of itself it does measure the clinical outcome, which is dependent on clinical actions based on the outcome of the screening.
- Measure 2797 is a process measure with HIGH level of clinical evidence. This rating is based largely on a 2014 systematic review by National Heart Lung and Blood Institute's (NHLBI) strong recommendation on the same topic and nearly identical criteria. Transcranial doppler screening of sickle cell patients is a step in stroke prevention and while screening does not guarantee the desired outcome, the clinical evidence clearly demonstrates that lack of annual screening is strongly associated with a poor outcome.
- Addressing a question posed to the committee: The difference between the NHLBI's recommendation (annual screening beginning at age 2 and continuing until at least age 16) and the measure's specification (annual screening for children ages 2 through 15 years old) most likely reflects the measure developer's attempt to translate the recommendation into an operational definition. Suggest forwarding this question to the measure developer.

1b. Performance Gap.

- Gaps identified, but not clear about disparities. Should collect this data if implemented.
- Based on the Medicaid data provided there are no gender or SES identified, however younger children were more likely to receive TCD screening than older children. As most children with sickle cell disease are eligible for Medicaid, there is likely limited data from commercial populations.
- Very little screening was done in the Medicald populations for which data was reported prior to 2005 . Performance improved annually for these groups, reaching 50% compliance by 2010. More recent data was not presented, but it is assumed that compliance is less than 100%.
- Based on the 2005-2010 Medicaid data submitted by the measure developer, there is high confidence that performance gaps currently persist.
- Addressing questions posed to the committee:
    - Re: more recent data. Data from the New York State Health Department could be used to assess if performance gaps persist.
    - Re: warrants a national performance measure. Yes - In this reviewer's opinion even though the target population is relatively small.
    - Re: evidence of disparities. This question can best be answered as more data is collected. Expect that the measure will be disparities sensitive.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The numerator for this measure is the number of children ages 2 through 15 years old with sickle cell anemia who received at least one TCD screening within the measurement year. The denominator is the number of children ages 2 through 15 years with sickle cell anemia within the measurement year.
- The CPT, ICD-9, and ICD-10 codes are included in the numerator and denominator details.
- The calculation algorithm is stated in S.18 and appears straightforward.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

- Reliability testing was conducted at the performance measure level, but not the individual element level. Per the algorithm, the rating may be HIGH or MODERATE.
- Testing was conducted using signal to noise analysis assessing the reliability to confidently distinguish the performance of one state's Medicaid program from that of another state.
- The reliability statistics ranged from range 0.96-0.99, indicating a high degree of reliability
- The developer also performed validity testing at the data element level for the identification of TCD and denominator criteria (case definition of three sickle cell claims), effectively meaning it also tested reliability at the data element level.
- The measure is both specified and tested at the health plan level.

Question for the Committee (as appropriate):

- *Do the results demonstrate sufficient reliability so that differences in performance can be identified? To distinguish between a HIGH vs. MODERATE rating, the Committee is asked to assess: Is there HIGH or MODERATE certainty or confidence that the performance measure scores are reliable?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The NHLBI guidline states that screening should occur annually from age 2 to at least age 16. This measure is specified for ages 2 to 15. Other than the age discrepancy, the specifications are linked with the evidence, which states that children with SCD should be screened annually; this measure assesses whether children with SCD were screened annually.

Question for the Committee:

- *Does the Committee wish to request clarification from the developer on the age discrepancy?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- Empirical validity testing was performed for both the critical data elements (eligible for MODERATE rating) and the performance measure score (ELIGIBLE FOR MODERATE OR HIGH RATING).
- For the critical element testing, administrative claim accuracy was assessed through comparison with medical charts (N=91, 2012 data).
    - Agreement between the claims and medical record (2005-2010) to identify TCD screening/no screening was 96.7% with kappa = 0.93, 95% confidence interval (CI): 0.86.
    - Using administrative claims to identify receipt of TCD screening resulted in a sensitivity of 94% (95% CI: 83%-99%), specificity of 100% (95% CI: 91%-100%), NPV of 93% (95% CI: 81%-99%), PPV of 93% (95% CI: 92%-100%) compared with the gold standard of medical records.
    - Inter-rater reliability of paper records was conducted by examining 10 charts; the two trained abstractors had 100% agreement for receipt of TCD screening from the medical records, resulting in a kappa of 1.00
- Empirical validity testing at the performance score level (rate of screening) was conducted using 2007-2009 MAX data (only national data set, which served as the gold standard) to the Medicaid data obtained directly from Michigan. Rates of TCD screening using each data source were calculated and compared using z-tests for two proportions; for these tests, the null hypothesis was that the rate in each year would be the same in both Michigan Medicaid data and MAX data. The correlation coefficient and squared correlation coefficient were calculated to identify the extent of the relationship between the two sources. Results indicated strong agreement.
    - 2007: z-score = -0.685, p-value =0.497; 2008: z-score = 0.223, p-value = 0.223; 2009: z-score = 1.079, p-value = 0.280
    - Pearson correlation coefficient = 0.98; squared coefficient = 0.96
- Face validity also was established by a panel of national experts and parent advocates, as well as measurement and state Medicaid experts. The developer provided information on one team as the developer and a second group that assessed face validity. The latter panel scored the face validity as very high, with an average of 8.5 out of 9. The panel concluded performance on the measure would both distinguish good from bad care and also improve the quality of care provided.

Question for the Committee (as appropriate):

- *Do you agree that the score from this measure as specified is an indicator of quality? Empirical validity testing at the performance measure score level (national MAX data vs. Medicaid data) means eligibility for a HIGH or MODERATE rating. Is there HIGH or MODERATE certainty or confidence that the performance measure scores are a valid indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- There are no exclusions for this measure.

Question for the Committee (as appropriate):

- *Should there be any exclusions for this measure?*

2b4. Risk adjustment:

- There is no risk adjustment or stratification for this measure.

Question for the Committee:

- Should there be any risk adjustment of stratification?

2b5. Meaningful difference:

e developer reports:

- 2005-2010 MAX data were used. Logistic regression was used to estimate the associations between each year and receipt of TCD screening (2005 reference). Trends in TCD screening were assessed over time using linear regression.
- The measure distinguishes performance across years and distinguishes changes over time within a state. The measure also distinguishes differences between states.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

- No information was provided by the developer.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- Need to make sure a stringent diagnosis specification for SCD is made.
- How reliable is the measure of TCD itself?

2a2. Reliability testing

- The data elements are clearly defined and the appropriate CPT, ICD-9, and ICD-10 codes are included. The calculation algorithm is logical and clear. This measure lends to consistent implementation.
- This is a straight forward measurement, and tests to date indicate that there is good reliability.
- This reviewer chooses to divide this question into two parts:
  - First is reliably assessing whether a transcranial doppler study was performed on an annual basis. Rated as high reliability. This is based on use of a matching billing code for the examination in question and the developer's data showing that these codes are used in nearly every case the desired doppler study was performed.
  - Second is reliably identifying children with sickle cell disease and not related conditions. Rated as moderate reliability. This is based on the difficulty of using diagnostic codes and the developer's data showing that such coding data is frequently incomplete.
- Addressing questions posed to the committee:
  - Re: Data elements defined: Yes
  - Re: Clear algorithm: Yes
  - Re: Can measure be consistently implemented: Yes
  - Re: Can performance differences be identified: Yes

2b1. Validity Specifications

- The age discrepancy between the measure and the clinical practice guideline should be clarified.
- Validity could be improved by assessing reasons for why screening was not performed in populations/settings that otherwise show high level of performance.

2b2. Validity Testing

- Seems pretty valid. Frequency of screen is not supported by data? What do we know about the natural history?
- Empirical validity testing was prefromed at both the critical data ements and the performance measure score. Face validity was also estabished by a panel of national experts, parent advocates and measurement and state Medicaid experts. There this measure is eligible for a high or moderate rating.
- Empirical and face validity strong.

2b3-2b7. Threats to Validity

- Not sure if there should be stratification. Are there significant influences on risk--are all ages the same risk?
- There are no exclusions, which seems appropriate for this measure.

- Using Algorithm #3, the measure is rated as insufficient since some potential threats to validity persist. For example, the available evidence suggests a valid link between the measure and the desired results only span a moderate window (11 to 50%). Suggest asking the developers whether any sites achieved a score of >90% and if not, why? One can readily imagine that providers and families might opt against annual transcranial doppler screening under certain circumstances such as prior strokes or factors that make transfusion therapy an undesirable intervention. This could be addressed by adding exclusions to the measure.
- Addressing questions posed to the committee:
  - Re: measure as an indicator of quality? The measure may only be valid within a certain range of scores (e.g., 0-50%), especially since 100% compliance may not be a realistic target.
  - Re: exclusions. Reviewing isolated instances where transcranial doppler is not being used would provide insight into potential need for exclusions. This could be addressed as measure enters widespread use.
  - Re: risk adjustment. No need is identified
  - Re: meaningful differences in quality. Again within certain ranges the measure is readily linked to quality of care, but the linkage is subject to threats as noted above.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are in defined fields in electronic claims.
- The measure was tested using Medicaid administrative claims data. At the state level, states can use their own data; at a multistate level, MAX data are available from CMS.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Seems like the data is available.
- Data elements are routinely used in administrative claims and routinely generated and used during clinical care.
- Should be fairly straightforward Health Plan measure as the required elements are routinely coded.
- Rated as high
- Addressing questions posed to the committee:
  - Re: Data elements generated and used during care delivery: Yes
  - Re: Data collection strategy ready for operational use. Yes

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- The measure is currently in use for surveillance purposes by the New York State Health Department.
- The developers have not indicated any specific plans for the measure's use in public reporting or value-based purchasing.
- The developers stated no unintended consequences were noted during testing.

Questions for the Committee:

- *Does New York State publicly report the surveillance results?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- I am not sure how important the frequency issue is? Does the risk change over time? If one has three negative evaluations, does the risk on a fourth measurement instance justify ongoing surveillance?
- Although the measure is currently being used for surveillance purposes in New York State Health Department, there is no information provided on public reporting.
- Health Plan measure. Only issue may be the size of the population, which may lead to difficulties in comparing performance.
- Rated as high
- Addressing questions posed to the committee:
  - Re: Does NY publicly report its results? Worthwhile asking the developer to help address this question.
  - Re: Potential unintended consequences? Consider minimal since transcranial doppler has minimal risk beyond cost, inconvenience, false negative and false positive results

Criterion 5: Related and Competing Measures

- There are no related or competing measures.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: **Transcranial Doppler Ultrasonography Screening Among Children with Sickle Cell Anemia**

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/25/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Transcranial Doppler ultrasonography screening among children with sickle cell anemia

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Transcranial Doppler (TCD) ultrasonography measures the blood flow velocity in cerebral arteries, specifically the distal internal carotid artery and the proximal middle cerebral artery. High blood velocities are indicative of an upcoming stroke and the need to begin stroke prevention efforts among children with sickle cell anemia. Stroke prevention efforts result in a substantial reduction in the incidence of stroke among children with sickle cell anemia.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

National Heart, Lung, and Blood Institute. Evidence Based Management of Sickle Cell Disease. 2014; http://www.nhlbi.nih.gov/health-pro/guidelines/sickle-cell-disease-guidelines/sickle-cell-disease-report.pdf.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Pages 20-21:

Recommendations
1. In children with SCA, screen annually with TCD according to methods employed in the STOP studies, beginning at age 2 and continuing until at least age 16. (***Strong Recommendation, Moderate-Quality Evidence***)
2. In children with conditional (170–199 cm/sec) or elevated (>200 cm/sec) TCD results, refer to a specialist with expertise in chronic transfusion therapy aimed at preventing stroke. (***Strong Recommendation, High-Quality Evidence***)
3. In children with genotypes other than SCA (e.g., HbSβ⁺-thalassemia or HbSC), do not perform screening with TCD. (***Strong Recommendation, Low-Quality Evidence***)
4. In asymptomatic children with SCD, do not perform screening with MRI or CT. (***Moderate Recommendation, Low-Quality Evidence***)
5. In asymptomatic adults with SCD, do not perform screening with neuroimaging (TCD, MRI, or CT). (***Moderate Recommendation, Very Low-Quality Evidence***)

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Strong Recommendation, Moderate-Quality Evidence

Grade of Recommendation	Clarity of Risk/ Benefit	Quality of Supporting Evidence	Implications
Strong recommendation Moderate-quality evidence	Benefits clearly outweigh harms and burdens, or vice versa	Evidence from RCTs with important limitations (inconsistent results, methodological flaws, indirect or imprecise evidence), or unusually strong evidence from unbiased observational studies	Recommendation can apply to most patients in most circumstances. Further research (if performed) is likely to have an impact on our confidence in the estimate of effect and may change the estimate.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Exhibit 4. GRADE Recommendations–A Closer Look

Grade of Recommendation	Clarity of Risk/ Benefit	Quality of Supporting Evidence	Implications
Strong recommendation High-quality evidence	Benefits clearly outweigh harms and burdens, or vice versa	Consistent evidence from well-performed RCTs or exceptionally strong evidence from unbiased observational studies*	Recommendation can apply to most patients in most circumstances. Further research is very unlikely to change our confidence in the estimate of effect.
Strong recommendation Moderate-quality evidence	Benefits clearly outweigh harms and burdens, or vice versa	Evidence from RCTs with important limitations (inconsistent results, methodological flaws, indirect or imprecise evidence), or unusually strong evidence from unbiased observational studies	Recommendation can apply to most patients in most circumstances. Further research (if performed) is likely to have an impact on our confidence in the estimate of effect and may change the estimate.
Strong recommendation Low-quality evidence	Benefits clearly outweigh harms and burdens, or vice versa	Evidence for at least one critical outcome from observational studies, from RCTs with serious flaws, or indirect evidence	Recommendation may change when higher quality evidence becomes available. Further research (if performed) is likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Strong recommendation Very low-quality evidence (very rarely applicable)	Benefits clearly outweigh harms and burdens, or vice versa	Evidence for at least one of the critical outcomes from unsystematic clinical observations or very indirect evidence	Recommendation may change when higher quality evidence becomes available; any estimate of effect, for at least one critical outcome, is very uncertain.
Weak recommendation High-quality evidence	Benefits closely balanced with harms and burdens	Consistent evidence from well-performed RCTs or exceptionally strong evidence from unbiased observational studies	The best action may differ depending on circumstances or patient or societal values. Further research is very unlikely to change our confidence in the estimate of effect.

Grade of Recommendation	Clarity of Risk/ Benefit	Quality of Supporting Evidence	Implications
Weak recommendation Moderate-quality evidence	Benefits closely balanced with harms and burdens	Evidence from RCTs with important limitations (inconsistent results, methodological flaws, indirect or imprecise evidence), or unusually strong evidence from unbiased observational studies	Alternative approaches likely to be better for some patients under some circumstances. Further research (if performed) is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Weak recommendation Low-quality evidence	Uncertainty in the estimates of benefits, harms, and burdens; benefits may be closely balanced with harms and burdens	Evidence for at least one critical outcome from observational studies, from RCTs with serious flaws, or indirect evidence	Other alternatives may be equally reasonable. Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Weak recommendation Very low-quality evidence	Major uncertainty in the estimates of benefits, harms, and burdens; benefits may or may not be balanced with harms and burdens	Evidence for at least one critical outcome from unsystematic clinical observations or very indirect evidence	Other alternatives may be equally reasonable. Any estimate of effect, for at least one critical outcome, is very uncertain.

Source: Reprinted with permission of the American Thoracic Society. Copyright © 2012 American Thoracic Society. Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, Fahy BF, Gould MK, Horan KL, Krishnan JA, Manthous CA, Maurer JR, McNicholas WT, Oxman AD, Rubenfeld G, Turino GM, Guyatt G; ATS Documents Development and Implementation Committee. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med*. 2006 Sep 1;174(5):605-14. Official Journal of the American Thoracic Society.[29]

* Exceptionally strong evidence from unbiased observational studies includes: (1) evidence from studies that yield estimates of the treatment effect that are large and consistent; (2) evidence in which all potential biases may be working to underestimate an apparent treatment effect, and therefore, the actual treatment effect is likely to be larger than that suggested by the study data; and (3) evidence in which a dose-response gradient exists

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The body of evidence summarized in these responses is from the National Heart, Lung, and Blood Institute Evidence-based Clinical Guidelines for the Management of Sickle Cell Disease (evidence tables: Table 9. Transcranial Doppler Results): https://www.nhlbi.nih.gov/sites/www.nhlbi.nih.gov/files/scd_screening.pdf

The specific service addressed in this evidence review was TCD screening.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

See 1a.4.3

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See 1a.4.4

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1991 to 2011

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

A total of 2 randomized control trials (RCTs) and 50 observational studies are included in the body of evidence (8 retrospective observational studies, 23 prospective observational studies, 18 cross-sectional studies, and 1 case series).

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

As stated within the NHLBI Clinical Guidelines: "Two RCTs and 50 observational studies on the use of TCD were included. The two RCTs evaluated the efficacy of early intervention and demonstrated that screening coupled with prophylactic transfusion can markedly reduce the risk of stroke in children with SCA whose cerebral blood flow velocity measurements are considered at high risk. The fifty observational studies enrolled more than 11,000 patients and assessed the use of TCD as a screening test in children with SCD. The quality of evidence supporting screening with TCD was considered moderate to high."

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Receipt of TCD screening does not directly impact the risk of stroke among children with sickle cell anemia; however, the indication of high-risk of stroke obtained from TCD screening (blood flow velocity>200cm/sec) prompts the initiation of primary stroke prevention efforts in the form of blood transfusions. For brevity, we have included estimates of benefit and consistency among studies within the body of evidence directly related to

the process of TCD screening and the health-related outcome of primary stroke prevention among children with sickle cell anemia.

The majority of the studies used a standard definition of an abnormal TCD screening result (blood flow velocity>200cm/sec). A handful of studies used a looser definition, classifying velocities of over 170cm/sec as abnormal; however, these children would have been included in the definition of conditional TCD screening result in the other studies. Studies reported between 2% and 33% abnormal TCD screening results within their study populations; this large range may be attributable to differing study population inclusion criteria

All studies investigating the relationship between blood flow velocity as detected by TCD screening and stroke risk show that children with high blood flow velocities in the cerebral vessels are at a significantly increased risk of stroke. Adams (1992) reported in a prospective observational study that among 7 children who had a stroke within the study period (overall n=190), 6 children had an abnormal TCD screening result (Fisher's exact p-value<0.00001). Adams (2004) also reported that among 2,342 children with SCD who received a TCD screen, risk of stroke with abnormal TCD was much higher than with normal results (p-value<.001), conditional findings (p-value<.001), or inadequate TCD results (p-value=.002).

All studies that assessed stroke rates pre- and post-TCD screening recommendations found a significantly decreased rate of first stroke among children with sickle cell anemia post-TCD recommendations when compared with the pre-TCD recommendation time period. Armstrong-Wells (2008) reported a stroke rate of 0.44 per 100 pre-TCD recommendations and a stroke rate of 0.19 per 100 person-years post-TCD recommendations; Enningful-Eghan (2010) reported a stroke rate of 0.67 per 100 person-years pre-TCD recommendations and a post-TCD stroke rate of 0.06 per 100 person-years (p-value<0.0001). In addition, McCarville (2008) showed significantly decreasing stroke rates with increasing TCD use (p-value=0.045).

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

No harm is expected through the receipt of TCD screening; therefore, there is no negative affect of TCD screening on the net benefit.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Q-METRIC_SCD_TCD_EvidenceAttachment.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Children with sickle cell anemia (HbSS) have over three hundred times the stroke risk than children with normal hemoglobin (Verduzco and Nathan, 2009). Without intervention, approximately 11% of children with sickle cell anemia will have a stroke by age 20 (Verduzco and Nathan, 2009; Ohene-Frempong et al., 1998). Transcranial Doppler (TCD) ultrasonography measures the blood velocities within the cerebral vessels (Adams et al., 1997; Adams et al., 1992). Children over the age of 2 with a time-average mean maximum blood flow velocity of 200cm/sec or greater as measured by TCD ultrasonography have been shown to have 27 times the risk of stroke than children with velocities less than 200cm/sec. This corresponds to a 40% risk of stroke among those with high velocities within 3 years (Adams et al., 1997). Initiation of chronic blood transfusions reduces the risk of stroke by 92% among children at highest risk of stroke as identified through TCD screening (Adams et al., 1997; Adams et al., 1992). TCD screening is a reasonable method to assess stroke risk among children with sickle cell anemia, as it is safe, non-invasive and low cost (Markus, 2000). Although other predictors of stroke have been examined, such as hematocrit levels and white blood cell count, TCD velocities have been shown to be the only independent predictor of stroke (Adams et al., 1992). Given the importance of TCD screening to stroke prevention among children with sickle cell anemia, the National Heart, Lung, and Blood Institute (NHLBI) recommends each child with sickle cell anemia receive one TCD screen per year from ages 2 to 16 years (National Heart, Lung, and Blood Institute, 2014). Although the benefits of TCD screening among children with sickle cell anemia have been known since the late nineties, prior studies indicate that TCD screening rates are low. However, these reports are limited in their generalizability, as they are often focused on a single healthcare provider or registry. This measure establishes a claims-based method for identifying receipt of TCD screening among larger and broader populations of children with sickle cell anemia. The measure specifications are reflective of the guidelines from the NHLBI, and the performance scores calculated through this measure will identify areas in need of improvement in receipt of TCD screening among children with sickle cell anemia.

Citations:

Adams RJ, McKie VC, Carl EM, et al. Long-term stroke risk in children with sickle cell disease screened with transcranial Doppler. Ann Neurol. Nov 1997;42(5):699-704.

Adams R, McKie V, Nichols F, et al. The use of transcranial ultrasonography to predict stroke in sickle cell disease. N Engl J Med. Feb 27 1992;326(9):605-610.

Markus HS. Transcranial Doppler ultrasound. Br Med Bull. 2000;56(2):378-388.

National Heart Lung and Blood Institute. Evidence Based Management of Sickle Cell Disease. 2014; http://www.nhlbi.nih.gov/health-pro/guidelines/sickle-cell-disease-guidelines/sickle-cell-disease-report.pdf. Accessed 11/11, 2014.

Ohene-Frempong K, Weiner SJ, Sleeper LA, et al. Cerebrovascular accidents in sickle cell disease: rates and risk factors. Blood. Jan 1 1998;91(1):288-294.

Verduzco LA, Nathan DG. Sickle cell disease and stroke. Blood. Dec 10 2009;114(25):5117-5125.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is*

required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
Rates of transcranial Doppler screening among children with sickle cell anemia enrolled in Medicaid, by state, 2005-2010

Florida
(Year: Numerator/Denominator = Rate)
2005: 113/526 = 21.5%
2006: 121/489 = 24.7%
2007: 133/449 = 29.6%
2008: 171/502 = 34.1%
2009: 264/697 = 37.9%
2010: 339/734 = 46.2%

Illinois
(Year: Numerator/Denominator = Rate)
2005: 65/250 = 26.0%
2006: 85/276 = 30.8%
2007: 70/278 = 25.2%
2008: 78/291 = 26.8%
2009: 90/338 = 26.6%
2010: 86/302 = 28.5%

Louisiana
(Year: Numerator/Denominator = Rate)
2005: 150/364 = 41.2%
2006: 141/321 = 43.9%
2007: 164/322 = 50.9%
2008: 167/334 = 50.0%
2009: 164/356 = 46.1%
2010: 168/361 = 46.5%

Michigan
(Year: Numerator/Denominator = Rate)
2005: 27/240 = 11.3%
2006: 35/219 = 16.0%
2007: 26/243 = 10.7%
2008: 49/228 = 21.5%
2009: 93/259 = 35.9%
2010: 104/240 = 43.3%

South Carolina
(Year: Numerator/Denominator = Rate)
2005: 41/214 = 19.2%
2006: 37/189 = 19.6%
2007: 41/173 = 23.7%
2008: 48/124 = 38.7%
2009: 38/102 = 37.3%
2010: 68/134 = 50.7%

Texas
(Year: Numerator/Denominator = Rate)
2005: 18/258 = 7.0%
2006: 15/292 = 5.1%
2007: 56/343 = 16.3%
2008: 89/352 = 25.3%
2009: 140/372 = 37.6%

2010: 146/370 = 39.5%

Total
(Year: Numerator/Denominator = Rate)
2005: 414/1852 = 22.4%
2006: 434/1786 = 24.3%
2007: 301/1326 = 22.7%
2008: 313/1297 = 24.1%
2009: 357/1359 = 26.3%
2010: 431/1329 = 32.4%

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
There are no gender disparities in TCD screening among children with sickle cell anemia (chi-square=1.2, p-value=0.28). The data used for performance scores is state Medicaid programs; therefore, there are no disparities identified by insurance or socioeconomic status. Younger children (ages 2-6) were more likely to receive TCD screening than older children (chi-square=99.01, p-value<0.0001). For those 2 to 6 years old, 36% received a TCD screen; for those ages 7 to 11 years, 31% received a TCD screen; and for those ages 12-15 years, 25% were screened.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Per NQF Review: Not currently an evaluation criterion.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
See Citations in 1b.1.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cardiovascular : Screening, Neurology : Stroke/Transient Ischemic Attack (TIA), Prevention : Screening

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Disparities

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://chear.org/sites/default/files/stories/pdfs/scd13_speconly.pdf

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Q-METRIC_SCD_Code_Table_ICD9_ICD10.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator is the number of children ages 2 through 15 years old with sickle cell anemia who received at least one TCD screening within the measurement year.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Each measurement year extends from January 1 to December 31 (12 months).

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Cases from target population with target process (Receipt of TCD screening): Receipt of TCD screening is identified as the presence of at least one CPT code for any of five acceptable ultrasonography tests within the measurement year among children in the target population. Acceptable CPT codes are: 93886 (complete study), 93888 (limited study), 93890 (vasoreactivity study), 93892 (emboli detection without intravenous microbubble injection), and 93893 (emboli detection with intravenous microbubble injection).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*

The denominator is the number of children ages 2 through 15 years with sickle cell anemia within the measurement year.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Children with sickle cell anemia are identified through the presence of at least three separate healthcare encounters related to sickle cell anemia (defined as hemoglobin [Hb]SS) within the measurement year. Sickle cell anemia-related healthcare encounters are identified through ICD codes. The ICD-9-CM codes to identify HbSS-related healthcare encounters are as follows: 282.61 (Hb-SS disease w/o crisis) and 282.62 (Hb-SS disease with crisis). The ICD-10-CM codes for HbSS-related healthcare encounters are as follows: D57.00 (Hb-SS disease with crisis, unspecified); D57.01 (Hb-SS disease with acute chest syndrome); and D57.02 (Hb-SS disease with splenic sequestration). Children ages 2 through 15 years are included within the target population (i.e., must not have a 2nd or 16th birthday within the measurement year).

It is important to note that accurate calculation of this measure requires that the target population be selected from among children who have all of their health services for the measurement year included in the administrative claims data set. For children who have dual enrollment in other health plans, their claims may not be complete since some of their health services may have been paid for by another health plan. Inclusion of children with other health insurance would potentially cause this measure to be understated. As a consequence, this measure requires that children must not only be continuously enrolled within the health plan from which claims are available, the enrollment files must also be assessed to determine whether other forms of health insurance existed during the measurement year. Children with evidence of other insurance during the measurement year (i.e., coordination of benefits) are excluded from the target population.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
There are no denominator exclusions.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion

If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify the denominator: Determine the eligible population using administrative claims. The eligible population is all individuals who satisfy all specified criteria, including age, continuous enrollment, and diagnosis requirements within the measurement year.

2. Identify the numerator: Identify numerator events using administrative claims for all individuals in the eligible population (denominator) within the measurement year.

3. Calculate the rate (numerator / denominator).

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure does not involve sampling; all sickle cell anemia cases meeting the inclusion criteria are included.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
N/A

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Other
If other: Any setting represented with claims data

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

N/A
2a. Reliability – See attached Measure Testing Submission Form **2b. Validity – See attached Measure Testing Submission Form** Q-METRIC_SCD_TCD_NQF_TestingAttachment-635799133636176811.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: **Transcranial Doppler Ultrasonography Screening Among Children with Sickle Cell Anemia**
Date of Submission: 9/25/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☒ other: Michigan Newborn Screening

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Existing Datasets:

- Michigan Medicaid administrative claims data provided by the Michigan Department of Health and Human Services (MDHHS)
- Medicaid Analytic eXtract (MAX) administrative claims data for 6 state Medicaid programs provided by the Centers for Medicare & Medicaid Services (CMS)

Other data used for testing (not existing datasets):

- Medical record data from Children's Hospital of Michigan (CHM), Detroit, Michigan; Hurley Medical Center (HMC), Flint, Michigan; and University of Michigan Health Services (UMHS), Ann Arbor, Michigan
- Michigan Newborn Screening (NBS) Results

1.3. What are the dates of the data used in testing?
Michigan Medicaid data 2007-2011; MAX data: 2005-2010; CHM, HMC, and UMHS medical record data: 2012; Michigan NBS: 1987-2010

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of:	**Measure Tested at Level of:**

(*must be consistent with levels entered in item S.26*)	
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other:	☐ other:

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

- The Michigan Medicaid data consisted of all Medicaid claims for Medicaid enrollees within the state of Michigan (2007-2011);
- The MAX data consisted of all Medicaid claims reported to CMS for Medicaid enrollees within 6 state Medicaid programs with moderate to high prevalence of sickle cell anemia: Florida, Illinois, Louisiana, Michigan, South Carolina, and Texas (2005-2010);
- The medical record data were obtained from three hospitals: CHM, HMC, and UMHS (2012). These three large medical centers are located in urban areas in Michigan which are reflective of the residence of the vast majority of children with sickle cell anemia living in Michigan:
    - CHM is a tertiary medical center located in Detroit, Michigan;
    - HMC is a tertiary medical center located in Flint, Michigan; and
    - UMHS is an academic medical center located in Ann Arbor, Michigan;
- The Michigan NBS data consisted of all births within the state of Michigan (1987-2010).

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

- The Michigan Medicaid data from 2007 to 2009 was a complete census of all children ages 2-16 years with sickle cell anemia that met eligibility criteria within each year (Table 1). The population was equally divided between sexes; approximately 98% were black.

Table 1: Number of children ages 2 to 16 years with sickle cell anemia enrolled in Michigan Medicaid, 2007-2009

2007	2008	2009
359	334	359

- The Michigan Medicaid data from 2010 and 2011 provided a complete census of all children ages 1-18 years with at least one sickle cell disease (SCD)-related administrative claim, continuously enrolled annually within Michigan Medicaid in 2010 and/or 2011, with a newborn screening result available. This included 938 children in 2010 and 924 children in 2011. The population was equally divided between sexes; approximately 75% were black and the average age was approximately 10 years.

- The MAX data included all children ages 2-16 years with sickle cell anemia that met eligibility criteria within each year for Medicaid claims reported by selected states (Table 2). The population was equally divided between sexes; approximately 98% were black.

Table 2: Number of children enrolled in Medicaid with sickle cell anemia, MAX data by state, 2005-2010

State	2005	2006	2007	2008	2009	2010
Florida	526	489	449	502	697	734
Illinois	250	276	278	291	338	302
Louisiana	364	321	322	334	356	361
Michigan	240	219	243	228	259	240
South Carolina	214	189	173	124	102	134
Texas	258	292	343	352	370	370

- A sample of abstracted medical records from 91 children with sickle cell anemia ages 2-16 years who were enrolled in Michigan Medicaid was drawn at three sickle cell centers in Michigan (CHM, HMC, UMHS) for children meeting the transcranial Doppler (TCD) screening measure specification criteria during 2012.
- The Michigan NBS data included all children born in the state of Michigan from 1987-2010 with a positive and confirmed screening result that had at least 1 SCD-related claim and were continuously enrolled in Michigan Medicaid in either 2010 or 2011.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

- Reliability testing data: MAX

- Validity testing data: Michigan Medicaid, MAX, Michigan NBS, and medical records

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The data do not include patient-level sociodemographic (SDS) variables; however, all children included in the data were enrolled in Medicaid.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The reliability of MAX data to evaluate TCD screening is of high importance since this is the only national source of state Medicaid data available upon which state-to-state comparisons may be conducted. The reliability of this measure was calculated using a signal-to-noise analysis. The signal-to-noise analysis was focused on assessing the reliability to confidently distinguish the performance of one state's Medicaid program from that of another state. For this approach, reliability was estimated with a beta-binomial model (RAND Corporation, TR-653-NCQA, 2009).

See section **2b2** for validity testing of data elements.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

State-specific reliability results for receipt of TCD screening among children with sickle cell anemia are detailed in Table 3. These results show that the reliability based on signal-to-noise analysis ranged from 0.96 to 0.99, with a median of 0.98.

Table 3. State-specific reliability for measure
Between State Variance: 0.0056

State	Numerator	Denominator	Reliability Statistic
Florida	1141	3397	0.99
Illinois	474	1735	0.98
Louisiana	954	2058	0.98
Michigan	334	1429	0.98
South Carolina	273	936	0.96
Texas	464	1985	0.98
Median (range)			0.98 (0.96-0.99)

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

State-specific reliability is very good; observed reliability was consistently greater than 0.95. In general, reliability scores can range from 0.0 (all variation is attributable to measurement error) to 1.0 (all variation is caused by real differences). While there is not a clear cut-off for minimum reliability level, values above 0.7 are considered sufficient to distinguish differences between some states and the mean; reliability values above 0.9 are considered sufficient to see differences between states (RAND Corporation, TR-653-NCQA, 2009). The median reliability observed across states was 0.98 (range: 0.96-0.99), which is consistent with a high degree of reliability.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Critical Data Elements

Numerator: The accuracy of administrative claims in identifying receipt of TCD screening was assessed through comparison to the gold standard of medical charts. An audit was conducted by trained medical record abstractors to compare administrative claims data with corresponding medical records data. Medical records were abstracted for all children meeting the TCD screening measure specification criteria; agreement between the medical records and the administrative claims was assessed using kappa. Furthermore, the sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) of administrative claims for receipt of TCD screening were calculated; the medical charts were the gold standard for comparison. In addition, the reliability of the data element abstracted from the medical chart was assessed by identifying a subset of the charts to be re-abstracted by another trained medical record abstractor; the results of the two abstractors were compared using percent agreement and kappa.

Denominator: The accuracy of the case definition (at least 3 claims for sickle cell anemia (HbSS) within the measurement year) to identify children with sickle cell anemia was assessed through comparison to the gold standard of newborn screening results for the state of Michigan for children enrolled in Michigan Medicaid in 2010 and 2011 with at least one SCD-related healthcare claim within their enrollment year(s). The area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, PPV, and NPV were calculated for the case definition. As a comparison, these values were also calculated for those with a minimum of at least 1 or 2 HbSS claims within each year.

Conversion of ICD-9 to ICD-10

The goal of ICD-9 to ICD-10 conversion was to convert this measure to a new code set, fully consistent with the intent of the original measure. All ICD-9 diagnosis codes were converted to the corresponding ICD-10 codes using the CMS 2015 diagnosis code General Equivalence Mappings (GEMs) and diagnosis code description files (accessed on August 26, 2015); these mapping files were created by CMS. The target ICD-9 codes were converted to ICD-10 using the GEM file and manually reviewed for consistency using the diagnosis code descriptions for the source ICD-9 and converted ICD-10 codes. In addition, the resultant ICD-10 codes were back-translated to ICD-9 to verify the accuracy of the coding. Source files from CMS were acquired from these files:

1. ICD 9 to 10 diagnosis GEM -*2015_I9gem.txt*
   https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html
2. ICD 10 to 9 diagnosis GEM - 2015_10gem.txt
   https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html
3. ICD 9 description file *CMS32_DESC_SHORT_DX.txt*
   https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html
4. ICD 10 description file - *icd10cm_order_2015.txt*
   https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html

The ICD-9 code 282.61 (Hb-SS disease without crisis) mapped to the ICD-10 code of D57.1 (sickle-cell disease without crisis). This ICD-10 code was not included in the measure specification, as it is not specific to sickle cell anemia (HbSS). The ICD-9 code 282.62 (Hb-SS disease with crisis) mapped to ICD-10 D57.00 (Hb-SS disease with crisis, unspecified) and was included in the specification. Subsequent verification using the GEMs indicated that ICD-10 codes D57.01 (Hb-SS disease with acute chest syndrome) and D57.02 (Hb-SS disease with splenic sequestration) were also appropriate to include in the measure specification to identify the study population (denominator).

Empirical Validity Testing of Performance Measure

Although a state would typically have direct access to its own Medicaid data, it is unlikely that a state would have similar access to other states' data for comparison. However, CMS develops and maintains standardized Medicaid Analytic eXtract (MAX) data for public use using administrative claims submitted by each state Medicaid program. The MAX data are the only national, person-level administrative claims dataset available for the Medicaid program. As a consequence, MAX data, rather than data acquired directly from individual Medicaid programs, are likely to be used to perform cross-state comparisons of TCD screening among children with sickle cell anemia. Since states submit their Medicaid data to CMS for conversion into the MAX datasets, a state's own Medicaid data can be considered the authoritative source for administrative claims.

Our empirical validity testing of this performance measure compared the MAX data for the state of Michigan (obtained from CMS) to the gold standard of Michigan Medicaid data (obtained directly from Michigan's claims data warehouse) for the same time period (2007-2009). Note that the testing time period was constraint to align with the most recent MAX data available from CMS at the time of this analysis. Rates of TCD screening using each source of data were calculated and compared using z-tests for two proportions; for these tests, the null hypothesis was that the rate in each year would be the same in both Michigan Medicaid data and MAX data. Additionally, the correlation coefficient and squared correlation coefficient were calculated to identify the extent of the linear relationship between the two data sources.

Face Validity of Performance Measure Score

The face validity of this measure was established by a panel of national experts and advocates for families of children with SCD convened by the Quality Measurement, Evaluation, Testing, Review, and Implementation Consortium (Q-METRIC). The Q-METRIC expert panel included nationally recognized experts in SCD, representing hematology, pediatrics, and SCD family advocacy. In addition, measure validity was considered by experts in state Medicaid program operations, health plan quality measurement, health informatics, and health care quality measurement. In total, the Q-METRIC SCD panel included 14 experts, providing a comprehensive perspective on SCD management and the measurement of quality metrics for states and health plans. The expert panel assessed whether the performance of the measure would result in improved quality of care for children with sickle cell disease. Specifically in respect to TCD screening, the panel weighed evidence to determine if the performance of TCD as outlined in the measure would improve the quality of care provided to patients. The voting process to prioritize the measure was based on the ability of the measure to distinguish good from poor quality.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Critical Data Elements

Numerator: For this comparison, 91 children with sickle cell anemia who were enrolled within Michigan Medicaid were successfully matched with their Michigan Medicaid administrative claims data. Among these children, TCD screening was identified in both the administrative claims data and the medical record review for 47 (51.6%) cases (Table 4). Similarly, 41 (45.1%) cases were classified as not having a TCD in both data sources, yielding an overall agreement of 96.7% (kappa = 0.93, 95% confidence interval (CI): 0.86, 1). Using administrative claims to identify receipt of TCD screening resulted in a sensitivity of 94% (95% CI: 83%-99%),

a specificity of 100% (95% CI: 91%-100%), a NPV of 93% (95% CI: 81%-99%), and a PPV of 93% (95% CI: 92%-100%) compared with the gold standard of medical records. Ten charts were also chosen for exploration of inter-rater reliability; the two trained abstractors had 100% agreement with each other for abstracting receipt of TCD screening from the medical records, resulting in a kappa of 1.00.

Table 4: Michigan validation testing (administrative claims vs. medical records) for transcranial Doppler screening among children with sickle cell anemia

Transcranial Doppler Screening in Medicaid Claims Data	Transcranial Doppler Screening in Medical Record (n=91)		
	Yes	No	Total
Yes	51.6% (47)	0	51.6% (47)
No	3.3% (3)	45.1% (41)	48.4% (44)
Total	54.9% (50)	45.1% (41)	100% (91)

Denominator: For this comparison, 865 children met eligibility criteria in 2010 (at least 1 SCD-related claim ages 1-18, continuous enrollment in Michigan Medicaid in 2010, a newborn screening result available); 836 children met eligibility criteria in 2011. In 2010, a case definition of 3 HbSS claims within the year was 91.4% sensitive and 80% specific in identifying children with sickle cell anemia (HbSS) (PPV: 80.4%; NPV: 91.3%). These results were replicated with nearly identical precision among the study population in 2011 (Table 5). In comparison, using a case definition of at least 1 HbSS claim or at least 2 HbSS claims to identify the study population resulted in substantially less specificity.

Table 5. Accuracy of case definition of at least 1, 2 and 3 HbSS claims within a year to identify children with sickle cell anemia as compared to the gold standard of newborn screening

Algorithm	Area under the ROC Curve	# True Positives	# False Positives	# True Negatives	# False Negatives	Sensitivity	Specificity	PPV	NPV
Results - 2010									
≥1 HbSS Claim	0.50	409	456	0	0	100.0%	0.0%	47.3%	NA
≥2 HbSS Claims	0.82	391	144	312	18	95.6%	68.4%	73.1%	94.5%
≥3 HbSS Claims	0.86	374	91	365	35	91.4%	80.0%	80.4%	91.3%
Results - 2011									
≥1 HbSS Claim	0.50	397	439	0	0	100.0%	0.0%	47.5%	NA
≥2 HbSS Claims	0.79	377	163	276	20	95.0%	62.9%	69.8%	93.2%
≥3 HbSS Claims	0.87	363	97	342	34	91.4%	77.9%	78.9%	91.0%

Empirical Validity Testing of Performance Measure

The comparison of rates of TCD screening from the gold standard of Michigan Medicaid data as compared to MAX data can be seen in Table 6. This illustrates that the number of TCD cases among children with sickle cell anemia ranged from 45 to 114 screenings in the claims acquired directly from the Medicaid data warehouse, versus a range of 26 to 93 screenings from MAX data for the same time period.

Table 6: Comparison of transcranial Doppler screening by source of Medicaid claims data for the state of Michigan, 2007-2009

Source	Rate	2007	2008	2009

	Components			
MAX data	Numerator	26	49	93
	Denominator	243	228	259
	Percentage	10.7%	21.5%	35.9%
Michigan Medicaid data	Numerator	45	58	114
	Denominator	359	334	359
	Percentage	12.5%	17.4%	31.8%

Figure 1 illustrates the TCD screening rates observed between the Michigan Medicaid data from the state warehouse and MAX data from CMS for each overlapping year noted, respectively: 12.5% versus 10.7% (2007); 17.4% versus 21.5% (2008); and 31.8% versus 35.9% (2009).

Figure 1: Comparison of transcranial Doppler screening by source of Medicaid claims data, Michigan

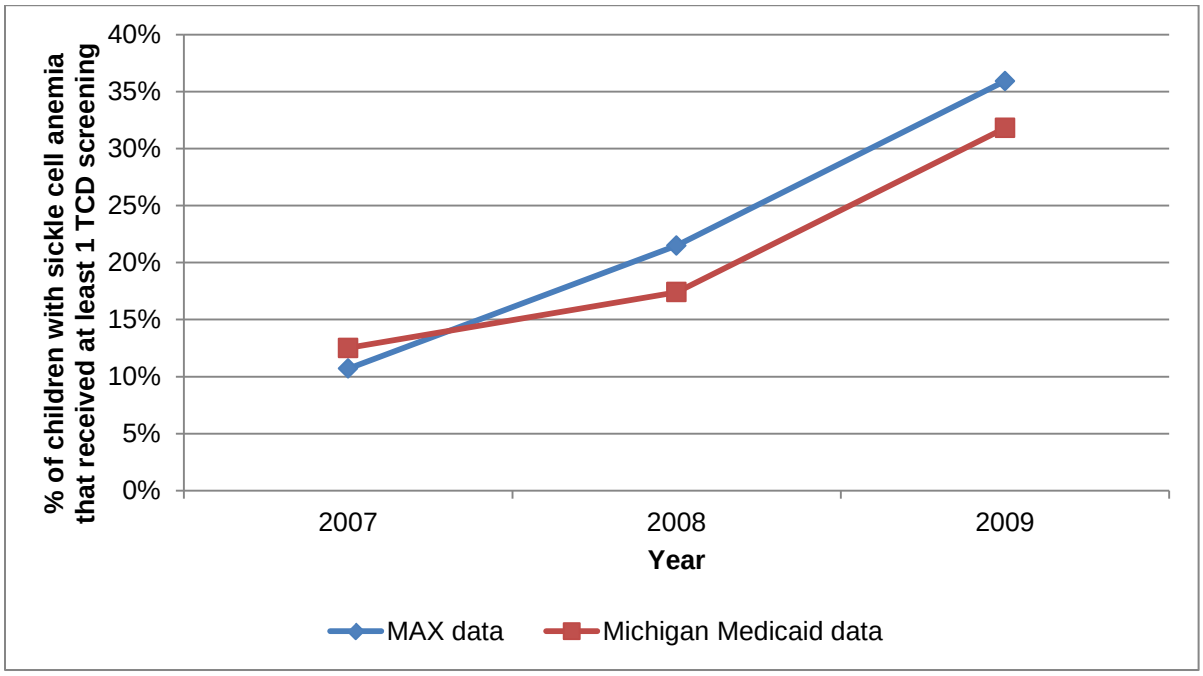


Table 7 reports the z-scores and p-values from the two-sample z-tests comparing the proportion of children that received screening each year between Michigan Medicaid and MAX data.

Table 7: Comparison of transcranial Doppler screening by source of Medicaid claims data, Michigan

	2007	2008	2009
z-score	-0.685	1.223	1.079
p-value	0.4965	0.2225	0.2801

Additionally, the data comparison revealed a Pearson correlation coefficient of 0.98, corresponding to a squared correlation coefficient of 0.96.

Face Validity of Performance Measure Score

The Q-METRIC expert panel concluded that this measure has a very high degree of face validity through a detailed review of concepts and metrics considered to be essential to effective SCD management and treatment. Concepts and draft measures were rated by this group for their relative importance. This measure was among the most highly rated, receiving an average score of 8.5 (with 9 as the highest possible score). In

addition, the expert panel concluded that the performance of TCD as outlined in this measure would improve the quality of care provided to patients, and the measure would be able to distinguish good from poor quality.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Critical Data Elements

Numerator: A kappa of greater than .81 is considered almost perfect agreement (Landis and Koch, 1997). In addition, the sensitivity, specificity, NPV and PPV are high. Given this evidence, we believe the validity of administrative claims in assessing receipt of TCD screening is very high.

Denominator: A sensitivity of over 90% and a specificity of approximately 80%, as well as the reliability across years, allow us to conclude that the denominator is valid for accurately identifying children with sickle cell anemia within administrative claims. These results indicate that the case definition used has a very high ability to correctly identify true cases and a somewhat lower ability to distinguish false positives. However, other less stringent case definitions resulted in substantially more misclassification than the chosen definition of at least 3 HbSS claims within the measurement year.

Empirical Validity Testing of Performance Measure

Our results suggest that, compared with the gold standard of Michigan Medicaid data, MAX data has a very high degree of validity. When TCD screening was assessed for the same state (Michigan) from these two data sources for the same time period (2007-2009), no differences in rates were observed (all p-values >0.20). Additionally, the high values of the correlation coefficient and the squared correlation coefficient indicate a high level of reliability. Correlation coefficients of greater than 0.70 indicate a strong positive linear relationship; therefore, our results suggest that compared with Michigan Medicaid data, MAX data is highly valid. The squared correlation coefficient value of 0.96 indicates that nearly 96% of the variability in the MAX data from CMS for the state of Michigan can be explained by variation in the data received directly from the Michigan Medicaid program. This finding indicates that the strength of the relationship between the two data sources is extremely strong.

Face Validity of Performance Measure Score

Given the high rating of the Q-METRIC expert panel, we feel this measure has a very high degree of face validity.

2b3. EXCLUSIONS ANALYSIS

☒ **no exclusions — *skip to section 2b4***

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the*

effect on the performance score is transparent, e.g., scores with and without exclusion)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

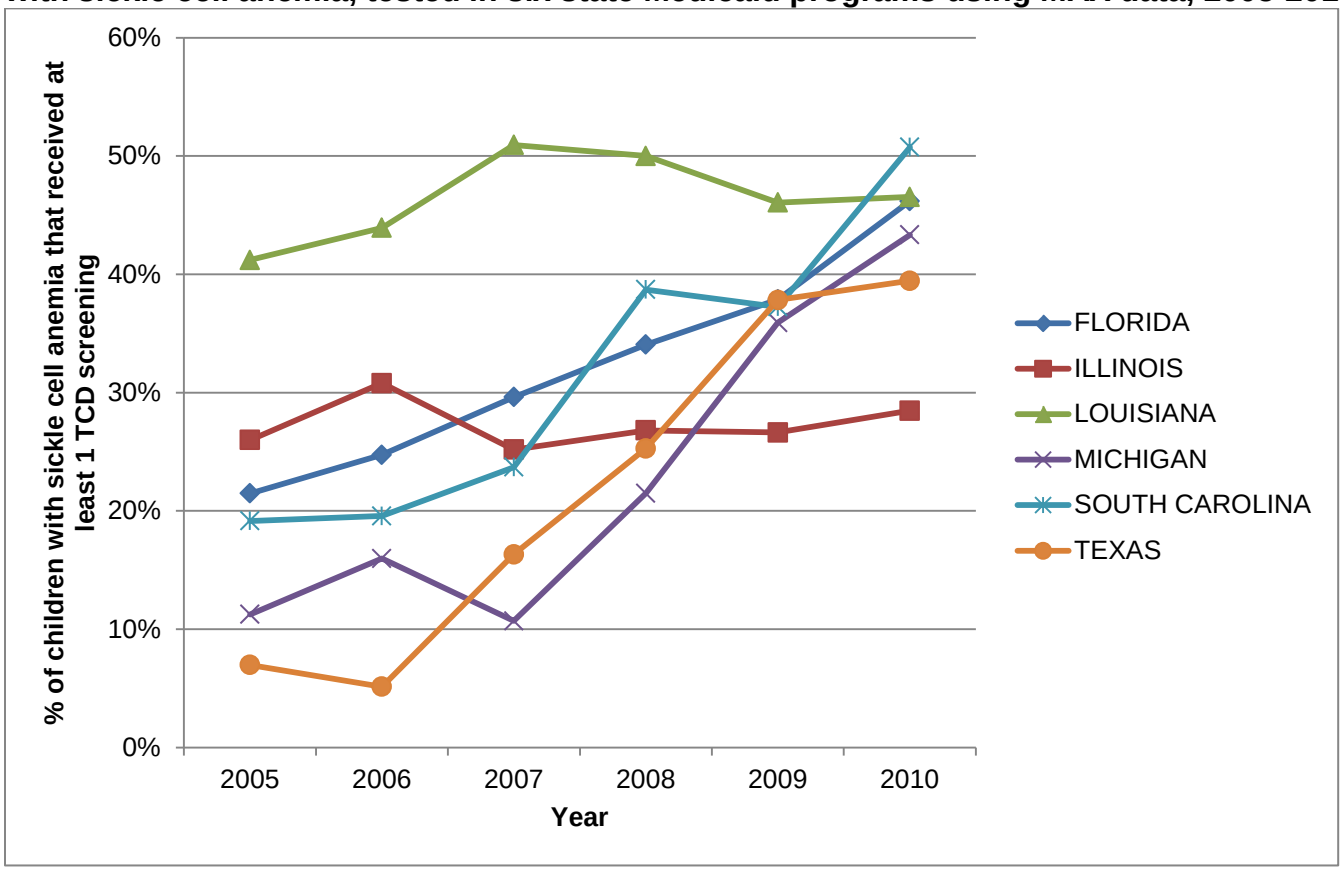
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Using the MAX data, the proportion of children receiving annual TCD screening was calculated for each year in the study period (2005 - 2010). We examined differences in performance across the 6 years included within this dataset. Logistic regression was used to estimate the associations between each year and receipt of TCD screening, with 2005 used as the reference category. Generalized estimating equation (GEE) models with robust standard errors were used to account for the correlation among children. Odds ratios with 95% confidence intervals were used to assess the final associations. The presence of trends in TCD screening rates were also assessed over time using linear regression. For all models, regression diagnostics were performed to assess normality of error variances.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The proportion of children receiving TCD screening ranged from 7% to 51% (Figure 2).

Figure 2. Trends for transcranial Doppler screening within the measurement year for children with sickle cell anemia, tested in six state Medicaid programs using MAX data, 2005-2010



Compared with 2005, children had higher odds of receiving TCD screening; these odds were statistically significant starting in 2007 (Table 8). Results from the linear regression model indicated that these rates did increase over time (p=0.0001).

Table 8. Odds of receipt of TCD screening among children with sickle cell anemia enrolled in 6 state Medicaid programs by year using MAX data, 2005-2010

Year	Odds Ratio	95% Confidence Interval	p-value
2005	Reference	Reference	N/A
2006	1.09	0.96, 1.25	0.17
2007	1.26	1.10, 1.44	0.0008
2008	1.60	1.40, 1.83	<0.0001
2009	1.94	1.69, 2.22	<0.0001
2010	2.36	2.10, 2,70	<0.0001

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

This measure was successfully able to distinguish differences in performance across years; the measure was also able to detect changes over time. As children in all years after 2005 had increased

odds of receipt of TCD screening compared with children in 2005, these results demonstrate that the likelihood of receiving a TCD screening did increase significantly over time.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the*

selected approach for missing data and what are the norms for the test conducted; <u>if no empirical analysis</u>, provide rationale for the selected approach for missing data)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
This measure was tested using Medicaid administrative claims data. The primary information needed for this measure includes a unique member identifier, health plan enrollment information, date of birth, dates of service, diagnosis codes, and procedure codes. These data are widely available, although obtaining them may require a restricted-use data agreement. For multiple-state comparisons, Medicaid Analytic eXtract (MAX) data are available from the Centers for Medicare & Medicaid. When the measure is used at the single-state level, state health departments can use their own Medicaid data.

The Quality Measurement, Evaluation, Testing, Review, and Implementation Consortium (Q-METRIC) testing determined that this measure is feasible using existing data from administrative claims systems. While Q-METRIC testing efforts support the feasibility of implementing this measure, the testing process demonstrated the technical challenges that may exist when identifying sickle cell anemia cases from very large administrative claims files, such as MAX data.

This measure was also tested using Medicaid administrative claims data acquired directly from the state of Michigan. Acquisition of data directly from state Medicaid programs requires the cooperation of those jurisdictions, as well as modification of the statistical

programming code developed for use with MAX files. Such modifications are necessary given the unique structure of the data files obtained directly from state Medicaid programs.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
N/A

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	Public Health/Disease Surveillance https://www.health.ny.gov/ New York State Health Department

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Name: New York State Health Department; Sponsor: Dr. David Anders
Purpose: Assess rates of TCD screening among children with sickle cell anemia in the state of New York
Geographic Area: Children with sickle cell anemia born from 2006-2014 enrolled in Medicaid in the state of New York

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in

use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No unintended negative consequences to individuals or populations were identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Q-METRIC – University of Michigan
Co.2 **Point of Contact:** Sarah, Reeves, sleasure@umich.edu, 734-615-9755-
Co.3 **Measure Developer if different from Measure Steward:** Q-METRIC – The University of Michigan
Co.4 **Point of Contact:** Gary, Freed, gfreed@med.umich.edu, 734-232-0657-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The face validity of this measure was established by a national panel of experts and advocates for families of children with sickle cell disease (SCD) convened by the Quality Measurement, Evaluation, Testing, Review, and Implementation Consortium (Q-METRIC) at the University of Michigan. The Q-METRIC Representative Panel included nationally recognized experts in SCD, representing hematology, pediatrics, and SCD family advocacy. The Q-METRIC Feasibility Panel included experts in state Medicaid program operations, health plan quality measurement, health informatics, and health care quality measurement. In total, the Q-METRIC SCD panels included 14 experts, providing a comprehensive perspective on SCD management and the measurement of quality metrics for states and health plans.

The Q-METRIC expert panels concluded that this measure has a very high degree of face validity through a detailed review of concepts and metrics considered to be essential to effective SCD management and treatment. Concepts and draft measures were rated by this group for their relative importance. This measure was among the most highly rated, receiving an average score of 8.5 (with 9 as the highest possible score).

Sickle Cell Disease Representative Panel:
Samir Ballas, MD, Professor, Division of Hematology, Thomas Jefferson University, Philadelphia, PA
Mary E. Brown, President and Chief Executive Officer, Sickle Cell Disease Association, Los Angeles, CA
George Buchanan, MD, Pediatric Hematologist, University of Texas Southwest Medical Center at Dallas, TX
Peter Lane, MD, Pediatric Hematologist-Oncologist, Children's Healthcare of Atlanta Pediatric Hospital, Atlanta, GA
Suzette Oyeku, MD, Assistant Professor of Pediatrics, Albert Einstein College, Bronx, NY
Lynnie Reid, Parent Representative, Boston, MA
Elliott Vichinsky, MD, Pediatric Hematology-Oncology, Children's Hospital and Research Center, Oakland, CA
Winfred Wang, MD, Hematologist, St. Jude Children's Hospital, Memphis, TN

Sickle Cell Disease Feasibility Panel:

Cathy Call, BSN, MSC, Senior Policy Analyst and Director for Health Quality Research, Altarum Institute, Alexandria, VA
J. Mitchell Harris, PhD, Director Research and Statistics, Children's Hospital Association, (formerly NACHRI), Alexandria, VA
Kevin Johnson, MD, MS, Professor and Vice Chair of Biomedical Informatics, Vanderbilt University, Nashville, TN
Don Lighter, MD, MBA, FAAP, FACHE, Director, The Institute for Health Quality Research and Education, Knoxville, TN
Sue Moran, BSN, MPH, Director of the Bureau of Medicaid Program Operations and Quality Assurance, Michigan Department of Community Health, Lansing, MI
Joseph Singer, MD, Vice President Clinical Affairs, HealthCore, Inc., Wilmington, DE
C. Jason Wang, MD, PhD, Associate Professor of Pediatrics, Stanford School of Medicine, Stanford, CA

Q-METRIC Investigators:
Kevin J. Dombkowski, DrPH, MS, Research Associate Professor of Pediatrics, School of Medicine, University of Michigan, Ann Arbor, MI
Gary L. Freed, MD, MPH, Professor of Pediatrics, School of Medicine and Professor of Health Management and Policy, School of Public Health, University of Michigan, Ann Arbor, MI (principal investigator)

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** N/A
Ad.7 **Disclaimers:** This work was funded by the Agency for Healthcare Research and Quality (AHRQ) and the Centers for Medicare & Medicaid Services (CMS) under the CHIPRA Pediatric Quality Measures Program Centers of Excellence grant number U18 HS020516. AHRQ, in accordance to CHIPRA 42 U.S.C. Section 1139A(b), and consistent with AHRQ´s mandate to disseminate research results, 42 U.S.C. Section 299c-3, has a worldwide irrevocable license to use and permit others to use products and materials from the grant for government purposes, which may include making the materials available for verification or replication by other researchers and making them available to the health care community and the public, if such distribution would significantly increase access to a product and thereby produce substantial or valuable public health benefits. The Measures can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the measures require a license agreement between the user and the Quality Measurement, Evaluation, Testing, Review and Implementation Consortium (Q-METRIC) at the University of Michigan (U-M). Neither Q-METRIC/U-M nor their members shall be responsible for any use of the Measures. Q-METRIC/U-M makes no representations, warranties or endorsement about the quality of any organization or physician that uses or reports performance measures, and Q-METRIC/U-M has no liability to anyone who relies on such measures. The Q-METRIC performance measures and specifications are not clinical guidelines and do not establish a standard of medical care.

This statement is signed by Gary L. Freed, MD, MPH, who, as the principal investigator of Q-METRIC, is authorized to act for any holder of copyright on the submitted measure.

Gary L. Freed, MD, MPH
Percy and Mary Murphy Professor of Pediatrics, School of Medicine
Professor of Health Management and Policy, School of Public Health
Principal Investigator, Q-METRIC
Child Health and Evaluation Research (CHEAR) Unit
Division of General Pediatrics
University of Michigan Hospital and Health Systems
Ann Arbor, MI 48109-5456

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2799
Measure Title: Use of Multiple Concurrent Antipsychotics in Children and Adolescents
Measure Steward: National Committee on Quality Assurance
Brief Description of Measure: The percentage of children and adolescents 1–17 years of age who were on two or more concurrent antipsychotic medications.
Developer Rationale: This measure addresses inappropriate prescribing patterns as one facet of safe and judicious use of antipsychotics in children and adolescents. Antipsychotic prescribing for youth has increased rapidly in recent decades. Although antipsychotic medications may serve as effective treatment for a narrowly defined set of psychiatric disorders in youth, less is known about the safety and effectiveness of antipsychotic prescribing patterns in community use (e.g., combinations of medications). Risks of multiple concurrent antipsychotics in comparison to monotherapy have not been systematically investigated. Existing evidence about the harms of multiple concurrent antipsychotic use in children appears largely in case reports and includes increased risk of serious drug interactions, delirium, serious behavioral changes, cardiac arrhythmias and death.

Numerator Statement: Children and adolescents who are on two or more antipsychotic medications concurrently for at least 90 days.
Denominator Statement: Children and adolescents who received 90 days or more of continuous antipsychotic medication treatment.
Denominator Exclusions: N/A

Measure Type: Process
Data Source: Administrative claims
Level of Analysis: Health Plan, Integrated Delivery System, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this process measure:

- The rate in this measure (multiple concurrent antipsychotics) relates to the desired outcome (optimal mental and physical outcomes) in the following way: Healthcare provider does not prescribe multiple concurrent antipsychotics >>> Patient receives safer treatment for psychiatric condition present >>> Patient avoids adverse side effects associated with use of multiple concurrent antipsychotic medications >>> Patient experiences improvement in mental and physical outcomes (desired outcome).

- The developer states that "The specific recommendation upon which this measure is based addresses the use of multiple antipsychotics concurrently and notes that the use of multiple antipsychotics has not been studied rigorously and should be avoided. This recommendation is based on established risks of antipsychotics, such as dangerous drug interactions, delirium, serious behavioral changes, cardiac arrhythmias, and death. These risks are in addition to the established side effects of antipsychotic medications that include metabolic disturbance, a serious concern for children."
- The measure is based on clinical practice guidelines. Four guidelines from three organizations are referenced, three of which are ratings. The ratings are:
    - American Academy of Child and Adolescent Psychiatry (AACAP) – not endorsed: ineffective or contraindicated
    - AACAP endorsed best practice principles: Best-practice principles that underlie medication prescribing, to promote the appropriate and safe use of psychotropic medications
    - TMAY Ratings – uses Oxford Centre for Evidence-Based Medicine, guideline is rated C (Level 4 studies or extrapolations from level 2 or 3 studies), very strong recommendation
- While there are several guidelines in this area, the developer focuses on the AACAP guideline since it is most relevant to the measure focus:
    - **Recommendation 8:** "The simultaneous use of multiple concurrent AAAs has not been studied rigorously and generally should be avoided." – Based on a literature review of 147 publications that included clinical trials, meta-analysis, practice guidelines, RCTs, systematic literature reviews, and case reports and series.
    - **Principle 12: "**The prescriber needs a clear rationale for using medication combinations…there is limited evidence in children and adolescents for the use of two antidepressants or two antipsychotics as an initial treatment approach or as a specific endpoint for treatment." – Based on a literature review of 147 publications that included clinical trials, meta-analysis, practice guidelines, RCTs, systematic literature reviews, and case reports and series.
- The developer indicates that the quality of evidence for avoiding multiple concurrent psychotic medications is high.
- No exact estimate exists of the benefits of avoidance of the multiple use of antipsychotic medications, but the short- and long-term risks of these medications in general is well-established.

Questions for the Committee

- *Is the relationship of this measure to patient outcomes supported by the evidence and, if so, how strong is the evidence for this relationship?*
- *The measure specifies concurrent use of medications for 90 days, but the guidelines do not appear to specify a timeframe. Is the timeframe reasonable? Does the Committee wish to explore this further with the developer?*

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following information:

- In a review of the literature, one systematic review found that among youth prescribed any antipsychotic, about one in 10 (9.6 percent, SD 7.2 percent) received multiple concurrent antipsychotics (Toteja et al., 2013). Other studies of multiple concurrent antipsychotics among youth prescribed any antipsychotic have found that prevalence among adolescents is twice that of younger children, and that the rate among adolescents has increased two-fold from the 1990s to the 2000s (Toteja et al., 2013). Another study of a large state Medicaid fee-for-service program found that about 7 percent of children 6–17 years of age on any antipsychotic were prescribed two or more antipsychotics for longer than 60 days (Constantine et al., 2010).
- The developer assessed use of multiple concurrent antipsychotics in Medicaid children using 2008 MAX data from 11 states. It found average rates of 6 percent, with a range of 2.9 to 9.4 percent (a lower rate indicates better care). For children in foster care, the average rate was 6.8 percent, with a range of 1.9 to 10.6 percent. In additional field-testing in Medicaid health plan data from one state, the average percentage of children 0-20

years with use of multiple concurrent antipsychotics was 4.4 percent, with a range of 1.8 percent to 7.0 percent.
- Disparities were noted. In particular, eight states at higher rates of multiple concurrent antipsychotic use in the foster care population compared with the general Medicaid population. Use was higher in adolescents than younger children.
- In both the general and foster care populations, rates were higher for black children and adolescents than Hispanic and white children. For the general population, rates were higher for metropolitan children as opposed to children in rural areas, but for the foster care population, higher rates were seen in rural areas.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- Supported. The concurrent use for 90 days makes sense.
- While there is evidence to say polypharmacy has increased risk, there is little evidence to say that high dose of 1 medication (4 mg risperdal) is safer than low doses of 2 medications (1 mg risperdal in am and 25 mg seroquel at night)... while this is not the norm, there may be rare clinical times where it is more appropriate and potentially safer and more effective. in general, i imagine these complex cases get referred to psychiatrists.
- "The context of this measure makes it inherently complex. While it's clear there is professional consensus that more than one drug is to be avoided, on a case-by-case basis this is likely to be considered in a small number of children with quite significant mental health conditions.
- The expert opinion and professional consensus appears to be strong given the guidelines referenced, but the direct evidence to support is much weaker. Some specific examples:
- Even the guidelines themselves state that ""use of multiple concurrent AAA's has not been studied rigorously and generally should be avoided."" This is not direct evidence of harm, but rather absence of evidence of safety. Their finding is that the use (of multiple AAA's is ""not endorsed,"" which speaks to lack of evidence rather than affirmative evidence of harm.
- Principle 12 only says that ""prescriber needs a clear rationale for using medication combinations…"" and that the principle applies to prescribing ""as an initial treatment approach."" I don't think the measure limits itself to initial treatment.
- The developers refer to a review of 147 studies, but it is unclear how many of these really relate to the specific point of harm from >1 antipsychotic. The AACAP-AAA review did not produce estimates of the benefit of avoidance. In sec 1.c.3 the developer state: ""Risks of multiple concurrent antipsychotics in comparison to mono therapy have not been systematically investigated."" In summary, I do not agree with the statement from the developers that the ""quality of the evidence in support of avoiding multiple concurrent antipsychotic medications is high.""""
- Measure of a process - I don't see clear evidence described that links concomitant use directly to poor outcomes for the child. There are myriad studies describing adverse effects of antipsychotic use in children, however, so it reasons that concomitant use will amplify these effects. Several clinical practice guidelines cited which directly address the focus of this measure. I suspect 90 days was selected to allow for some 'washout' period if a child is being transitioned from one antipsychotic to another. While individual practice will vary on the timeline to transition from one medicine to another, 90 days is sufficient time for any transition to have occurred.
- While there is little research on the use of multiple antipsychotics with children and adolescents, it is well established that the use of antipsychotics can increase metabolic disorder, cardiac issues, behavioral changes, and other significant problems.
- The Measure proposed looks to support the contention that multiple concurrent antipsychotic use in Children and Adolescents may result in numerous negative affects. This measure and the collection of data proposed if this Measure is endorsed has a goal to decrease the use of multiple concurrent antipsychotic drugs in these populations. the Developers state strong evidence against this multiple drug use. The evidence relates well to the process

Measure proposed and is supported by the stated rationale. Recommendation 8 describes the evidence used to support this measure.

1b. Performance Gap.

- There is overall less than optimal performance. The developer stated there were racial disparities, but did not provide specific numbers.
- Performance gap exists. the measure describes disparities by population subgroups (race), but may also be explained by other psychosocial risk factors (high stress home environment, lack of supports/resources, attachment issues, etc.).
- A performance gap is demonstrated by variability in rates seen at the state level (using Medicaid MAX data, 2008)-- they were overall 6% with a range of 2.9 to 9.4. Similar ranges were seen at the health plan level (Medicaid plans from one state). The mean rate is not "0" but given the concerns above, it's not clear what the right number is that would balance risks and benefits. Especially when it's possible that some children have already failed single medication treatment. While the relative difference are great (3-fold between lowest and highest) the absolute differences are less dramatic. However, there does appear to be implicit consensus of experts that the current rate is too high.
- Disparities are demonstrated by race/ethnicity and by age (African American children and adolescents more likely to receive >1 agent).
- Performance data is provided, including a systematic review, a single state review of Medicaid FFS data and an 11-state MAX data review. There is a clear gap in care and data cited to suggest this gap is worsening over time. There appear to be clear disparities in certain subpopulations (black children, adolescents, children in foster care) and I would indicate this as disparities sensitive.
- The rate of using multiple antipsychotics with this age group has doubled in the past twenty years.
- There appears to be higher utilization of multiple antipsychotics in minority youth and those in foster care.
- Performance data on the measure was provided including state, health plans and other data. Data at the state level is cited. For example, the Developer looked at Medicaid recipients, comparing those in Foster Care and total Medicaid recipients. in these cases higher use of multiple drugs in Foster Care in 8 states. Also, higher use in Black patients than other groups is seen in some groups. From their discussion a gap in care is seen comparing Medicaid to other groups.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The numerator for this measure is children and adolescents who are on two or more antipsychotic medications concurrently for at least 90 days. The denominator is children and adolescents who received 90 days or more of continuous antipsychotic medication treatment.
- The numerator and denominator details provide steps to identify patients for inclusion and include a list of medications. No codes are needed to calculate the measure.
- The measure is stratified by age, but is not risk adjusted.

Questions for the Committee.

- *Are all appropriate medications included?*
- *Is the logic or calculation algorithm clear?*

- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

- This measure was tested at the performance measures score level using a beta-binomial signal-to-noise analysis. For this type of testing, a score of zero implies that all the variability in a measure is attributable to measurement error. A score of 1.0 implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one reporting entity from another. A score of 0.7 or higher indicates adequate reliability to distinguish performance between two entities and is considered acceptable.
- Per the NQF algorithm, reliability testing at the computed performance measure score may be rated HIGH, MODERATE, or LOW depending on the testing results.
- The developer reports the following testing results:
  - The average state level reliability was 0.99, and the minimum was 0.96, suggesting high reliability at the state level.
  - The reliability for Medicaid health plans averaged 0.64, with a minimum of 0.28.
  - The reliability for commercial health plans averaged 0.42 average, with a minimum of 0.08.

Questions for the Committee

- *The developer concludes the measure is reliable only at the state level. Does the Committee concur?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The specifications are consistent with the evidence. The goal of the measure is to assess inappropriate prescribing of antipsychotic medication to children and adolescents. The evidence provided supports the specifications.

Question for the Committee

- *Are the appropriate medications included in the specifications?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- The measure was tested at the performance measure score level using both empirical testing and face validity.
- For the empirical testing, the developer assessed construct validity with two types of analyses: correlations among measures and rankings of health plans and states on measures on the three antipsychotic medication measures.
  - Correlations were tested using health plans, as there was not enough entities to test between states. The results showed that among Medicaid health plans, there were no statistically significant correlations between the *Multiple Concurrent* measure and the other measures addressing antipsychotic use in children and adolescents. Among national commercial plans, there was moderate negative correlation between the *Follow-up Visit* and *Multiple Concurrent* measures (r=-.58, p=0.02).
  - The developer states that "Among MAX states and one state's Medicaid plans, we found good consistency in the states and plans, respectively, with the best and worst performance." Their interpretation is that the results show that plans and states can be approximately ranked based on profiles of performance across multiple measures. The consistent performance across measures suggest the measures are assessing a dimension of quality.
- Per the NQF algorithm, validity testing at the computed performance measure score may be rated HIGH,

MODERATE, or LOW depending on the testing results.

- The developer used its standardized HEDIS process to test face validity of the measure construct, but does not explicitly call out face validity of the computed performance score, as required by NQF.
  - The developer worked with five expert panels to identify the most appropriate method for assessing the use of multiple concurrent antipsychotics among this patient population. All of the panels concluded this measure was specified to assess multiple concurrent use of antipsychotics.
  - The draft measure was put out for public comment and brought to the developer's Committee on Performance Measurement.
  - The developer states that the measure has sufficient face validity.

Questions for the Committee

- *Do the results of the empiric testing demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you believe that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- There are no exclusions.

Questions for the Committee

- *Should there be any exclusions for this measure?*
- *Does the Committee believe there are other threats to validity?*

2b4. Risk adjustment:

- The measure is not risk adjusted.

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer states that the results indicate that there is 2.1% gap in performance between Medicaid plans at the 25th and 75th percentiles, a 3.2% gap in performance among commercial plans and a 4.4% gap in performance among states at the 25th and 75th percentiles. This means states at the 75th percentile have on average 504 more children and adolescents receiving multiple concurrent antipsychotics than states at the 25th percentile.

Question for the Committee

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- This is not needed.

2b7. Missing Data

- The measure is collected using all administrative data sources. According to the developer there are no missing data, so this is not applicable.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- Would like clarification on the 90 day continuous use specification for numerator hits. Must the child be on the same two for 90 consecutive days for both, or would a child who is on one for 90 days and a second one for the first 45 days and a different "second" one for the remaining 45 days be included in the numerator.
- I'm curious if there is one scheduled med (ability for example), and one prn med (risperdal m-tab as needed for agitation) - if that gets included in numerator.
- The language of the numerator and denominator could be more precise. Children who are "on" these medications is imprecise. Rather, the numerator (and denominator) are based on pharmacy dispensings covering a 90-day period. The numerator calculation is complex but understandable. I wonder if there is a typo: it says that if the number of days between the end date (of dispensing 1) and the start day of dispensing 2 "= 15" days, the gap days should be counted. My guess is that this should say "<=15 days".

It is not clear how the generic drug names are to be translated into the measure calculation. Should there be a list of NDC codes, or is it left to each plan to determine how to capture these.
For the denominator, it's not clear why a 32 day gap is allowed (compared to a 15 day gap in the numerator).
Review for whether all appropriate medications are included must include content experts (e.g. pharmacists), or reliance on the process used by the developers.

- Numerator and denominator are clearly defined. Appropriate medications are included. The calculation algorithm is clear and appears it can be consistently implemented.
- The measure specifies both first and second generation antipsychotics and appears thorough.
- Reliability specifications were submitted in detail. It is my understanding that codes with descriptors were not provided. One presented analysis in which a score of 1.0 indicates high reliability was presented. This Measure was shown to have a score of .99 average at the state level (.96 minimum) indicating high reliability for the use of this Measure. It appears to me that this measure can be consistently implemented.

2a2. Reliability testing

- Reliability was tested with the MAX data set (11 states), 17 Medicaid health plans within one state, and a sample of commercial plans. The method for reliability testing is the beta-binomial-signal to noise method, which is appropriate. The reliability is acceptable in the very large state-level analysis. But, it is of borderline acceptability in the Medicaid plans (that have larger sample sizes) but was not acceptable (minimum reliability .08) in some of the commercial plans. This is directly related to the number of children meeting denominator criteria. For example, 24 of 72 commercial plans had less than 30 children (so were excluded). 25% of the included plans had less than 90. This highlights the need for using this measure only in settings with sufficient samples of children meeting the denominator criteria. By the algorithm I would rate the reliability as Low at the health plan level and Moderate at the state level.
- Calculations suggest reliability is high only at the state level.
- My understanding of this work is limited. It does appear that reliability was completed and it was determined that measurements were reliable at the state level only.

2b1. Validity Specifications

- What does the developer attribute to the disparity between the reliability for Medicaid/commercial health plans compared to state level.
- The measure specifications are consistent with the intent, and consistent with the evidence, at the level presented-- with the caveats above. The list of medications appears reasonable, but requires review by individuals with content expertise.
- Appropriate medications are included.
- The testing suggests that this measures is a valid measure for assessing the rate at which providers prescribe more than one antipsychotic medication in youth.
- The reliability testing evidenced a score of .9, suggesting the measure is highly reliable at detecting differences at the state level. Reliability estimates for health plans was significantly lower.
- It does not appear to me that there are specifications inconsistent with the evidence. It also appears that the target population (children and adolescents) values and would be served by more consistent avoidance of multiple antipsychotic drug use if this measure is endorsed.

2b2. Validity Testing

- There is good face validity on the measure based on the developers use of five expert panels and opportunity for public comment. The empirical validity studies are less convincing. Correlation with other measures in this topic area is is poor using Spearman correlation coefficients. The ranking method suggests only rough stability of rankings (of health plans and states) across related measures. In a sense, this is as much about reliability as validity, and is difficult to interpret with no quantification of what would be considered "good" validity. I would consider the results of the empirical testing Low, but the face

validity as Moderate.

- Developers demonstrate consistency with this measure in comparison to other antipsychotic measures, as well as standardized method to demonstrate face validity. Score on this measure is an indicator of quality.
- The measure appears to distinguish between low and high performer states well.
- Measure was ranked as a high priority by expert panel (face validity).
- As with 2a2, my understanding of this work is limited. Validity was tested on several levels.

2b3-2b7. Threats to Validity

- Because this is based on pharmacy claims, it's likely the data are complete. However, some caveats need to be made. It relies on patients having pharmacy benefits, and always using the same plan for dispensings. Because the numerator and denominator both rely on this, I do not believe it is a major concern. There is no information on medicines that the patient did not take (or were discontented by the prescriber). The developers should address whether any of these are concerns.
- Developers state this is administrative data, and therefore there are no missing data.
- Authors indicate no missing data
- There were no exclusions included in this measure. 2b5 - Gaps were seen that this measure proposed to decrease.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer notes:

- These elements are all generated through normal process of care and are in defined fields in electronic claims.
- The measure is a part of HEDIS, which has a standardized collection and calculation process, as well as a system to collect real-time feedback from measure users.
- Field testing results showed the measure is feasible to be collected by health plans and states using administrative claims data.
- As part of HEDIS, the data elements are subject to that program's data collection and audit requirements.
- This is not an eMeasure.

Questions for the Committee

- *Are the required data elements routinely generated and used during care delivery?*
- *Does the testing data collection strategy indicate the measure is ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- While polypharmacy is not desired, there may be relatively rare clinical cases where it is justified and does not represent poor quality of care. in general, complex cases should involve specialists.
- Feasibility of the measure is good, given that it is based in pharmacy claims only, and is currently being used (on a voluntary basis) as a HEDIS measure. Usability is good based on these as well.
- All required data elements are routinely generated and used in the course of normal care delivery. No concerns about putting this measure into operational use.
- Feasible because it relies on administrative claims data. Measure can be obtained through data that is secured through routine daily care.
- The data elements required in this measure are routinely generated and used during care delivery. No

concerns regarding collection strategy.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- The measure is currently in use in for both quality improvement with benchmarking and public reporting.
- It is included in Quality Compass for Medicaid 2015, a tool that displays health plan-level performance rates for HEDIS measures. It is used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance.
- The measure also is reported on in The State of Health Care Quality Report, a national report produced by the developer including the results from HEDIS measures.
- This is a new measure and improvement results are not yet available.
- No unintended consequences have been reported thus far.

Question for the Committee

- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Would like to hear developers provide perspective on HEDIS 2015 analysis for this measure. Are the rates cited in the results of the testing of the measure rate per 1,000 or rate per 100.
- The measure has been approved for use in the Quality Compass for Medicaid.

Criterion 5: Related and Competing Measures

- This measure, 2799, is related to one NQF-endorsed measure, 2337: Antipsychotic Use in Children Under 5 Years Old.
- This measure has a different target population of those who have continuous use of antipsychotics for 90 days or more, includes more children (up to age 18 years), and has a different focus (i.e., a specific type of non-recommended practice [multiple concurrent use] as opposed to any use).

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** N/A

Measure Title: Use of Multiple Concurrent Antipsychotics in Children and Adolescents

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 10/9/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Multiple concurrent antipsychotic medication avoided for those with continuous antipsychotic medication treatment

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

N/A

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

N/A

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The rate in this measure (multiple concurrent antipsychotics) relates to the desired outcome (optimal mental and physical outcomes) in the following way:

Health care provider does not prescribe multiple concurrent antipsychotics >>> Patient receives safer treatment for psychiatric condition present >>> Patient avoids adverse side effects associated with use of multiple concurrent antipsychotic medications >>> Patient experiences improvement in mental and physical outcomes (desired outcome).

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

- American Academy of Child and Adolescent Psychiatry. Practice Parameter for the Use of Atypical Antipsychotic Medications in Children and Adolescents. http://www.aacap.org/App_Themes/AACAP/docs/practice_parameters/Atypical_Antipsychotic_Medications_Web.pdf (July 12, 2012)
- American Academy of Child and Adolescent Psychiatry. September 2009. Practice parameter on the use of psychotropic medication in children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry.* 48(9):961–73.
- Scotto, Rosato N., C.U. Correll, E. Pappadopulos, A. Chait, S. Crystal, P.S. Jensen. June 2012. Treatment of maladaptive aggression in youth: CERT guidelines II. Treatments and ongoing management. *Pediatrics.* 129(6):e1577–86.
- Texas Department of Family and Protective Services and University of Texas at Austin College of Pharmacy. 2013. Psychotropic Medication Utilization Parameters for Foster Children. http://www.dfps.state.tx.us/documents/Child_Protection/pdf/TxFosterCareParameters-September2013.pdf (October 22, 2013)

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Guideline (Date)	Population	Recommendation or Statement	Type/Grade
AACAP-AAA (2011) Practice parameter for the use of atypical antipsychotic medications in children and adolescents	5-18 years	"The simultaneous use of multiple concurrent AAAs has not been studied rigorously and generally should be avoided." (Recommendation 8)	Not Endorsed
AACAP-PsyMed (2009) Practice parameter on the use of psychotropic medication in children and adolescents	≤18 years	"The prescriber needs a clear rationale for using medication combinations….there is limited evidence in children and adolescents for the use of two antidepressants or two antipsychotics as an initial	Best practice principle

		treatment approach or as a specific endpoint for treatment." (Principle 12)	
TMAY (2012) Center for Education and Research on Mental Health Therapeutics—Treatment of maladaptive aggression in youth	≤18 years	Use of two simultaneous psychotropic medications should be avoided (Recommendation 18)	Evidence: C Strength of Recommendation: Very Strong
TX (2010) Texas Department of Family and Protective Services – Psychotropic medication utilization parameters for foster children	Children (age un-specified)	Prescribing multiple antipsychotics is a situation that warrants clinical review.	Not specified*

*TX (2010) did not specify the use of a rating system.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

Guideline Developer	Definition
AACAP	**Not endorsed:** Ineffective or contraindicated.
AACAP endorsed best-practice principles	**Best-practice principles** that underlie medication prescribing, to promote the appropriate and safe use of psychotropic medications
TMAY Ratings	Oxford Centre for Evidence-Based Medicine grade of evidence (A-D) C: Level 4 studies or extrapolations from level 2 or 3 studies
	Strength of Recommendation: Very strong (≥90% agreement)

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Guideline Developer	Definition
AACAP	**Minimal Standard/ Clinical Standard:** Rigorous/ substantial empirical evidence (meta-analyses, systematic reviews, RCTs) and/or overwhelming clinical consensus; expected to apply more than 95 percent of the time
	Clinical guidelines: Strong empirical evidence (non-randomized controlled trials, cohort or case-control studies), and/or strong clinical consensus; expected to apply in most cases (75% of the time)
	Options: Acceptable but not required; there may be insufficient evidence to support higher recommendation (uncontrolled trials, case/series reports)

Guideline Developer	Definition
TMAY Ratings	Oxford Centre for Evidence-Based Medicine grade of evidence (A-D)
	A: Consistent level 1 studies
	B: Consistent level 2 or 3 studies or extrapolations from level 1 studies
	D: Level 5 evidence or troublingly inconsistent or inconclusive studies of any level
	Strength of Recommendation: Strong (70-89% agreement)
	Strength of Recommendation: Fair (50-69% agreement)
	Strength of Recommendation: Weak (<50% agreement)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

OCEBM Levels of Evidence Working Group. 2011. The Oxford 2011 levels of evidence. http://www.cebm.net/index.aspx?o=5653 (October 12, 2013)

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

N/A

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

N/A

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

N/A

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

N/A

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

N/A

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

N/A

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Several guidelines address the use of multiple concurrent antipsychotics in children and adolescents; each guideline cautions against the use of multiple concurrent antipsychotics given the lack of evidence supporting this type of use. While we list the full range of guidelines in sections 1a.4.2 and 1a.4.3 above, we focus on and describe in more detail the American Academy of Child and Adolescent Psychiatry (AACAP) Guideline in the remaining sections, as it is most closely relevant to the specified measure. The AACAP guideline addresses the use of antipsychotic medications in children and adolescents. The specific recommendation upon which this measure is based addresses the use of multiple antipsychotics concurrently and notes that the use of multiple antipsychotics has not been studied rigorously and should be avoided. This recommendation is based on established risks of antipsychotics, such as dangerous drug interactions, delirium, serious behavioral changes, cardiac arrhythmias, and death. These risks are in addition to the established side effects of antipsychotic medications that include metabolic disturbance, a serious concern for children.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

See table under 1a.4.2 for the level of evidence grade given to each guideline. See table under 1a.4.3 for the definition of the level of evidence grade given to each guideline.

AACAP Strength of Empirical Evidence

AACAP rates the strength of the empirical evidence in descending order as follows:

- (rct) Randomized, controlled trial is applied to studies in which subjects are randomly assigned to two or more treatment conditions
- (ct) Controlled trial is applied to studies in which subjects are non-randomly assigned to two or more treatment conditions
- (ut) Uncontrolled trial is applied to studies in which subjects are assigned to one treatment condition

- (cs) Case series/report is applied to a case series or a case report

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See table under 1a.4.4 for the definition of the level of evidence grade not given to the guidelines.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1990-2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

The guidelines listed in our table above address antipsychotic polypharmacy among children. The AACAP-AAA recommendation is rated a "Not Endorsed," indicating there is no rigorous/substantial empirical evidence and/or overwhelming clinical consensus to support prescribing multiple concurrent antipsychotics for children and adolescents. The guideline states that "due to the possibility of significant risks associated with these agents [atypical antipsychotics], the use of more than one agent is not recommended and is not supported in the scientific literature". AACAP includes an additional, broader guideline around use of multiple concurrent psychotropic medications in youth; we focus on the antipsychotic-specific AACAP-AAA guideline here and describe the body of evidence for each relevant recommendation below.

When developing their guidelines, AACAP limited its evidence review to clinical trials, meta-analysis, practice guidelines, randomized controlled trials (RCTs), systematic literature reviews, and case reports and series. AACAP selected a total of 147 publications for careful examination based on their weight in the hierarchy of evidence attending to the quality of individual studies, relevance to clinical practice and the strength of the entire body of evidence. AACAP did not provide a breakdown of specific numbers of each publication type. We have identified where there are certain publication types available to support each guideline.

Recommendation 8: "The simultaneous use of multiple concurrent AAAs has not been studied rigorously and generally should be avoided."

This recommendation is based on a literature review conducted by a medical professional society on the established metabolic impacts of antipsychotics and other health risks and evidence of efficacy of psychosocial treatments. The literature review contained a total of 147 publications that included clinical trials, meta-analysis, practice guidelines, RCTs, systematic literature reviews, and case reports and series.

- American Academy of Child and Adolescent Psychiatry. Practice parameter on the use of psychotropic medications in children and adolescents. *J Am Acad Child Adolesc Psychiatry.* 2009;48:961-973.

Principle 12: "The prescriber needs a clear rationale for using medication combinations….there is limited evidence in children and adolescents for the use of two antidepressants or two antipsychotics as an initial treatment approach or as a specific endpoint for treatment."

This principle is based on a literature review conducted by a medical professional society on the established metabolic impacts of antipsychotics and other health risks and evidence of efficacy of psychosocial treatments. The literature review contained a total of 147 publications that included clinical trials, meta-analysis, practice guidelines, RCTs, systematic literature reviews, and case reports and series.

- American Academy of Child and Adolescent Psychiatry. Practice parameter on the use of psychotropic medications in children and adolescents. *J Am Acad Child Adolesc Psychiatry.* 2009;48:961-973.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The evidence review used by AACAP prioritized study designs less subject to bias and studies that represent the best scientific evidence. The evidence review included a large number of studies with large numbers of patients from various populations. Overall, the quality of the evidence in support of avoiding multiple concurrent antipsychotic medication is high.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

The AACAP-AAA review did not include an exact estimate of benefits of avoiding multiple concurrent antipsychotics in youth. However, the evidence has established that use of multiple concurrent antipsychotic is associated with adverse short-term psychotic, behavioral, cardiovascular, and other side effects in youth and to negative long-term health outcomes throughout the lifespan.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The AACAP review did not examine the potential harms of avoiding multiple concurrent antipsychotics in youth, which have not been thoroughly investigated. However, the harms of unnecessary antipsychotic use in general in kids has been well established (Andrade et al. 2011; Bobo et al., 2013; Correll, 2008; Correll et al., 2009; Crystal et al., 2009; Daniels, 2006; Lean and Pajonk, 2003; Safer et al., 2003; Srinivasan et al. 2002; Van Bennekom et al., 2013).

<u>Citations</u>

Andrade, S.E., J.C. Lo, D. Roblin, et al. December 2011. Antipsychotic medication use among children and risk of diabetes mellitus. Pediatrics. 128(6):1135–41.

Bobo, W.V., W.O. Cooper, C.M. Stein, et al. October 1, 2013. Antipsychotics and the risk of type 2 diabetes mellitus in children and youth. JAMA Psychiatry. 70(10):1067–75.

Correll, C.U. 2008. Antipsychotic use in children and adolescents: minimizing adverse effects to maximize outcomes. FOCUS: The Journal of Lifelong Learning in Psychiatry. 6(3):368–78.

Correll, C. U., Manu, P., Olshanskiy, V., Napolitano, B., Kane, J. M., & Malhotra, A. K. 2009. Cardiometabolic risk of second-generation antipsychotic medications during first-time use in children and adolescents. Journal of the American Medical Association. 302(16):1765-1773.

Crystal, S., M. Olfson, C. Huang, H. Pincus and T. Gerhard. 2009. Broadened use of atypical antipsychotics: Safety, effectiveness, and policy challenges. Health Affairs. 28:w770–81.

Daniels, S.R. 2006. The consequences of childhood overweight and obesity. The future of children. 16(1):47–67.

Lean, M.E., and F.G. Pajonk. 2003. Patients on Atypical Antipsychotic Drugs Another high-risk group for type 2 diabetes. Diabetes Care. 26(5), 1597–605.

Safer, D.J., J.M. Zito, S. DosReis. 2003. Concomitant psychotropic medication for youths. *American Journal of Psychiatry.* 160(3): p. 438–49.

Srinivasan, S. R., Myers, L., & Berenson, G. S. 2002. Predictability of childhood adiposity and Insulin for developing insulin resistance syndrome (syndrome X) in young adulthood the Bogalusa heart study. Diabetes. 51(1):204-209.

Van Bennekom, M., H. Gijsman, F. Zitman. 2013. Antipsychotic polypharmacy in psychotic disorders: A critical review of neurobiology, efficacy, tolerability and cost effectiveness. *Journal of Psychopharmacology.* 27: 327.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

To our knowledge, there have been no new studies that contradict the current body of evidence.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

N/A

1a.8.2. Provide the citation and summary for each piece of evidence.

N/A

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Multiple_Concurrent_Evidence.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure addresses inappropriate prescribing patterns as one facet of safe and judicious use of antipsychotics in children and adolescents. Antipsychotic prescribing for youth has increased rapidly in recent decades. Although antipsychotic medications may serve as effective treatment for a narrowly defined set of psychiatric disorders in youth, less is known about the safety and effectiveness of antipsychotic prescribing patterns in community use (e.g., combinations of medications). Risks of multiple concurrent antipsychotics in comparison to monotherapy have not been systematically investigated. Existing evidence about the harms of multiple concurrent antipsychotic use in children appears largely in case reports and includes increased risk of serious drug interactions, delirium, serious behavioral changes, cardiac arrhythmias and death.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
New measure: not applicable

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
In a review of the literature, one systematic review found that among youth prescribed any antipsychotic, about one in 10 (9.6 percent, SD 7.2 percent) received multiple concurrent antipsychotics (Toteja et al., 2013). Other studies of multiple concurrent antipsychotics among youth prescribed any antipsychotic have found that prevalence among adolescents is twice that of younger children, and that the rate among adolescents has increased two-fold from the 1990s to the 2000s (Toteja et al., 2013). Another study of a large state Medicaid fee-for-service program found that about 7 percent of children 6–17 years of age on any antipsychotic were prescribed two or more antipsychotics for longer than 60 days (Constantine et al., 2010).

As part of the measure's field-testing, we assessed use of multiple concurrent use of antipsychotic medications in Medicaid children, using the Medicaid Analytic eXtract (MAX) data files. Analysis of administrative claims data from 11 states demonstrated that the average percentage of children with use of multiple concurrent antipsychotics was 6.0 percent, with a range of 2.9 to 9.4 percent (a lower rate indicates better care). For children in foster care, the average rate was 6.8 percent, with a range of 1.9 to 10.6 percent. In additional field-testing in Medicaid health plan data from one state, the average percentage of children 0-20 years with use of multiple concurrent antipsychotics was 4.4 percent, with a range of 1.8 percent to 7.0 percent.

Citations

Constantine, R., M. Bengtson, T. Murphy, et al. 2012. Impact of the Florida Medicaid Prior-Authorization Program on use of antipsychotics by children under age six. Psychiatric Services. 12: DOI: 10.1176/appi.ps.201100346.

Toteja, N., J.A. Gallego, E. Saito, et al. 2013. Prevalence and correlates of antipsychotic polypharmacy in children and adolescents receiving antipsychotic treatment. International Journal of Neuropsychopharmacology. DOI: 10.1017/S1461145712001320.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We tested the measure and evaluated disparities in its performance using Medicaid Analytic eXtract (MAX) data. We assessed performance by age group (0-5, 6-11, 12-20), race/ethnicity, foster care status, and rurality/urbanicity.

MAX DATA DESCRIPTION AND RESULTS
Our MAX dataset was composed of 2008 service data from 11 states. The analysis population included all Medicaid enrollees aged 0-20 on December 31, 2008 in the 11 states. Both fee-for-service and managed care enrollees were included. Data files included person summary, outpatient claims, inpatient claims and prescription claims. States were chosen due to completeness of their data for managed care enrolled beneficiaries.

Of the 11 states, eight had higher rates of multiple concurrent antipsychotic use in the foster care population compared with the general Medicaid population. In both the general population and foster care population, multiple concurrent antipsychotic use was highest among adolescents compared with the lower age strata. In the general population, rates of multiple concurrent antipsychotic use were slightly higher among Black Non-Hispanic children and adolescents (7.5 percent) than Hispanic (6.1 percent) and White Non-Hispanic (6.5 percent) children and adolescents. Similarly, in the foster care population, rates of multiple concurrent antipsychotic use were slightly higher among Black Non-Hispanic children and adolescents (8.6 percent) than Hispanic (6.7 percent) and White Non-Hispanic (7.6 percent) children and adolescents. For the general population of children, higher rates of multiple concurrent antipsychotic use were seen in metropolitan areas (6.8 percent) than rural areas (5.7 percent). However, within the foster care population, higher rates were seen in rural areas (9.5 percent), compared with metropolitan areas (6.6 percent).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
A leading cause of morbidity/mortality, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Antipsychotic prescribing for children has increased rapidly in recent decades, driven by new prescriptions and by longer duration of use (Patten et al., 2012; Cooper et al., 2006). Although some evidence supports the efficacy of antipsychotics in youth for certain narrowly defined conditions, less is known about the safety and effectiveness of antipsychotic prescribing patterns in community use (e.g., combinations of medications, off-label prescribing, dosing outside of recommended ranges).

Children and adolescents prescribed antipsychotics are more at risk for serious health concerns, including weight gain, extrapyramidal side effects, hyperprolactinemia and some metabolic effects (Correll et al., 2011). Girls treated with certain antipsychotics may also be at increased risk for gynecological problems (Talib et al., 2013) and osteoporosis (Cohen et al., 2012).

Risks of multiple concurrent antipsychotics in comparison to monotherapy have not been systematically investigated; existing evidence appears largely in case reports, and includes increased risk of serious drug interactions, delirium, serious behavioral changes, cardiac arrhythmias and death (Safer et al., 2003). Research demonstrating that the pharmacokinetics of antipsychotics may vary by developmental stage (Correll et al., 2011) suggests that use of multiple concurrent antipsychotics may pose greater risks for children and adolescents compared to adults.

The financial impact of multiple concurrent antipsychotic use in children has not been examined; however, antipsychotics are a costly

form of drug therapy. Atypical antipsychotics have the greatest mean prescription cost ($132) of any psychotropic medication (Martin & Leslie, 2003) and until recently were the most costly drug class within the Medicaid program (Crystal et al., 2009). Additionally, there are substantial long-term costs of treating side effects associated with antipsychotic medications, including treatment of obesity, diabetes and dyslipidemias. There is some evidence that these health conditions, such as new onset diabetes, do not always resolve after discontinuation of the antipsychotic (Lean and Pajonk, 2003). Although this is an understudied area, it is reasonable to assume that unresolved side effects from antipsychotics would be associated with the long-term increases in health care costs that have been established for obesity and diabetes.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Cohen, D., O. Bonnot, N. Bodeau, et al. 2012. Adverse effects of second-generation antipsychotics in children and adolescents. Journal of Clinical Psychopharmacology. 32:309–16.

Cooper, W.O., P.G. Arbogast, H. Ding, G.B. Hickson, D.C. Fuchs, and W.A. Ray. 2006. Trends in prescribing of antipsychotic medications for US children. Ambulatory Pediatrics. 6(2):79–83.

Correll, C.U., C.J. Kratochvil, J.S. March. 2011. Developments in pediatric psychopharmacology: Focus on stimulants, antidepressants, and antipsychotics. Journal of Clinical Psychiatry. 72:655–70.

Crystal, S., M. Olfson, C. Huang, H. Pincus, and T. Gerhard. 2009. Broadened use of atypical antipsychotics: Safety, effectiveness, and policy challenges. Health Affairs. 28:w770–81.

Lean, M.E., F.G. Pajonk. 2003. Patients on Atypical Antipsychotic Drugs Another high-risk group for type 2 diabetes. Diabetes Care. 26(5), 1597–605.

Martin, A., D. Leslie. 2003. Trends in psychotropic medication costs for children and adolescents, 1997-2000. Archives of Pediatric Adolescent Medicine. 157(10):997–1004.

Patten, S.B., W. Waheed, L. Bresee. 2012. A review of pharmacoepidemiologic studies of antipsychotic use in children and adolescents. Canadian Journal of Psychiatry. 57:717–21.

Safer, D.J., J.M. Zito, S. DosReis. 2003. Concomitant psychotropic medication for youths. American Journal of Psychiatry. 160(3): p. 438–49.

Talib, H.J., E.M. Alderman. 2013. Gynecologic and reproductive health concerns of adolescents using selected psychotropic medications. Pediatric Adolescent Gynecology. 26:7–15.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

Behavioral Health, Mental Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Safety, Safety : Medication Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
None

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Children and adolescents who are on two or more antipsychotic medications concurrently for at least 90 days.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months (January 1 – December 31)

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Children and adolescents who are on two or more antipsychotic medications (Table APC-A) concurrently for at least 90 consecutive days during the measurement year (January 1 – December 31).

To identify the numerator: for each patient in the eligible population, by drug, identify all prescription events, start dates and end dates. Then identify the number of concurrent antipsychotic medication treatment events.
Step 1: For each patient, identify the first day during the measurement year where the patient was being treated with two or more different antipsychotic medications; this is the concurrent antipsychotic medication treatment event start date.
Step 2: Beginning with (and including) the start date, identify the number of consecutive days where the patient remains on two or more different antipsychotic medications. If the number of days =90 days, the patient is numerator compliant.
Step 3: If the number of consecutive days on multiple antipsychotic medications is <90 days, identify the end date and identify the next day during the measurement year where the patient was being treated with two or more different antipsychotic medications. If the number of days between the end date and the next start date is =15 days, include the days in the concurrent antipsychotic medication treatment event (concurrent antipsychotic medication treatment events allow for a 15-day gap).
Step 4: If the number of days between the end date and the next start date exceeds 15 days, end the event; using the new start date, continue to assess for concurrent antipsychotic medication treatment events.
Step 5: Continue this process until the number of concurrent antipsychotic medication treatment days is =90 consecutive days (i.e., the patient is numerator compliant) or until the measurement year is exhausted (i.e., no concurrent antipsychotic medication treatment events were identified during the measurement year).

Table APC-A: Antipsychotic Medications
First-generation antipsychotic medications: Chlorpromazine HCL; Fluphenazine HCL; Fluphenazine decanoate; Fluphenazine enanthate; Haloperidol; Haloperidol decanoate; Haloperidol lactate; Loxapine HCL; Loxapine succinate; Molindone HCL; Perphenazine; Pimozide; Promazine HCL; Thioridazine HCL; Thiothixene; Thiothixene HCL; Trifluoperazine HCL; Triflupromazine HCL

Second-generation antipsychotic medications: Aripiprazole; Asenapine; Clozapine; Iloperidone; Lurasidone; Olanzapine; Olanzapine pamoate; Paliperidone; Paliperidone palmitate; Quetiapine fumarate; Risperidone; Risperidone microspheres; Ziprasidone HCL; Ziprasidone mesylate

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Children and adolescents who received 90 days or more of continuous antipsychotic medication treatment.

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Children and adolescents age 1-17 as of December 31 of the measurement year (January 1 – December 31) who received 90 days or more of continuous antipsychotic medication (Table APC-A) treatment.

The instructions outlined here are designed to identify through pharmacy claims children with at least 90 days of continuous antipsychotic use. The measure allows for a 32-day gap in order to account for missed prescription fills, which can be common among children, particularly for off-label use for behavioral control.

Step 1: Identify patients in the specified age range who were dispensed an antipsychotic medication during the measurement year (January 1 – December 31).
Step 2: For each patient, identify all antipsychotic prescriptions during the measurement year. For each drug, identify start and end dates of the prescriptions. Starting with the first prescription in the measurement year determine if there is a second dispense date of that same drug. If there is no second dispensing event with the same Drug ID, the start date is the first prescription's dispense date and the end date is the start date plus the days supply minus one. If there is a second dispensing event of the same drug, determine if there are gap days (a 32-day gap is allowed). Calculate the number of days between (but not including) the first prescription's dispense date and the second prescription's dispense date. If the number of days is less than or equal to the first prescription's days supply plus 32 days, the gap is less than or equal to 32 days and is allowed. The start date is the first prescription's dispense date and the end date is the second prescription's dispense date plus days supply minus one.
Step 3a: Continue assessing all subsequent dispensing events with allowable gaps for the same drug and adjust end dates as needed. If there is a second dispensing event of the same drug and there is a gap that exceeds the allowable gap, assign an end date for this drug event and begin with the next prescription to again assess if there is 90 days of continuous use. A patient can have multiple start and end dates per drug during the measurement year.
Step 3b: Continue assessing each dispensed prescription for each drug until all dispensing events are exhausted. If a dispensing event goes beyond December 31 of the measurement year, assign the end date as December 31.
Step 4: For each patient, identify if they were dispensed at least 90 consecutive treatment days of antipsychotics during the measurement year.

Table APC-A: Antipsychotic Medications
First-generation antipsychotic medications: Chlorpromazine HCL; Fluphenazine HCL; Fluphenazine decanoate; Fluphenazine enanthate; Haloperidol; Haloperidol decanoate; Haloperidol lactate; Loxapine HCL; Loxapine succinate; Molindone HCL; Perphenazine; Pimozide; Promazine HCL; Thioridazine HCL; Thiothixene; Thiothixene HCL; Trifluoperazine HCL; Triflupromazine HCL
Second-generation antipsychotic medications: Aripiprazole; Asenapine; Clozapine; Iloperidone; Lurasidone; Olanzapine; Olanzapine pamoate; Paliperidone; Paliperidone palmitate; Quetiapine fumarate; Risperidone; Risperidone microspheres; Ziprasidone HCL; Ziprasidone mesylate

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
N/A

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1*

page should be provided in an Excel or csv file in required format with at S.2b)
Report three age stratifications and a total rate:
1–5 years
6–11 years
12–17 years
Total (sum of the age stratifications)

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
N/A

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Determine the eligible population, or the denominator.
Step 1a: Identify patients in the specified age range who were dispensed an antipsychotic medication during the measurement year (January 1 – December 31).
Step 1b: For each patient, identify all antipsychotic prescriptions during the measurement year. For each drug, identify start and end dates of the prescriptions. Starting with the first prescription in the measurement year determine if there is a second dispense date of that same drug. If there is no second dispensing event with the same Drug ID, the start date is the first prescription's dispense date and the end date is the start date plus the days supply minus one. If there is a second dispensing event of the same drug, determine if there are gap days (a 32-day gap is allowed). Calculate the number of days between (but not including) the first prescription's dispense date and the second prescription's dispense date. If the number of days is less than or equal to the first prescription's days supply plus 32 days, the gap is less than or equal to 32 days and is allowed. The start date is the first prescription's dispense date and the end date is the second prescription's dispense date plus days supply minus one.
Step 1c: Continue assessing all subsequent dispensing events with allowable gaps for the same drug and adjust end dates as needed. If there is a second dispensing event of the same drug and there is a gap that exceeds the allowable gap, assign an end date for this drug event and begin with the next prescription to again assess if there is 90 days of continuous use. A patient can have multiple start and end dates per drug during the measurement year.
Step 1d: Continue assessing each dispensed prescription for each drug until all dispensing events are exhausted. If a dispensing event goes beyond December 31 of the measurement year, assign the end date as December 31.
Step 1e: For each patient, identify if they were dispensed at least 90 consecutive treatment days of antipsychotics during the measurement year.

Step 2: Determine the numerator. For each patient in the eligible population, by drug, identify all prescription events, start dates

and end dates. Identify the number of concurrent antipsychotic medication treatment events.
Step 2a: For each patient, identify the first day during the measurement year where the patient was being treated with two or more different antipsychotic medications; this is the concurrent antipsychotic medication treatment event start date.
Step 2b: Beginning with (and including) the start date, identify the number of consecutive days where the patient remains on two or more different antipsychotic medications. If the number of days =90 days, the patient is numerator compliant.
Step 2c: If the number of consecutive days on multiple antipsychotic medications is <90 days, identify the end date and identify the next day during the measurement year where the patient was being treated with two or more different antipsychotic medications. If the number of days between the end date and the next start date is =15 days, include the days in the concurrent antipsychotic medication treatment event (concurrent antipsychotic medication treatment events allow for a 15-day gap).
Step 2d: If the number of days between the end date and the next start date exceeds 15 days, end the event; using the new start date, continue to assess for concurrent antipsychotic medication treatment events.
Step 2e: Continue this process until the number of concurrent antipsychotic medication treatment days is =90 consecutive days (i.e., the patient is numerator compliant) or until the measurement year is exhausted (i.e., no concurrent antipsychotic medication treatment events were identified during the measurement year).

Step 3: Divide the numerator by the denominator to calculate the rate.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is part of the Healthcare Effectiveness Data and Information Set (HEDIS). As part of HEDIS, this measure pulls from administrative claims collected in the course of providing care to health plan members. NCQA collects the HEDIS data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

This measure has also been tested at the state level and could be reported by states if added to a relevant program.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Ambulatory Care : Clinician Office/Clinic, Behavioral Health/Psychiatric : Outpatient
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Multiple_Concurrent_Testing_10-12-15.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 2799 (New Measure)
Measure Title: Use of Multiple Concurrent Antipsychotics in Children and Adolescents
Date of Submission: 10/9/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

This measure was tested using administrative claims data from the following sources.

- State analyses
  - Medicaid Analytic eXtract (MAX)
- Health plan analyses
  - Medicaid health plans from one state
  - Commercial health plans nationwide

For more information about MAX, refer to http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Data-and-Systems/MAX/MAX-General-Information.html.

1.3. What are the dates of the data used in testing? Click here to enter date range

MAX data 2008, Medicaid health plan data for 17 plans 2010, and commercial health plan data for 73 plans 2012.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency

☒ health plan	☒ health plan
☒ other: State; Integrated Delivery System	☒ other: State; Integrated Delivery System

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
As part of the Pediatric Quality Measures Program (PQMP), NCINQ had access to the Medicaid Analytic eXtract (MAX) for conducting state analyses. In addition, NCINQ was able to test this measure in Medicaid health plan data from one large mid-Atlantic state. In order to assess the measure's use for HEDIS, we conducted an additional analysis in commercial data from a large administrative database. Our samples were as follows.

- State analyses
  - 2008 claims data from the MAX for 11 states
- Health plan analyses
  - 2010 claims data from 17 Medicaid health plans from one mid-Atlantic state
  - 2012 claims data from 73 commercial health plans nationwide

These administrative data sources included claims for all of the data elements needed to capture this measure, including claims for health care system encounters, laboratory codes, and pharmacy codes.

For our MAX analysis, the 11 states were chosen on the basis of Mathematica Policy Research reports that suggested that they provided adequate encounter/managed care data (Byrd & Dodd, 2012; Byrd & Dodd, 2013).

Citations
Byrd VLH, Dodd AH. *Assessing the usability of encounter data for enrollees in comprehensive managed care across MAX 2007-2009*. December 2012 2012.

Byrd VLH, Dodd AH. Assessing the Usability of MAX 2008 Encounter Data for Comprehensive Managed Care. *Medicare & Medicaid Research Review*. 2013;3(1).

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
We tested a set of several measures related to antipsychotic use in three datasets described above. Our analyses included enrollees who met continuous enrollment and measure-specific criteria. Our commercial health plan analyses included enrollees age 0-17 years during the measurement year. All other analyses included enrollees ages 0 to 20 during the measurement year. The age ranges varied slightly as our draft concepts were refined and in order to make the measures relevant to states (children/adolescents typically defined as age up to 21 years) and health plans (children/adolescents typically defined as age up to 18 years). We excluded enrollees who were dually eligible for Medicaid and Medicare. In the MAX data, a total of 126,018 children and adolescents met the denominator criteria and were included in the sample for this measure. Across the 17 Medicaid plans, the total number of children and adolescents who met denominator criteria was 13,294, and across 49 commercial plans that had sufficient denominators (>30), the total was 11,895.

Below are descriptions of the patient samples in terms of denominator sizes across the entities measured. They include the mean denominator, minimum denominator, maximum denominator, and the 25th, 50th (or median), and 75th percentiles.

Denominator Size Distribution Across 11 States (MAX) (2008)

Mean	11,456
Minimum	1,545
25th	5,951
Median	10,393
75th	15,569
Maximum	24,161

Denominator Size Distribution Across 17 State Medicaid Health Plans from One State (2010)

Mean	783
Minimum	123
25th	319
Median	680
75th	976
Maximum	2,582

Denominator Size Distribution Across 49* Commerical Health Plans Nationwide (2012)

Mean	243
Minimum	31
25th	92
Median	168
75th	290
Maximum	1,566

* Of the 73 commercial plans included in the testing of this measure, 49 had sufficient denominators (>30)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Reliability of the measure score was tested using a beta-binomial calculation and this analysis included the entire data samples described in the sections above (MAX state data, Medicaid heath plan, commercial health plan).

Validity was demonstrated through a systematic assessment of face validity. Per NQF instructions we have described the composition of the technical expert panels which assessed face validity in the data sample questions above. In addition, validity was demonstrated through two types of analyses: correlations among measures using Spearman Correlation Coefficients (using commercial health plan data sample and Medicaid health plans data sample) and rankings of health plans and states on measures (using MAX state data sample and Medicaid health plan data sample). This analysis is described further in section 2b2.3.

For identifying statistically significant & meaningful differences in performance, all three data samples were used (MAX state data, Medicaid heath plan, commercial health plan).

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We assessed differences across multiple age strata (0-5, 6-11, 12-17, and total [0-17]), race/ethnicity (Hispanic; White, non-Hispanic; Black, non-Hispanic), and foster care status.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability Testing of Performance Measure Score: The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the "signal." The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents "noise." The reliability of the measure is represented as the ratio of signal to noise.

- A score of 0 indicates none of the variation (signal) is attributable to the plan
- A score of 1.0 indicates all of the variation (signal) is attributable to the plan
- A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY

The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan's variation for the average plan's variation. Thus, the reliability for some plans may be more or less than the overall reliability across plans.

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

This measure achieved a reliability score above 0.7 for state-level reliability and about 0.7 for Medicaid health-plan level reliability. This measure achieved a higher level and narrower range of reliability in the state data compared to the health plan data.

	Average Reliability	Minimum Reliability

MAX States	.99	.96
Medicaid Health Plans	.64	.28
Commercial Health Plans	.42	.08

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

As stated in 2a2.2, we estimated reliability with a beta-binomial model (Adams, 2009). A score of zero implies that all the variability in a measure is attributable to measurement error. A score of 1.0 implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one reporting entity from another. A score of 0.7 or higher indicates adequate reliability to distinguish performance between two entities and is considered acceptable. The testing results suggest that this measure has high reliability at the state level.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Face Validity

The health-plan level of this measure was assessed for use in the HEDIS Health Plan Measure Set. As part of this process, NCQA assessed the face validity of the measure using its HEDIS process. NCQA staff shared the measure concepts, supporting evidence and field test results with its standing Behavioral Health Measurement Advisory Panel, Technical Measurement Advisory Panel and additional panels. We posted the measures for Public Comment, a 30-day period of review that allowed interested parties to offer feedback about the measure. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle. This measure has undergone the following steps associated with that cycle.

Step 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA's Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

Step 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a

detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

Step 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQAs Board of Directors will be included in the next HEDIS year and reported as first-year measures.

Step 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

Empirical Validity

As part of field testing, we assessed construct validity, which considers whether measures are capturing important aspects of a quality concept. We conducted two types of analyses: correlations among measures and rankings of health plans and states on measures.

We first tested for construct validity by exploring whether this measure was correlated with other related measures, including the *Follow-up Visit for Children and Adolescents on Antipsychotics* measure. The *Follow-up Visit* measure assesses the percentage of youth who have a follow-up visit with a prescriber within 30 days after the start of a new antipsychotic prescription. We hypothesized that organizations that perform well on one of these measures should perform well on the other measure. We calculated correlations using the Spearman correlation coefficients. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

We then explored whether entities that manage one aspect of antipsychotic prescribing for children and adolescents well, such as avoiding multiple concurrent antipsychotics, also manage other aspects of care well. This test shows if plans and states can be approximately ranked based on profiles of performance across multiple measures. Consistency of performance across measures suggests that the measures are assessing a dimension of quality. For state rankings, we compared the *Multiple Concurrent* measure with the *Use of Antipsychotics in Very Young Children* measure and the *Metabolic Monitoring for Children and Adolescents on Antipsychotics* measure. For the Medicaid health plan rankings we compared the Multiple Concurrent measure with the *Use of Higher than Recommended Doses of Antipsychotics in Children and Adolescents* measure and the *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* measure. The *Antipsychotics in Very Young* measure assesses the percentage of youth under age six who were prescribed antipsychotics (a lower rate indicates better performance). The *Metabolic Monitoring* measure assesses the percentage of youth with ongoing antipsychotic use who had metabolic monitoring. The *High Dose* measure assesses the percentage of youth prescribed a higher-than-recommended dose of an antipsychotic. The

Psychosocial Care measure assesses the percentage of youth who receive first-line psychosocial care when newly prescribed an antipsychotic (among those youth that do not have a primary indication for an antipsychotic).

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face Validity Results

Step 1: This measure was developed to address inappropriate prescribing patterns for children and adolescents on antipsychotics. NCQA and five expert panels worked together in 2013 and 2014 to identify the most appropriate method for assessing the use of multiple concurrent antipsychotics among this patient population. Across the multiple expert panels that reviewed this measure, all panels concluded this measure was specified to assess multiple concurrent use of antipsychotics.

Step 2: The measure was written and field-tested in 2013 and 2014. After reviewing field test results, the CPM recommended to send the measure to public comment with a majority vote in January 2014.

Step 3: The measure was released for Public Comment in 2014 prior to publication in HEDIS. This measure was rated a high priority by many commenters. Of 74 comments received, the majority (65 percent) supported it as-is or with suggested modifications. The CPM recommended moving this measure to first year data collection by a majority vote in May 2014.

Step 4: The measure was introduced in HEDIS 2015. Organizations voluntarily reported this measure in the first year (2014) and the results were analyzed for public reporting in the following year (2015). The measure was approved in September 2015 by the CPM for public reporting in HEDIS 2016 for Medicaid plans.

Empirical Validity Results

Correlations

When determining correlations among measures, we focused on health plans, as there were not enough entities to measure correlations with the state data.

The results showed that among Medicaid health plans, there were no statistically significant correlations between the *Multiple Concurrent* measure and the other measures addressing antipsychotic use in children and adolescents. Among national commercial plans, there was moderate negative correlation between the *Follow-up Visit* and *Multiple Concurrent* measures (r=-.58, p=0.02). In addition to assessing correlations among the measures in this set, we examined correlations between performance on the measures and rates of hospitalization for mental health and substance use problems. However, we did not find consistent correlations.

Ranking

Among MAX states and one state's Medicaid plans, we found good consistency in the states and plans, respectively, with the best and worst performance.

MAX State Performance Rankings

State	Multiple Concurrent Antipsychotics[1]	Antipsychotics in Very Young Children[1]	Metabolic Monitoring[2]
1	5.7	**0.3**	14.2
2	6.6	**0.3**	19.4
3	**9.4**	**0.3**	20.6
4	7.7	0.2	**6.5**

5	3.3	**0.1**	**4.8**
6	**2.9**	**0.3**	18.7
7	**8.1**	0.2	20.0
8	7.1	**0.1**	14.8
9	7.7	**0.0**	**29.1**
10	4.1	**0.1**	19.6
11	**3.0**	**0.1**	**36.2**
Mean	**6.0**	**0.2**	**18.5**

[1]Lower rate indicates better performance
[2]Higher rate indicates better performance

Medicaid Health Plan Performance Rankings for One State

Plan	Multiple Concurrent Antipsychotics[1]	Higher than Recommended Doses[1]	First-Line Psychosocial Care[2]
3	3.8	**11.7**	41.7
9	**7.0**	8.3	48.6
6	**6.6**	4.9	30.1
17	3.3	**9.6**	**26.4**
2	5.1	4.4	**27.4**
8	4.6	5.4	43.5
4	3.3	5.8	46.9
5	3.9	4.9	42.4
1	5.6	5.6	51.6
11	5.1	5.7	43.8
16	3.3	4.0	56.6
15	6.3	5.7	28.0
12	4.3	4.7	43.3
13	4.5	**3.3**	30.7
7	**2.3**	4.6	**67.7**
14	4.6	5.4	64.3
10	**1.8**	**2.7**	**67.0**
Mean	**4.4**	**5.7**	**44.7**

[1]Lower rate indicates better performance
[2]Higher rate indicates better performance

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Face Validity

The expert panels consulted showed good agreement that the measure as specified will accurately differentiate quality across states and health plans. Additionally this measure was rated as a high priority measure by the expert panels and by those who responded to the public comment. Our interpretation of these results is that this measure has sufficient face validity.

Empirical Validity

Correlations

Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. The results confirmed our hypothesis that commercial plans that performed well on providing follow-up visits (higher rates indicate better performance) also performed well on avoiding multiple concurrent prescribing for those on antipsychotics (lower rates indicate better performance).

Ranking

The results show that plans and states can be approximately ranked based on profiles of performance across multiple measures. The consistent performance across measures suggest the measures are assessing a dimension of quality.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified

(*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR). The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Variation in Performance Rates across MAX States (2008 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
6.0	3.0	3.3	6.6	7.7	8.1	4.4

IQR: Interquartile range

Variation in Performance Rates across Medicaid Plans from one State (2010 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
4.4	2.9	3.3	4.5	5.4	6.4	2.1

IQR: Interquartile range

Variation in Performance Rates across Commercial Plans Nationwide (2012 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
3.1	.7	1.4	3.1	4.6	5.1	3.2

IQR: Interquartile range

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The results indicate that there is 2.1% gap in performance between Medicaid plans at the 25th and 75th percentiles, a 3.2% gap in performance among commercial plans and a 4.4% gap in performance among states at the 25th and 75th percentiles. This means states at the 75th percentile have on average 504 more children and adolescents receiving multiple concurrent antipsychotics than states at the 25th percentile.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than***

one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
States and plans collect this measure using all administrative data sources, for all intents and purposes, there are no missing data in administrative data. We have done no assessment to look for the distribution of missing data. For plans reporting on this measure for HEDIS, NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
N/A

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Field testing results, more fully described in the Scientific Acceptability section, showed the measure is feasible to be collected by health plans and states using administrative claims data. Further, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the managed care organization´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity

4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Through our Policy Clarification Support System, NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures, including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation, information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Quality Compass http://www.ncqa.org/tabid/177/Default.aspx http://www.ncqa.org/tabid/836/Default.aspx The State of Health Care Quality Report Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx The State of Health Care Quality Report http://www.ncqa.org/tabid/836/Default.aspx

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

QUALITY COMPASS: This measure has just been approved for use in Quality Compass for Medicaid, a tool that displays health plan-level performance rates for HEDIS measures. It is used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. The Quality Compass 2015 Medicaid tool includes data for 182 public reporting Medicaid health plan products, serving approximately 20 million covered lives. Benchmarks are calculated from a total pool of 244 public and non-public reporting health plan products, serving approximately 25 million covered lives.

THE STATE OF HEALTH CARE QUALITY REPORT: HEDIS measures are reported nationally and by geographic regions in the State of Health Care Quality Report, published by NCQA and summarizing findings on quality of care. In 2015 the report included measures on 15.4 million Medicare Advantage beneficiaries in 507 Medicare Advantage health plans, 103.9 million members in 413 commercial health plans, and 25.4 million Medicaid beneficiaries in 237 plans across 50 states.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No negative consequences have been reported since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are

compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
2337 : Antipsychotic Use in Children Under 5 Years Old

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
This new measure assesses multiple concurrent antipsychotic use children and adolescents who have continuous antipsychotic use. Measure 2337 is Antipsychotic Use in Children under 5 Years Old and assesses whether children under 5 are prescribed an antipsychotic at some point during the measurement year. Both measures are specified for the health plan level and use administrative claims as the data source. Both measures assess antipsychotic use; however, our measure has a broader age range (up to 18 years). In addition, the target population for this new measure is also focused only on those who have continuous use of antipsychotics for 90 days or more. In terms of measure focus, measure 2337 is focused on the utilization of antipsychotics among very young children for 30 days or more. The Use of Multiple Concurrent Antipsychotics in Children and Adolescents measure is focused on the receipt of multiple antipsychotics concurrently for at least 90 days during the measurement year. While both measures are assessing overuse/appropriateness of antipsychotics in children, what is being measured (or considered overuse) is different. While measure 2337 looks at any prescription for antipsychotics, our measure looks for a specific type of non-recommended practice (multiple concurrent use).

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee on Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-
Co.3 **Measure Developer if different from Measure Steward:** National Committee on Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
NCQA Behavioral Health Measurement Advisory Panel
Bruce Bobbitt, PhD, LP, Optum
Peter Delany, PhD, LCSW-C, Substance Abuse and Mental Health Services Administration
Ben Druss, MD, MPH, Emory University
Frank A. Ghinassi, PhD, ABPP, Western Psychiatric Institute and University of Pittsburgh Medical Center
Rick Hermann, MD, Tufts Medical Center and UpToDate, Inc.
Connie Horgan, ScD, Brandeis University
Neil Korsen, MD, Maine Health
Charlotte Mullican, BSW, MPH, Agency for Healthcare Research and Quality
Harold Pincus, MD, Columbia University and RAND Corporation
Bruce L. Rollman, MD, MPH, University of Pittsburgh School of Medicine
Michael Schoenbaum, PhD, National Institute of Mental Health
John H. Straus, MD, Massachusetts Behavioral Health Partnership and Beacon Health Options

NCQA Committee on Performance Measurement (CPM)
Bruce Bagley, MD, American Academy of Family Physicians
Andrew Baskin, MD, Aetna
Patrick Conway, MD, MMSc, Center for Medicare & Medicaid Services
Jonathan D. Darer, MD, Geisinger Health System
Helen Darling, National Business Group on Health
Rebekah Gee, MD, MPH, FACOG, LSU School of Medicine and Public Health
Foster Gesten, MD, NYSDOH Office of Managed Care
David Grossman, MD, MPH, Group Health Physicians
Christine Hunter, MD (Co-Chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Bernadette Loftus, MD, The Permanente Medical Group
J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, RN, PhD, AARP
Eric C. Schneider, MD, MSc (Co-Chair), RAND Corporation
Marcus Thygeson, MD, MPH, Blue Shield of Califorina

NCQA Technical Measurement Advisory Panel
Andy Amster, MSPH, Kaiser Permanente
Kathryn Coltin, MPH, Independent Consultant
Lekisha Daniel-Robinson, Centers for Medicare and Medicaid Services
Marissa Finn, MBA, Cigna HealthCare
Scott Fox, MS, MEd, Independence Blue Cross
Carlos Hernandez, CenCalHealth
Kelly Isom, MA, RN, Aetna
Harmon Jordan, ScD, RTI International
Ernest Moy, MD, MPH, Agency for Healthcare Research and Quality
Patrick Roohan, New York State Department of Health
Lynne Rothney-Kozlak, MPH, Rothney-KozlakConsulting, LLC
Natan Szapiro, Independence Blue Cross

National Collaborative for Innovation in Quality Measurement (NCINQ) Measurement Advisory Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Katie Brookler, Colorado Department of Health Care Policy and Financing
Cathy Caldwell, MPH, Alabama Department of Public Health
Jennifer Havens, MD, NYU School of Medicine
Ted Ganiats, MD, University of California, San Diego
Darcy Gruttadaro, JD, National Allegiance on Mental Illness
Virginia Moyer, MD, MPH, FAAP, Baylor College of Medicine, USPSTF
Edward Schor, MD, Lucile Packard Foundation for Children's Health
Xavier Sevilla, MD, FAAP, Whole Child Pediatrics
Gwen Smith, Illinois Department of Healthcare and Family Services/Health Management Associates
Janet (Jessie) Sullivan, MD, Hudson Health Plan
Kalahn Taylor-Clark, PhD, MPH, George Mason University
Craig Thiele, MD, CareSource
Charles Wibbelsman, MD, Kaiser Permanente Medical Group, Inc.
Jeb Weisman, PhD, Children's Health Fund

NCINQ Consumer Panel
Joan Alker, MPhil, Georgetown Center for Children and Families
Roni Christopher, MEd, OTR/L, PCMH-CCE, The Greater Cincinnati Health Collaborative
Daniel Coury, MD, Nationwide Children's Hospital
Eileen Forlenza, Colorado Medical Home Initiative, Children and Youth with Special Health Care Needs Unit
Michaelle Gady, JD, Families USA
Janis Guerney, JD, Family Voices
Jocelyn Guyer, MPA, Georgetown Center for Children and Families
Catherine Hess, MSW, National Academy for State Health Policy
Carolyn Muller, RN, Montgomery County Health Department
Cindy Pellegrini, March of Dimes
Judith Shaw, EdD, MPH, RN, VCHIP
Stuart Spielman, JD, LLM, Autism Speaks
Michelle Sternthal, PhD, March of Dimes

NCINQ Foster Care Panel
Kamala Allen, MHS, Center for Health Care Strategies
Mary Applegate, MD, Ohio Department of Job and Family Services
Samantha Jo Broderick, Foster Care Alumni of America
Mary Greiner, MD, Cincinnati Children's Hospital Medical Center
David Harmon, MD, FAAP, Superior HealthPlan
Patricia Hunt, Magellan Health Services
Audrey LaFrenier, MSW, Parsons Child and Family Center
Bryan Samuels, MPP, Chapin Hall
Phil Scribano, DO, MSCE, The Children's Hospital of Philadelphia
Lesley Siegel, MD, State of Connecticut Department of Children and Families
Chauncey Strong, MSW, LGSW, Fairfax County Department of Family Services/Foster Care and Adoption
Janet (Jessie) Sullivan, MD, Hudson Health Plan
Nora Wells, MS, National Center for Family/Professional Partnerships

NCINQ Mental Health Panel
Francisca Azocar, PhD, Optum Health Behavioral Solutions
Frank Ghinassi, PhD, Western Psychiatric Institute and Clinic of UPMC Presbyterian Shadyside
Jennifer Havens, MD, NYU Langone Medical Center
Danielle Laraque, MD, FAAP, Maimonides Infants and Children's Hospital of Brooklyn

NCINQ State Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Sharon Carte, MHS, State of West Virginia Children's Health Insurance Program

Susan Castellano, Minnesota Department of Human Services
Catherine Hess, MSW, National Academy for State Health Policy
Michael Hogan, PhD, New York State office of Mental Health
Barbara Lantz, MN, RN, State of Washington Department of Social and Health Services, Medicaid Purchasing Administration
Judy Mohr Peterson, PhD, Oregon Health Authority
Tracy Plouck, MPA, Ohio Department of Mental Health
Gina Robinson, Colorado Department of Health Care Policy and Financing
Janet Stover, Illinois Association of Rehabilitation Facilities
Eric Trupin, PhD, University of Washington

NCINQ Measure Development Partners
Shahla Amin, MS, Rutgers University
Scott Bilder, PhD, Center for Health Services Research, Rutgers University
Stephen Crystal, PhD, Institute for Health, Health Care Policy and Aging Research, Rutgers University
Molly Finnerty, PhD, NY State Office of Mental Health
Emily Leckman-Westin, PhD, NY State Office of Mental Health
Sheree Neese-Todd, MA, Institute for Health, Health Care Policy and Aging Research, Rutgers University

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2014

Ad.3 **Month and Year of most recent revision:**

Ad.4 **What is your frequency for review/update of this measure?** Every 3 years

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2016

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2800
Measure Title: Metabolic Monitoring for Children and Adolescents on Antipsychotics
Measure Steward: National Committee on Quality Assurance
Brief Description of Measure: The percentage of children and adolescents 1–17 years of age who had two or more antipsychotic prescriptions and had metabolic testing.
Developer Rationale: This measure addresses metabolic monitoring as one facet of safe and judicious use of antipsychotics in children and adolescents. Although antipsychotic medications offer the potential for effective treatment of psychiatric disorders in children, they can also increase a child's risk for developing serious metabolic health complications associated with poor cardiometabolic outcomes in adulthood. Despite the risk of such adverse side effects, research suggests that children and adolescents do not receive appropriate laboratory monitoring. Thus, this measure encourages metabolic monitoring of children who are on antipsychotic medications.

Numerator Statement: Children and adolescents who received glucose and cholesterol tests during the measurement year.
Denominator Statement: Children and adolescents who had ongoing use of antipsychotic medication (at least two prescriptions).
Denominator Exclusions: No exclusions

Measure Type: Process
Data Source: Administrative claims
Level of Analysis: Health Plan, Integrated Delivery System, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this process measure:

- The developer provides the following relationship between the process being measured and outcome: Child or adolescent has ongoing use of antipsychotic medication >>> Metabolic monitoring by a health care provider >>> Identification of metabolic issues/side effects >>> Health care provider addresses metabolic issue by, for example, adjusting antipsychotic medication regimen >>> Patient receives intervention for metabolic issues present >>> Metabolic issues reduced or eliminated >>> Improvement in metabolic functioning for patient (desired outcome).
- The measure is based on 11 evidence-based clinical practice guidelines and standards from five organizations,

including the American Academy of Child and Adolescent Psychiatry (AACAP), Canadian Alliance for Monitoring Effectiveness and Safety of Antipsychotics in Children (CAMESA), and others.

- The developer indicates that guidelines are all graded very strongly.
- The developer focuses on the guidelines from AACAP as it is most relevant to the measure. It is based on an evidence review of 147 published clinical trials, meta-analysis, practice guidelines, randomized controlled trials (RCTs), systematic literature reviews, and case reports and series. The measure is based on three recommendations:
    - **Recommendation 10 (and Table 2):** "The acute and long-term safety of these medications in children and adolescents has not been fully evaluated and therefore careful and frequent monitoring of side effects should be performed…*Ideally, monitoring of BMI, blood pressure, fasting glucose and fasting lipid profiles should follow, whenever feasible, the recommendations found in the consensus statement put forth by the American Diabetes Association and American Psychiatric Association."* This recommendation is based on expert opinion established during a consensus development conference for four medical professional societies.
    - **Recommendation 12:** "Careful attention should be given to the increased risk of developing diabetes with the use of AAA, and blood glucose and other parameters should be assessed at baseline and monitored at regular intervals." This recommendation is based on previous studies on various populations, including adults, focused on the association between diabetes/abnormal glucose regulation and the use of antipsychotics. It is also based on expert opinion, literature reviews, case series, an observational study, and a randomized clinical trial; double-blind, randomized, placebo-controlled study.
    - **Recommendation 13:** "In those patients with significant weight changes and/or a family history indicating high risk, lipid profiles should be obtained at baseline and monitored at regular intervals." This recommendation is based on a review of seven national, cross-sectional studies conducted between 1973 and 1994 that focused on the association between elevated lipid levels and the development of cardiovascular disease throughout the lifespan.
- The developer states that "Overall, the quality of the evidence regarding metabolic monitoring for children and adolescents on antipsychotics is high. The evidence provides a strong link between antipsychotic use and adverse metabolic side effects in youth and to negative long-term health outcomes throughout the lifespan."
- No studies that evaluate the benefit of monitoring are cited, however, there is evidence regarding the adverse side effects and long-term consequences. Harms noted included limitations on time due to increased medical appointments. Thus, in the absence of evidence of other harms, AACAP estimated there is less harm through increased vigilance and regular metabolic monitoring.

Questions for the Committee

- *Is the relationship between the measure to patient outcomes clear and reasonable?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer reports the following:

- A review of the literature identified one study of Medicaid-enrolled children in three states that found that only 31 percent of youth starting an atypical antipsychotic received a glucose test, and only 14 percent received a lipid test—far lower than rates reported for adults (Morrato et al., 2010). A second study found the association of atypical antipsychotics with diabetes is greater among children and adolescents than adults (Hammerman et al., 2008).
- During field testing, the developer used MAX data from 2008 for 11 states. It found the percentage of children receiving metabolic screening within 30 days of a new antipsychotic medication prescription was 6.0 percent, with a range of 0.4 percent to 14.0 percent. For children and adolescents who had ongoing antipsychotic use,

the percentage that received metabolic monitoring was on average 18.5 percent, with a range of 4.8 percent to 36.2 percent.

- In an examination of claims data from 17 Medicaid health plans in one state, the developer found that the average percentage of children receiving baseline metabolic screening within 30 days of a new antipsychotic medication prescription among the general population of children in health plans was 10.3 percent, with a range of 0.2 to 17.8 percent. For ongoing metabolic monitoring during the measurement year, the data suggest similar gaps in care. The percentage of children with ongoing antipsychotic use receiving metabolic monitoring during the measurement year was 30.9 percent, with a range of 2.3 to 40.0 percent.
- Performance rates for the measure were constructed and tested for three age strata (0-5, 6-11, 12-17), race/ethnicity and foster care status. Of the 11 states, eight had higher rates of metabolic monitoring for the foster care population compared with the general Medicaid population. In both the general population and foster care population, metabolic monitoring was highest among adolescents compared with the lower age strata. In both the general and foster care populations, monitoring was higher for Hispanic children and adolescents than for black or white children and adolescents.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- What evidence is there for including newborns and infants in this measure? Is there any aspect of the new lipid guidelines issued by NHLBI that would warrant a change to this measure? Are the diabetes screening tests supported by recommendations from ADA?
- Good evidence.
- The measure addresses a process (monitoring using blood tests for metabolic conditions for children taking these medications. The evidence has led several professional societies to include this as a strong recommendation in their guidelines. However, there are two major issues.
    - 1. The evidence is all for the link between the medications and the metabolic derangement. The guideline recommendations are quite strong, but are based on "Expert Opinion" even the AACAP recommendations say "safety has not been fully evaluated"-- which is not direct evidence of harm (though monitoring may still be indicated). Rec 13 limits lipid testing to those with "significant weight changes and/or a family history.." The developers reference a significant body of evidence (used by the guideline group) that these drugs can be associated with metabolic derangements. They reference everything from case reports to RCT's. The developers write that they "did not feel comfortable" conducting a literature review themselves, but it would be good to have the results of an SR or at least some of the large studies to understand the incidence of metabolic derangement.
    - 2. All of the evidence is not direct evidence between the measure and outcome. The logic is right-- if these medications cause metabolic derangements, presumably the long term cardiovascular harm would be the same as in general populations. So, this is credible, but quite indirect evidence between the measure and long term outcomes.
- This will come down to whether there is sufficient evidence that metabolic derangements themselves are an "outcome". If so, it could be "moderate" (though I would like to see results of major studies). But, using Algorithm 1 strictly, I would rate this as "Insufficient Evidence with Exception."
- Process measure - relationship between testing for metabolic conditions and increased identification of metabolic conditions seems clear. Recommendation strongly endorsed by multiple national organizations and based on a large number of studies. The evidence cited directly addresses the proposed measure.
- The evidence suggests a strong link between antipsychotic use and adverse metabolic side effects in youth. Overall quality of evidence is high. Screening for metabolic side effects which leads to adjustments/intervention which then leads to improved metabolic functioning (outcome).

- The evidence provided directly applies to the process of care in the measure. The evidence is noted to be strong and I agree. The evidence s based largely on clinical practice guidelines (11) and standards of care related organizations (5). Yes the measured process if related to the healthcare action (goal of the measure.) The rationale strongly supports the measure process.

1b. Performance Gap.

- Overall less than optimal performance. Disparities demonstrated.
- Gap exists.
- A performance gap is demonstrated both by the low proportion tested in even the highest performing plans/state and by the very wide variability in testing rates seen at the state level (using Medicaid MAX data) and among Medicaid health plans, and commercial plans. The MAX data are from 2008 which is a bit problematic, but the other analyses are more recent (2010 and 2012). It should be noted that the denominators of eligible patients within these plans varies widely. In the commercial health plans, 25% of the plans had 100 patients or less eligible for the measure. This may contribute to the variability observed.
- Disparities are demonstrated by race/ethnicity and by age (adolescents more likely to be tested).
- A performance gap is suggested not only through cited studies, but also through the developer's analysis of MAX data from 11 states. The developers' analysis indicates higher rates of testing for adolescents compared to younger children, children in foster care as opposed to Medicaid and Hispanic children as compared to other groups. It appears to be a disparities sensitive measure.
- Evidence presented suggests that there is a stronger association between the use of atypical antipsychotics and diabetes among children and youth than in adults.
- Testing for metabolic issues in children that are on antipsychotics is low, suggesting a need for a national performance measure.
- Yes, performance data was provided including an 11 state review of targeted children for baseline metabolic (glucose and lipids) - 6.0% average, and with ongoing drug use -18.5 average. This is another data indicate a gap to be addressed by endorsement of this measure. Of not also is that the group with the highest performance of this monitoring (%age) was adolescents where younger children would seem to be more at risk for long-term negative effects of unmonitored use of multiple antipsychotic drugs. To me, there seems a clear need for improvement.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The numerator is children and adolescents who received glucose and cholesterol tests during the measurement year.
- The denominator is children and adolescents who had ongoing use of antipsychotic medication (at least two prescriptions).
- The numerator and denominator details spell out the tests and medications that should be included. Codes needed to calculate the measure (ie, CPT codes) are included in an Excel spreadsheet provided as an appendix.
- The algorithm logic is straightforward.
- The measure is stratified by age but is not risk adjusted.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Are the list of medications and tests appropriate?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer reports:

- The measure was tested with 2008 MAX data from 11 states, 2010 claims data from 17 Medicaid health plans from one state, and 2012 claims data from 73 commercial plans nationwide.
- This measure was tested at the performance measures score level using a beta-binomial signal-to-noise analysis. For this type of testing, a score of zero implies that all the variability in a measure is attributable to measurement error. A score of 1.0 implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one reporting entity from another. A score of 0.7 or higher indicates adequate reliability to distinguish performance between two entities and is considered acceptable.
    - The average reliability for states and plans was above 0.7 (ranging from 0.99 to 0.83), suggesting the measure is reliable, particularly at the Medicaid health plans and state levels.
- Per the NQF algorithm, reliability testing at the computed performance measure score may be rated HIGH, MODERATE, or LOW depending on the testing results.

Questions for the Committee

- *Does the Committee concur with the developer's conclusion that the results demonstrate sufficient reliability so that differences in performance can be identified?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The specifications are consistent with the evidence. The goal of the measure is to encourage metabolic monitoring of children who are on antipsychotic medications in order to reduce the risk of serious metabolic health complications. The evidence provided supports the specifications.

Question for the Committee

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- The developer states the measure was tested at the performance measure score level using both empirical testing and face validity.
- For the empirical testing, the developer assessed construct validity with two types of analyses: correlations among measures and rankings of health plans and states on measures on the three antipsychotic medication measures. The developer reports the following results:
    - Correlations were tested using health plans, as there was not enough entities to test between states. Among national commercial plans, there was a very slight positive correlation between the *First-line Psychosocial Care* and *Metabolic Monitoring* measures (r=0.12, p=.70) and high positive correlation between the *Metabolic Screening* and *Metabolic Monitoring* measures (r=0.82, p<0.0001).
    - Among Medicaid plans in one state, there was a slight positive correlation between the *Follow-up Visit* and *Metabolic Monitoring* measures (r=0.14, p=.58) and high positive correlation between the *Metabolic Screening* and *Metabolic Monitoring* measures (r=0.72, p<0.001).
    - Among MAX states and one state's Medicaid plans, the developer found good consistency in the states and plans, respectively, with the best and worst performance.
- Per the NQF algorithm, validity testing at the computed performance measure score may be rated HIGH, MODERATE,

or LOW depending on the testing results.

- The developer used its standardized HEDIS process to test face validity, but does not explicitly call out face validity of the computed performance score, as required by NQF.
    - The developer worked with five expert panels to identify the most appropriate method for assessing the use of multiple concurrent antipsychotics among this patient population. All of the panels concluded this measure was specified to assess multiple concurrent use of antipsychotics.
    - The draft measure was put out for public comment and brought to the developer's Committee on Performance Measurement.
    - The developer states that the measure has sufficient face validity.

Questions for the Committee

- *Do the results of the empiric testing demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you believe that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- There are no exclusions.

Questions for the Committee

- *Should there be any exclusions for this measure?*
- *Does the Committee believe there are other threats to validity?*

2b4. Risk adjustment:

- This measure is not risk adjusted.

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer states that the results indicate that there is 6.7% gap in performance between Medicaid plans at the 25th and 75th percentiles, a 6.4% gap in performance among commercial plans and a 6.4% gap in performance among states at the 25th and 75th percentiles.

Question for the Committee

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- This is not needed.

2b7. Missing Data

- The measure is collected using all administrative data sources. According to the developer there are no missing data from admin data, so this is not applicable.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- What considerations were given to children/adolescents who were previously identified as already having type 1 or 2 diabetes. Should they be excluded from the denominator?
- Data elements appear clearly defined. The codes used for glycemic control and lipid testing are well-specified and appear complete.
- The list of included medications also appears complete, but must be reviewed by someone with content expertise (e.g. a pharmacist), unless we trust that the NCQA process did this well. The medications to be included may change over time.
- Given that the data are based on claims, it is likely that this could be consistently implemented. One issue is that the

developers do not provide evidence that these tests are always billed to health plans as distinct claims. The developers should address the potential situation of these tests done during an inpatient stay or for some other reason do not get billed as a separate procedure.

- Data elements are clearly defined - suggested testing codes are appropriate and inclusive. This measure seems that it can be consistently implemented.
- The measure specifications are consistent with the evidence.
- The data elements are clearly defined. All required related information is included in the document.
- It is my view that this measure can be consistently implemented although there are large gap from current (as reported) and the goal outcome level of meeting this measure. It was noted that this measure is already in use by the Quality Compass as the means to help selection of a health plan...so this measure can work.

2a2. Reliability testing

- Reliability was tested with the MAX data set (11 states), 17 Medicaid health plans within one state, and a sample of commercial plans. The method for reliability testing is the beta-binomial-signal to noise method, which is appropriate. However, this says nothing about reliability at practice or physician levels. The reliability is acceptable in the very large state-level analysis, and the Medicaid plans (that have larger sample sizes) but was not acceptable (minimum reliability .35) in some of the commercial plans. This highlights the need for using this measure only in settings with sufficient samples of children meeting the denominator criteria. By the algorithm I would rate the reliability as Moderate.
- I concur with the developer's conclusion that the results demonstrate good reliability such that differences in performance can be identified.
- It is my impression that the reliability testing was adequate. It included a large group - 2008 MAX data of 11 states, 17 Medicare health plans from 2010 and 2012 data from 73 commercial plans, nationwide. Reliability was tested by Beta-binomial signal to noise study. Reliability across the three groups ranged from .83 to .99 (with highest score 1.0) which was considered High. I believe this was at the data element level, but not sure.

2b1. Validity Specifications

- Yes, no concerns.
- The specifications are consistent with the evidence (given its limitations as described above).
- The measure specifications are consistent with the evidence.
- The testing suggests that this measure is valid for assessing the compliance with metabolic monitoring.
- I don't see inconsistencies between the specifications and the evidence.
- If this measure is endorsed it would set a goal to have metabolic monitoring at baseline, 12 months and then annually for glucose. It is noted that patients/families would have to have more trips to the doctor/labs which may not be well received. However, the risks associated with drug induced metabolic abnormalities should of higher concern to the target population. Drug induced obesity is surely to be avoided if possible.

2b2. Validity Testing

- Face validity is well demonstrated. The process of review by multiple expert panels adds to the face validity. Empirical evidence for validity is more limited. There is very low correlation with other measures that would reflect quality of care for the same patients (follow-up measure, psychosocial care). There was high correlation only between metabolic screening and metabolic monitoring which are so similar one would expect them to be highly correlated.
- The empirical validity testing would probably be rated as Low (or perhaps Moderate). But given the strong face validity, a score of Moderate seems most appropriate.
- The developers assessed construct validity and demonstrate good consistency with other established measures. Also used a standardized process to demonstrate face validity. The score from this measure appears to be an indicator of quality.
- Empirical tests showed only slight correlations between metabolic monitoring and psychosocial care and follow-up visits.
- Did not provide a computed performance score for face validity. Public comment was overwhelmingly in support of the measure.

- I believe the same data was used for the validity testing. It is also noted, based on expert panel consensus, that face validity is sufficient.
- It appears there was a high correlation at the health plan level between Metabolic Screening and the Metabolic measures, though my notes are not complete on this.

2b3-2b7. Threats to Validity

- What evidence is there to demonstrate that a change in antipsychotics as a result of metabolic tests is less harmful (from a behavioral health perspective) to the affected patients than remaining on the antipsychotic that is causing the issues.
- The developers statement that there are no missing data requires more explanation. Claims data are complete, in that everything submitted separately for payment is present. But, there are events that are not. In the current case, the developers must assume that every blood test sent is billed appropriately and paid by the insurer. Institutions do not always appropriately bill 100% (though I expect the rate to be high). There may be errors of omission, as well as tests sent during inpatient care (or other bundled services) that do not result in claims. The developers should address whether any of these are concerns.
- Developers state this is data acquired from administrative sources and therefore there are no missing data.
- Results show that health plans can be ranked based on their performance across multiple measures, suggesting this can measure quality in the area of metabolic monitoring.
- Results show those plans that score high on initial screenings also perform well on ongoing metabolic monitoring.
- This measure appears sensitive enough to show performance gaps between plans with regards to metabolic monitoring.
- Exclusions - none included but I question whether patients with known DM or glucose abnormalities as well as lipid disorders should be excluded.
- This, I believe, may be a consideration in Risk Adjustment.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer notes:

- These elements are all generated through normal process of care, and are in defined fields in electronic claims.
- The measure is a part of HEDIS, which has a standardized collection and calculation process, as well as a system to collect real-time feedback from measure users.
- Field testing results showed the measure is feasible to be collected by health plans and states using administrative claims data.
- As part of HEDIS, the data elements are subject to that program's data collection and audit requirements.
- This is not an eMeasure.

Questions for the Committee

- *Are the required data elements routinely generated and used during care delivery?*
- *Does the testing data collection strategy indicate the measure is ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- No concerns.
- Measure appears usable.
- This measure should be feasible to collect and report given current insurance claims systems, and is currently in use.
- All data elements are acquired and stored in the routine course of delivery. The measure appears ready to enter into

operational use.

- Feasible because it relies on administrative claims data. The field testing also suggests that the collection of this data is feasible.
- It appears that the required data elements are routinely generated. They are currently available in HEDIS measures through administrative health claims.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- The measure is currently in use in for both quality improvement with benchmarking and public reporting.
- It is included in Quality Compass for Medicaid 2015, a tool that displays health plan-level performance rates for HEDIS measures. It is used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance.
- The measure also is reported on in The State of Health Care Quality Report, a national report produced by the developer including the results from HEDIS measures.
- This is a new measure and improvement results are not yet available.
- No unintended consequences have been reported thus far.

Question for the Committee

- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- If there are data from use of this measure for HEDIS 2015, can the developer share the analysis and lessons learned.
- By measuring rates of performance for metabolic monitoring, more providers are likely to engage in metabolic screening and monitoring of these children and adolescents.

Criterion 5: Related and Competing Measures

- This measure is related to two other measures, 1932 : Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD) and 2337 : Antipsychotic Use in Children Under 5 Years Old. This measure has a different target population and focus.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** N/A

Measure Title: Metabolic Monitoring for Children and Adolescents on Antipsychotics

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 10/9/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Metabolic monitoring provided for ongoing use of antipsychotic medication

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

N/A

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

This measure assesses metabolic monitoring (i.e., the receipt of glucose and cholesterol tests) among children and adolescents that have ongoing antipsychotic use. Given the documented metabolic risks of antipsychotic medications, monitoring of metabolic indices is important to ensure appropriate management of side effect risk, especially in youth. The path envisioned is as follows.

Child or adolescent has ongoing use of antipsychotic medication >>> Metabolic monitoring by a health care provider >>> Identification of metabolic issues/side effects >>> Health care provider addresses metabolic issue by, for example, adjusting antipsychotic medication regimen >>> Patient receives intervention for metabolic issues present >>> Metabolic issues reduced or eliminated >>> Improvement in metabolic functioning for patient (desired outcome).

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

- American Academy of Child and Adolescent Psychiatry. Practice Parameter for the Use of Atypical Antipsychotic Medications in Children and Adolescents. http://www.aacap.org/App_Themes/AACAP/docs/practice_parameters/Atypical_Antipsychotic_Medications_Web.pdf. (July 12, 2012)
- American Academy of Child and Adolescent Psychiatry. July 2001. Practice parameter for the assessment and treatment of children and adolescents with schizophrenia. *Journal of the American Academy of Child and Adolescent Psychiatry.* 40(7 Suppl):4S-23S.2.
- McClellan, J., R. Kowatch, R.L. Findling. January 2007. Practice parameter for the assessment and treatment of children and adolescents with bipolar disorder. *Journal of the American Academy of Child and Adolescent Psychiatry.* 46(1):107–25.
- Pringsheim, T., C. Panagiotopoulos, J. Davidson, J. Ho. August 2011. Evidence-based recommendations for monitoring safety of second generation antipsychotics in children and youth. *Journal of the American Academy of Child and Adolescent Psychiatry.* 20(3):218–33.
- Gleason, M.M., H.L. Egger, G.J. Emslie, et al. December 2007. Psychopharmacological treatment for very young children: contexts and guidelines. *Journal of the American Academy of Child and Adolescent Psychiatry.* 46(12):1532–72.
- Scotto Rosato N., C.U. Correll, E. Pappadopulos, A. Chait, S. Crystal, P.S. Jensen. June 2012. Treatment of maladaptive aggression in youth: CERT guidelines II. Treatments and ongoing management. *Pediatrics.* 129(6):e1577–86.
- Texas Department of Family and Protective Services and University of Texas at Austin College of Pharmacy. 2013. Psychotropic Medication Utilization Parameters for Foster Children. http://www.dfps.state.tx.us/documents/Child_Protection/pdf/TxFosterCareParameters-September2013.pdf (October 22, 2013)

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Organization (Date)	Recommendation	Type/Grade
AACAP-AAA (2011) Practice parameter for the use of atypical antipsychotic medications in children and adolescents.	"The acute and long-term safety of these medications in children and adolescents has not been fully evaluated and therefore careful and frequent monitoring of side effects should be performed…*Ideally, monitoring of BMI, blood pressure, fasting glucose and fasting lipid profiles should follow, whenever feasible, the recommendations found in*	Clinical Guideline

Organization (Date)	Recommendation	Type/Grade
	the consensus statement put forth by the American Diabetes Association and American Psychiatric Association." Table: Fasting plasma glucose—Baseline, 12 wks, annually; Fasting lipid profile—Baseline, 12 wks (Recommendation 10, and Table 2)	
	"Careful attention should be given to the increased risk of developing diabetes with the use of AAA, and blood glucose and other parameters should be assessed at baseline and monitored at regular intervals."(Recommendation 12)	Clinical Standard
	"In those patients with significant weight changes and/or a family history indicating high risk, lipid profiles should be obtained at baseline and monitored at regular intervals."(Recommendation 13)	Clinical Guideline
AACAP-BP (2007) Practice parameter for the assessment and treatment of children and adolescents with bipolar disorder.	"Psychopharmacological interventions require baseline and follow-up symptom, side effect, and laboratory monitoring as indicated….*The American Diabetes Association's recommendations for managing weight gain for patients taking antipsychotics should be followed. This includes baseline BMI, waist circumference, blood pressure, fasting glucose, and a fasting lipid panel. The BMI should be followed monthly for 3 months and then quarterly. Blood pressure, fasting glucose and lipids should be followed up after 3 months then yearly*." (Recommendation 8)	Minimal Standard
AACAP-SZ (2001) Practice parameter for the assessment and treatment of children and adolescents with schizophrenia.	"The use of antipsychotic agents requires….. documentation any required baseline and follow-up laboratory monitoring..."	Minimal Standard
CAMESA (2011) Canadian Alliance for Monitoring Effectiveness and Safety of Antipsychotics in Children—Evidence-based recommendations for monitoring safety of second generation antipsychotics in children and youth.	The guideline provides antipsychotic medication-specific recommendations for monitoring physical examination maneuvers (height, weight, BMI, waist circumference, blood pressure, and neurological examination for extrapyramidal symptoms), and laboratory tests (glucose, insulin, lipid profile tests, AST, ALT, prolactin, and TSH) for children on AAAs. The GRADE rating system is used to rate each test, for each medication, at each time point examined (baseline, 3, 6, and 12 months). *Summary recommendation:* All children prescribed AAAs should be monitored for metabolic side effects at baseline, 3, 6, and 12 months with the following tests: fasting glucose, fasting insulin, and fasting lipid profile (total cholesterol, LDL, HDL, TG). *(Note: Fasting insulin is not recommended for youth on aripiprazole, but is appropriate for all other AAAs.)*	Ranges from 1A (strong) to not recommended depending on the specific medication, laboratory test and timeframe. Strongest evidence and recommendations are for baseline tests.

Organization (Date)	Recommendation	Type/Grade
	A baseline fasting glucose is recommended for all children and adolescents on AAAs (strong recommendation/low quality evidence all AAAs except Ziprasidone, weak recommendation/ consensus based).	1C (all AAA except Ziprasidone) 3 (Zip=3)
	A baseline fasting lipid profile is recommended for all children and adolescents on AAAs (strong recommendation with high to low evidence depending upon the AAA, except Ziprasidone, weak recommendation/consensus based).	1A-1C (all AAAs except Ziprasidone) 3 (Zip=3)
	A follow-up fasting glucose and fasting lipid panel (one or more of the tests within the panel) is strongly recommended for all children at one or more time points during the year. (strong recommendation/high-moderate-low evidence for all AAAs, except Ziprasidone, weak recommendation/consensus based).	1A-1C (all AAAs except Ziprasidone) 3 (Zip=3)
PPWG (2007) The AACAP-sponsored Preschool Psychopharmacology Working Group—Psychopharmacological treatment for very young children: Contexts and guidelines.	"Use of AAA should follow the AACAP practice parameter on AAAs. This practice parameter describes the minimum standards for monitoring vital signs, BMI, fasting blood glucose, extrapyramidal symptoms, lipid profiles, and electrocardiography." (Disruptive Behaviors Algorithm, Stage 2: Pharmacological Intervention).	Not specified
T-MAY (2012) Center for Education and Research on Mental Health Therapeutics—Treatment of maladaptive aggression in youth.	Practitioners should conduct appropriate, guideline-based laboratory monitoring.	Evidence: A, Recommendation: Very strong
TX (2010) Texas Department of Family and Protective Services—Psychotropic medication utilization parameters for foster children.	Practitioners should document appropriate monitoring of laboratory findings.	Not specified*

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

Guideline Developer	Definition
AACAP	*Minimal Standard/ Clinical Standard:* Rigorous/substantial empirical evidence (meta-analyses, systematic reviews, RCTs) and/or overwhelming clinical consensus; expected to apply more than 95 percent of the time
	Clinical guidelines: Strong empirical evidence (nonrandomized controlled trials, cohort or case-control studies), and/or strong clinical consensus; expect to apply in

Guideline Developer	Definition
	most cases (75% of the time)
CAMESA **TMAY Ratings**	**GRADE** *1A:* Strong recommendation, High-quality evidence Oxford Centre for Evidence-Based Medicine grade of evidence (A-D) A: Consistent level 1 studies *Strength of Recommendation:* Very strong (≥90% agreement)

*TX (2010) did not specify the use of a rating system.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Guideline Developer	Definition
AACAP	*Options:* Acceptable but not required; there may be insufficient evidence to support higher recommendation (uncontrolled trials, case/ series reports).
	Not endorsed: Ineffective or contraindicated.
AACAP endorsed best-practice principles	Best-practice principles that underlie medication prescribing, to promote the appropriate and safe use of psychotropic medications
CAMESA	**GRADE**
	1B: Strong recommendation, Moderate-quality evidence
	1C: Strong recommendation/ Low-quality evidence
	2A: Weak recommendation, High- or moderate-quality evidence
	2B: Weak recommendation, Low-quality evidence
	3: Weak recommendation, No evidence, consensus based
PPWG	*A:* Well controlled RCTs, large meta-analyses, or overwhelming clinical consensus
	B: Empirical evidence (open trials, case series) or strong clinical consensus
	C: Single case reports or no published reports, recommendation developed by expert consensus (informal)
TMAY Ratings	B: Consistent level 2 or 3 studies or extrapolations from level 1 studies
	C: Level 4 studies or extrapolations from level 2 or 3 studies
	D: Level 5 evidence or troublingly inconsistent or inconclusive studies of any level
	Strength of Recommendation: Strong (70-89% agreement)
	Strength of Recommendation: Fair (50-69% agreement)
	Strength of Recommendation: Weak (<50% agreement)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

- Andrews, J.C., H.J. Schunemann, A.D. Oxman, et al. July 2013. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *Journal of Clinical Epidemiology.* 66(7):726–35.
- Guyatt, G.H., A.D. Oxman, G.E. Vist, et al. April 26, 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal.* 336(7650):924–6.
- OCEBM Levels of Evidence Working Group. 2011. The Oxford 2011 levels of evidence. 2011; http://www.cebm.net/index.aspx?o=5653 (October 12, 2013)

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

N/A

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

This measure addresses metabolic monitoring as one facet of safe and judicious use of antipsychotics in children and adolescents. Given the documented metabolic risks of antipsychotic medications, monitoring of metabolic indices is important to ensure appropriate management of side effect risk, especially in youth. Numerous guidelines address the need for metabolic monitoring among youth on antipsychotic medications (See section 1a.4.2.). This measure is based on guidelines and evidence from the American Academy of Child and Adolescent Psychiatry (AACAP), Canadian Alliance for Monitoring Effectiveness and Safety of Antipsychotics in Children (CAMESA), and others. These organizations recommend metabolic testing for youth prescribed antipsychotics, with consensus that baseline and ongoing metabolic monitoring are standards of care for this population. While we list the full range of guidelines in sections 1a.4.2 and 1a.4.3 above, we focus on and describe in more detail the AACAP Guideline in the remaining sections, as it is most closely relevant to the specified measure.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The grade assigned by AACAP to the level of evidence varied by the guideline recommendation. The level of evidence varied from Clinical Guideline to Clinical Standard. See table under 1a.4.2 for the level of evidence grade given to each guideline. See table under 1a.4.3 for the definition of the level of evidence grade given to each guideline.

AACAP Strength of Empirical Evidence

AACAP rates the strength of the empirical evidence in descending order as follows:

- (rct) Randomized, controlled trial is applied to studies in which subjects are randomly assigned to two or more treatment conditions
- (ct) Controlled trial is applied to studies in which subjects are non-randomly assigned to two or more treatment conditions
- (ut) Uncontrolled trial is applied to studies in which subjects are assigned to one treatment condition
- (cs) Case series/report is applied to a case series or a case report

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See table under 1a.4.4 for the definition of the level of evidence grade not given to the guidelines.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1990-2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

The guidelines listed in our table above address metabolic monitoring in the context of antipsychotic prescribing for children. The AACAP-AAA guideline is rated a "Clinical Guideline," indicating it is based on strong empirical evidence and/or strong clinical consensus for most cases. AACAP includes several condition-specific guidelines around metabolic monitoring; we focus on the general AACAP-AAA antipsychotics guideline here and describe the body of evidence for each relevant recommendation below.

When developing their guidelines, AACAP limited its evidence review to clinical trials, meta-analysis, practice guidelines, randomized controlled trials (RCTs), systematic literature reviews, and case reports and series. AACAP selected a total of 147 publications for careful examination based on their weight in the hierarchy of evidence attending to the quality of individual studies, relevance to clinical practice and the strength of the entire body of evidence. However, AACAP did not provide a breakdown of specific numbers of each publication type. Given the number of studies selected we did not feel comfortable re-conducting the evidence review and delineating all the publication types for each guideline. Instead we have identified where there are certain publication types available to support each guideline.

Recommendation 10 (and Table 2): "The acute and long-term safety of these medications in children and adolescents has not been fully evaluated and therefore careful and frequent monitoring of side effects should be performed…*Ideally, monitoring of BMI, blood pressure, fasting glucose and fasting lipid profiles should follow, whenever feasible, the recommendations found in the consensus statement put forth by the American Diabetes Association and American Psychiatric Association."* Table 2: Fasting plasma glucose—Baseline, 12 weeks, annually; Fasting lipid profile—Baseline, 12 weeks.

This recommendation is based on expert opinion established during a consensus development conference for four medical professional societies. The four societies found that an increasing number of methodologically rigorous studies have assessed the effectiveness of antipsychotics for children and adolescents in specific clinical situations. However, the long-term safety profile of each antipsychotic used by youth has yet to be effectively evaluated and characterized. In the absence of such evidence, AACAP recommends increased vigilance.

- American Diabetes Association, American Psychiatric Association, American Association of Clinical Endocrinologists, North American Association for the Study of Obesity. Consensus development conference on antipsychotic drugs and obesity and diabetes. *Diabetes Care.* 2004;27:596-601.

Recommendation 12: "Careful attention should be given to the increased risk of developing diabetes with the use of AAA, and blood glucose and other parameters should be assessed at baseline and monitored at regular intervals."

This recommendation is based on previous studies on various populations, including adults, focused on the association between diabetes/abnormal glucose regulation and the use of antipsychotics.

- Expert opinion from four medical professional societies
  - American Diabetes Association, American Psychiatric Association, American Association of Clinical Endocrinologists, North American Association for the Study of Obesity. Consensus development conference on antipsychotic drugs and obesity and diabetes. *Diabetes Care.* 2004;27:596-601.
- Literature reviews, including case reports, case series, observational analytic epidemiologic studies, uncontrolled observations, large retrospective database analyses, and controlled experimental studies, such as randomized clinical trials
  - Casey DE, Haupt DW, Newcomer JW, et al. Antipsychotic-induced weight gain and metabolic abnormalities: implications for increased mortality in patients with schizophrenia. *J Clin Psychiatry.* 2004;65[suppl 7]:4-18.
  - Newcomer JW. Second-generation (atypical) antipsychotics and metabolic effects: a comprehensive literature review. *CNS Drugs.* 2005;19:1-93.
- Case series, including five cases
  - Bloch Y, Vardi O, Mendlovic S, Levkovitz Y, Gothelf D, Ratzoni G. Hyperglycemia from olanzapine treatment in adolescents. *J Child Adolesc Psychopharmacol.* 2003;13:97-102.
- Observational study

    - Hedenmalm K, Hagg S, Stahl M, Mortimer O, Spigset O. Glucose intolerance with atypical antipsychotics. *Drug Saf.* 2002;25:1107-1116.
- Randomized clinical trial; double-blind, randomized, placebo-controlled study
    - Henderson DC, Copeland PM, Daley TB, et al. A double-blind placebo-controlled trial of sibutramine for olanzapine associated weight gain. *Am J Psychiatry.* 2005;162:954-962.

Recommendation 13: "In those patients with significant weight changes and/or a family history indicating high risk, lipid profiles should be obtained at baseline and monitored at regular intervals."

This recommendation is based on a review of seven national, cross-sectional studies conducted between 1973 and 1994 that focused on the association between elevated lipid levels and the development of cardiovascular disease throughout the lifespan.

- Freedman DS, Dietz WH, Srinivasan SR, Berenson GS., The relation of overweight to cardiovascular risk factors among children and adolescents: the Bogalusa Heart Study. *Pediatrics.* 1999;103:1175-1182.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The evidence review used by AACAP prioritized study designs less subject to bias and studies that represent the best scientific evidence. The evidence review included a large number of studies with large numbers of patients from various populations. Overall, the quality of the evidence regarding metabolic monitoring for children and adolescents on antipsychotics is high. The evidence provides a strong link between antipsychotic use and adverse metabolic side effects in youth and to negative long-term health outcomes throughout the lifespan.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

AACAP did not cite any studies that directly evaluated the benefit of metabolic monitoring for children and adolescents on antipsychotics. However, the evidence demonstrates the adverse side effects, including diabetes, weight gain, and hyperlipidemia, as well as the concerns regarding the safety of long-term antipsychotics use in youth. Thus, AACAP estimates there is a greater benefit to be gained through increased vigilance and regular metabolic monitoring.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

AACAP did not cite any studies that directly evaluated the harm of metabolic monitoring for children and adolescents on antipsychotics. AACAP noted that some patients and their parents may face negative social consequences due to frequent medical appointments, including greater time constraints on school and work responsibilities. However, given the adverse side effects and concerns regarding the safety of long-term antipsychotics use in youth, regular metabolic monitoring is still vital in the follow-up of these patients. Thus, in the absence of evidence of other harms, AACAP estimated there is less harm through increased vigilance and regular metabolic monitoring.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

To our knowledge, there have been no new studies that contradict the current body of evidence.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

N/A

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Metabolic_Monitoring_Evidence.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure addresses metabolic monitoring as one facet of safe and judicious use of antipsychotics in children and adolescents. Although antipsychotic medications offer the potential for effective treatment of psychiatric disorders in children, they can also increase a child's risk for developing serious metabolic health complications associated with poor cardiometabolic outcomes in adulthood. Despite the risk of such adverse side effects, research suggests that children and adolescents do not receive appropriate laboratory monitoring. Thus, this measure encourages metabolic monitoring of children who are on antipsychotic medications.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
New measure: not applicable

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
In a review of the literature, one study of Medicaid-enrolled children in three states found that only 31 percent of youth starting an atypical antipsychotic received a glucose test, and only 14 percent received a lipid test—far lower than rates reported for adults (Morrato et al., 2010). A second study found the association of atypical antipsychotics with diabetes is greater among children and adolescents than adults (Hammerman et al., 2008).

As part of the measure's field-testing, we assessed the rates of baseline metabolic screening and ongoing monitoring in Medicaid children, using the Medicaid Analytic eXtract data files. Based on data from 2008 for 11 states, the percentage of children receiving metabolic screening within 30 days of a new antipsychotic medication prescription was 6.0 percent, with a range of 0.4 percent to 14.0 percent. For children and adolescents who had ongoing antipsychotic use, the percentage that received metabolic monitoring was on average 18.5 percent, with a range of 4.8 percent to 36.2 percent.

In an examination of claims data from 17 Medicaid health plans in one state, we found that the average percentage of children receiving baseline metabolic screening within 30 days of a new antipsychotic medication prescription among the general population of children in health plans was 10.3 percent, with a range of 0.2 to 17.8 percent. For ongoing metabolic monitoring during the measurement year, the data suggests similar gaps in care. The percentage of children with ongoing antipsychotic use receiving metabolic monitoring during the measurement year was 30.9 percent, with a range of 2.3 to 40.0 percent.

Citations

Hammerman, A., J. Dreiher, S.H. Klang, H. Munitz, A.D. Cohen, M. Goldfracht. September 2008. Antipsychotics and diabetes: an age-related association. Annals of Pharmacotherapy. 2(9):1316–22.

Morrato, E., G. Nicol, D. Maahs, B. Druss, D. Hartung, R. Valuck, et al. 2010. Metabolic screening in children receiving antipsychotic drug treatment. Archives of Pediatric and Adolescent Medicine. 164, 344–51.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We tested the measure and evaluated disparities in its performance using Medicaid Analytic eXtract (MAX)data.

MAX DATA DESCRIPTION AND RESULTS
Our MAX dataset was composed of 2008 service data from 11 states. The analysis population included all Medicaid enrolled persons aged 0-20 on December 31, 2008 in the 11 states. Both fee-for-service and managed care enrollees were included. Data files included person summary, outpatient claims, inpatient claims and prescription claims. States were chosen due to completeness of their data for managed care enrolled beneficiaries.

Performance rates for the measure were constructed and tested by three age strata (0-5, 6-11, 12-17), race/ethnicity and foster care status. Of the 11 states, eight had higher rates of metabolic monitoring for the foster care population compared with the general Medicaid population. In both the general population and foster care population, metabolic monitoring was highest among adolescents compared with the lower age strata. We found that in the general population, rates of metabolic monitoring were slightly higher among Hispanic children and adolescents (24.8 percent) than Black Non-Hispanic (19.4 percent) and White Non-Hispanic (19.1 percent) children and adolescents. Similarly, in the foster care population, rates of metabolic monitoring were slightly higher among Hispanic (31.3 percent) than Black Non-Hispanic (23.2 percent) and White Non-Hispanic (24.0 percent) children and adolescents.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
A leading cause of morbidity/mortality, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Antipsychotic prescribing for children has increased rapidly in recent decades, driven by new prescriptions and by longer duration of use (Patten et al., 2012, Cooper et al., 2006). While antipsychotic medications offer the potential for effective treatment of psychiatric disorders in children, they can also increase a child's risk for developing serious health concerns, including metabolic health complications. Antipsychotic medications are associated with a number of potentially adverse impacts, including weight gain (Correll et al., 2009) and diabetes (Andrade et al. 2011; Bobo et al., 2013); both can have serious implications for future health outcomes. For example, metabolic problems in childhood and adolescence are associated with poor cardiometabolic outcomes in adulthood (Srinivasan et al., 2002). Obesity and dyslipidemias in childhood carry increased long-term health risk into adulthood. The long-term consequences of pediatric obesity and other metabolic disturbances include higher risk of heart disease in adulthood (Baker et al., 2007), cancer and shortened life span (Daniels, 2006). Diabetes is associated with serious cardiovascular, neurological and renal complications, including heart disease, stroke, blindness, kidney failure and nervous system damage (Centers for Disease Control and Prevention, 2011). Other serious risks associated with antipsychotic medications in children include extrapyramidal side effects, sedation and somnolence, liver toxicity and cardiac arrhythmias (Correll, 2008). Due to the potential negative health consequences associated with children developing cardiometabolic side effects from these medications, it is important to continuously monitor metabolic indices to ensure appropriate management of side effects.

A multi-year study of youth enrolled in three health maintenance organizations found that exposure to atypical antipsychotics was associated with a fourfold risk of diabetes in the following year, compared with children not prescribed a psychotropic medication,

the broader class of medications under which antipsychotics fall (Andrade et al., 2011). Another study of youth enrolled in a state Medicaid plan found that those starting an antipsychotic had three times the risk of developing diabetes compared with youth starting other psychotropic medications (Bobo et al, 2013). The association of atypical antipsychotics with diabetes has been found to be greater among children and adolescents than among adults (Hammerman et al., 2008).

Although there is little research available on the fiscal burden associated with adverse effects of antipsychotic use among children and adolescents, one study of Medicaid-enrolled youth on antipsychotics found that health care costs for patients who developed cardiometabolic side effects were 34 percent higher than those who did not (Jerrell, 2009). Further, diabetes is one of the most expensive chronic conditions in children (Imperatore et al., 2012). Proper screening and monitoring can contribute to early detection and management of cardiometabolic side effects and thus reduce long-term costs.

1c.4. Citations for data demonstrating high priority provided in 1a.3

Andrade, S.E., J.C. Lo, D. Roblin, et al. December 2011. Antipsychotic medication use among children and risk of diabetes mellitus. Pediatrics. 128(6):1135-41.

Baker, J., L. Olesen, T. Sorensen. 2007. Childhood body mass index and the risk of coronary heart disease in adulthood. New England Journal of Medicine. 357:2329–37.

Bobo, W.V., W.O. Cooper, C.M. Stein, et al. October 1, 2013. Antipsychotics and the risk of type 2 diabetes mellitus in children and youth. JAMA Psychiatry. 70(10):1067–75.

Centers for Disease Control and Prevention. 2011. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Cooper, W.O., P.G. Arbogast, H. Ding, G.B. Hickson, D.C. Fuchs, and W.A. Ray. 2006. Trends in prescribing of antipsychotic medications for US children. Ambulatory Pediatrics. 6(2):79–83.

Correll, C.U. 2008. Antipsychotic use in children and adolescents: minimizing adverse effects to maximize outcomes. FOCUS: The Journal of Lifelong Learning in Psychiatry. 6(3):368–78.

Correll, C. U., Manu, P., Olshanskiy, V., Napolitano, B., Kane, J. M., & Malhotra, A. K. 2009. Cardiometabolic risk of second-generation antipsychotic medications during first-time use in children and adolescents. Journal of the American Medical Association. 302(16):1765-1773.

Daniels, S.R. 2006. The consequences of childhood overweight and obesity. The future of children. 16(1):47–67.

Hammerman, A., J. Dreiher, S.H. Klang, H. Munitz, A.D. Cohen, M. Goldfracht. September 2008. Antipsychotics and diabetes: an age-related association. Annals of Pharmacotherapy. 2(9):1316–22.

Jerrell, J.M., R.S. McIntyre. July–August 2009. Health-care costs of pediatric clients developing adverse events during treatment with antipsychotics. Value Health. 12(5):716–22.

Imperatore, G., J. Boyle, T. Thompson, et al. 2012. Projections of Type 1 and Type 2 Diabetes Burden in the U.S. Population Aged < 20Years Through 2050. Diabetes Care. 35: 2515–20.

Patten, S.B., W. Waheed, L. Bresee. 2012. A review of pharmacoepidemiologic studies of antipsychotic use in children and adolescents. Canadian Journal of Psychiatry. 57:717–21.

Srinivasan, S.R., L. Myers, G.S. Berenson. January 2002. Predictability of childhood adiposity and insulin for developing insulin resistance syndrome (syndrome X) in young adulthood: the Bogalusa Heart Study. Diabetes. 51(1):204–9.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Endocrine : Screening, Mental Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Medication Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
None

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** XXXX_APM_Value_Sets.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Children and adolescents who received glucose and cholesterol tests during the measurement year.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months (January 1 – December 31)

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Children and adolescents who received at least one test for blood glucose (Glucose Tests Value Set) or HbA1c (HbA1c Tests Value Set) and at least one test for LDL-C (LDL-C Tests Value Set) or cholesterol (Cholesterol Tests Other Than LDL Value Set) during the measurement year (January 1 – December 31). See attachment for all value sets (S.2b).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Children and adolescents who had ongoing use of antipsychotic medication (at least two prescriptions).

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Children and adolescents age 1-17 as of December 31 of the measurement year (January 1 – December 31) who had at least two antipsychotic medication dispensing events (Table APM-A) of the same or different medications, on different dates of service during the measurement year.

Table APM-A: Antipsychotic Medications
First-generation antipsychotic medications: Chlorpromazine HCL; Fluphenazine HCL; Fluphenazine decanoate; Fluphenazine enanthate; Haloperidol; Haloperidol decanoate; Haloperidol lactate; Loxapine HCL; Loxapine succinate; Molindone HCL; Perphenazine; Pimozide; Promazine HCL; Thioridazine HCL; Thiothixene; Thiothixene HCL; Trifluoperazine HCL; Triflupromazine HCL
Second-generation antipsychotic medications: Aripiprazole; Asenapine; Clozapine; Iloperidone; Lurasidone; Olanzapine; Olanzapine pamoate; Paliperidone; Paliperidone palmitate; Quetiapine fumarate; Risperidone; Risperidone microspheres; Ziprasidone HCL; Ziprasidone mesylate
Combinations: Olanzapine-fluoxetine HCL (Symbyax); Perphenazine-amitriptyline HCL (Etrafon, Triavil [various])

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
No exclusions

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Report three age stratifications and a total rate:
1–5 years.
6–11 years.
12–17 years.
Total (sum of the age stratifications).

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Determine the eligible population, or the denominator, by identifying the number of patients in the specified age range who had at least two antipsychotic medication dispensing events (Table APM-A) of the same or different medications, on different dates of service during the measurement year.
Step 2: Determine the numerator by identifying the number of patients in the eligible population who received at least one glucose and one cholesterol test during the measurement year.
Step 3: Divide the numerator by the denominator to calculate the rate.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is part of the Healthcare Effectiveness Data and Information Set (HEDIS). As part of HEDIS, this measure pulls from administrative claims collected in the course of providing care to health plan members. NCQA collects the HEDIS data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

This measure has also been tested at the state level and could be reported by states if added to a relevant program.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic, Behavioral Health/Psychiatric : Outpatient, Laboratory
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Metabolic_Monitoring_Testing_10-12-15.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 2800 (New Measure)
Measure Title: Metabolic Monitoring for Children and Adolescents on Antipsychotics
Date of Submission: 10/9/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

This measure was tested using administrative claims data from the following sources.

- State analyses
  - Medicaid Analytic eXtract (MAX)
- Health plan analyses
  - Medicaid health plans from one state
  - Commercial health plans nationwide

For more information about MAX, refer to http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Data-and-Systems/MAX/MAX-General-Information.html

1.3. What are the dates of the data used in testing? Click here to enter date range

MAX data 2008, Medicaid health plan data for 17 plans 2010, and commercial health plan data for 73 plans 2012.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency

☒ health plan	☒ health plan
☒ other: State; Integrated Delivery system	☒ other: State; Integrated Delivery system

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

As part of the Pediatric Quality Measures Program (PQMP), NCINQ had access to the Medicaid Analytic eXtract (MAX) for conducting state analyses. In addition, NCINQ was able to test this measure in Medicaid health plan data from one large mid-Atlantic state. In order to assess the measure's use for HEDIS, we conducted an additional analysis in commercial data from a large administrative database. Our samples were as follows.

- State analyses
  - 2008 claims data from the MAX for 11 states
- Health plan analyses
  - 2010 claims data from 17 Medicaid health plans from one mid-Atlantic state
  - 2012 claims data from 73 commercial health plans nationwide

The administrative data sources included claims for all of the data elements needed to capture this measure, including claims for health care system encounters, laboratory codes, and pharmacy codes.

For our MAX analysis, the 11 states were chosen on the basis of Mathematica Policy Research reports that suggested that they provided adequate encounter/managed care data (Byrd & Dodd, 2012; Byrd & Dodd, 2013).

Citations

Byrd VLH, Dodd AH. Assessing the usability of encounter data for enrollees in comprehensive managed care across MAX 2007-2009. December 2012 2012.

Byrd VLH, Dodd AH. Assessing the Usability of MAX 2008 Encounter Data for Comprehensive Managed Care. Medicare & Medicaid Research Review. 2013;3(1).

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We tested a set of several measures related to antipsychotic use in the three datasets described above. Our analyses included enrollees who met continuous enrollment and measure-specific criteria. Our commercial health plan analyses included enrollees age 0-17 years during the measurement year. All other analyses included enrollees ages 0 to 20 during the measurement year. The age ranges varied slightly as our draft concepts were refined and in order to make the measures relevant to states (children/adolescents typically defined as age up to 21 years) and health plans (children/adolescents typically defined as age up to 18 years). We excluded enrollees who were dually eligible for Medicaid and Medicare. In the MAX data, a total of 148,910 children and adolescents met the denominator criteria and were included in the sample for this measure. Across the 17 Medicaid plans, the total number of children and adolescents who met denominator criteria was 14,174, and across 52 commercial plans that had sufficient denominators (>30), the total was 15,227.

Below are descriptions of the patient samples in terms of denominator sizes across the entities measured. They include the mean denominator, minimum denominator, maximum denominator, and the 25th, 50th (or median), and 75th percentiles.

Denominator Size Distribution Across 11 States (MAX) (2008)

Mean	13,537
Minimum	1,784
25th	6,272
Median	12,372
75th	18,684
Maximum	28,997

Denominator Size Distribution Across 17 State Medicaid Health Plans from One State (2010)

Mean	834
Minimum	125
25th	306
Median	748
75th	1,082
Maximum	2,437

Denominator Size Distribution Across 52* Commerical Health Plans Nationwide (2012)

Mean	293
Minimum	33
25th	103
Median	206
75th	369
Maximum	1,870

* Of the 73 commercial plans included in the testing of this measure, 52 had sufficient denominators (>30)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Reliability of the measure score was tested using a beta-binomial calculation and this analysis included the entire data samples described in the sections above (MAX state data, Medicaid heath plan, commercial health plan).

Validity was demonstrated through a systematic assessment of face validity. Per NQF instructions we have described the composition of the technical expert panels which assessed face validity in the data sample questions above. In addition, validity was demonstrated through two types of analyses: correlations among measures using Spearman Correlation Coefficients (using commercial health plan data sample) and rankings of health plans and states on measures (using MAX state data sample and Medicaid health plan data sample). This analysis is described further in section 2b2.3.

For identifying statistically significant & meaningful differences in performance, all three data samples were used (MAX state data, Medicaid heath plan, commercial health plan).

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
We assessed differences across multiple age strata (0-5, 6-11, 12-17, and total [0-17]), race/ethnicity (Hispanic; White, non-Hispanic; Black, non-Hispanic), and foster care status.

2a2. RELIABILITY TESTING
***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
Reliability Testing of Performance Measure Score: The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the "signal." The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents "noise." The reliability of the measure is represented as the ratio of signal to noise.

- A score of 0 indicates none of the variation (signal) is attributable to the plan
- A score of 1.0 indicates all of the variation (signal) is attributable to the plan
- A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY
The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan's variation for the average plan's variation. Thus, the reliability for some plans may be more or less than the overall reliability across plans.

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)
This measure achieved a reliability score above 0.7 for both state- and plan-level reliability.

	Average Reliability	Minimum Reliability
MAX States	.99	.99
Medicaid Health Plan	.98	.89
Commercial Health	.83	.35

Plan		

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
As stated in 2a2.2, we estimated reliability with a beta-binomial model (Adams, 2009). A score of zero implies that all the variability in a measure is attributable to measurement error. A score of 1.0 implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one reporting entity from another. A score of 0.7 or higher indicates adequate reliability to distinguish performance between two entities and is considered acceptable. The testing results suggest that this measure has adequate reliability for states and health plans, with very high reliability for Medicaid health plans and states in particular.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☒ **Empirical validity testing**
☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

<u>Face Validity</u>
The health-plan level of this measure was assessed for use in the HEDIS Health Plan Measure Set. As part of this process, NCQA assessed the face validity of the measure using its HEDIS process. NCQA staff shared the measure concepts, supporting evidence and field test results with its standing Behavioral Health Measurement Advisory Panel, Technical Measurement Advisory Panel and additional panels. We posted the measures for Public Comment, a 30-day period of review that allowed interested parties to offer feedback about the measure. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle. This measure has undergone the following steps associated with that cycle.

Step 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure "Desirable"? The work-up is vetted by NCQA's Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

Step 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health

plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

Step 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQAs Board of Directors will be included in the next HEDIS year and reported as first-year measures.

Step 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

Empirical Validity

As part of field testing, we assessed construct validity, which considers whether measures are capturing important aspects of a quality concept. We conducted two types of analyses: correlations among measures and rankings of health plans and states on measures.

We tested for construct validity by exploring whether this measure was correlated with other related measures, including the *Metabolic Screening for Children and Adolescents on Antipsychotics* measure, the *Follow-up Visit for Children and Adolescents on Antipsychotics* measure and the *Use of First-line Psychosocial Care for Children and Adolescents on Antipsychotics* measure. The *Metabolic Screening* measure assesses the percentage of youth who undergo metabolic testing prior to or immediately after the start of a new antipsychotic prescription. The *Follow-up Visit* measure assesses the percentage of youth who have one or more visits with a prescriber within 30 days after the start of a new antipsychotic prescription. The *Psychosocial Care* measure assesses the percentage of youth who have psychosocial care provided before or soon after the start of a new antipsychotic prescription. A higher rate indicates better performance for all three measures.

We hypothesized that organizations that perform well on this measure should perform well on the other measures. We calculated correlations using the Spearman correlation coefficient, which estimates the strength of the linear association between two continuous variables. The magnitude of correlation ranges from -1 and +1.; a value of 1 indicates a perfect linear dependence in which increasing values on one variable are associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable are associated with decreasing values of the second variable.

We then explored whether entities that conduct metabolic monitoring also manage other aspects of antipsychotic-related care well. We looked to see if plans and states can be approximately ranked based on profiles of performance across multiple measures. Consistency of performance across measures suggests that the measures are assessing a dimension of quality.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face Validity Results

Step 1: This measure was developed to address the need for metabolic monitoring of children and adolescents who are on antipsychotics. NCQA and five expert panels worked together in 2013 and 2014 to identify the most appropriate method for assessing metabolic monitoring among this patient population. Across the multiple expert panels that reviewed this measure, all panels concluded this measure was specified to assess metabolic monitoring.

Step 2: The measure was written and field-tested in 2013 and 2014. After reviewing field test results, the CPM recommended to send the measure to public comment with a majority vote in January 2014.

Step 3: The measure was released for Public Comment in 2014 prior to publication in HEDIS. Of 67 comments received, nearly all (94 percent) supported it as-is or with suggested modifications. The CPM recommended moving this measure to first year data collection by a majority vote in May 2014.

Step 4: The measure was introduced in HEDIS 2015. Organizations voluntarily reported this measure in the first year (2014) and the results were analyzed for public reporting in the following year (2015). The measure was approved in September 2015 by the CPM for public reporting in HEDIS 2016 for Medicaid and commercial plans.

Empirical Validity Results

Correlations

When determining correlations among measures, we focused on health plans, as there were not enough entities to measure correlations with the state data.

Among national commercial plans, there was a very slight positive correlation between the *First-line Psychosocial Care* and *Metabolic Monitoring* measures (r=0.12, p=.70) and high positive correlation between the *Metabolic Screening* and *Metabolic Monitoring* measures (r=0.82, p<0.0001).

Measure	**Pearson Correlation Coefficients**		
	Psychosocial Care	Metabolic Screening	Metabolic Monitoring
Psychosocial Care	1	0.18	0.12
Metabolic Screening		1	0.82
Metabolic Monitoring			1

Among Medicaid plans in one state, there was a slight positive correlation between the *Follow-up Visit* and *Metabolic Monitoring* measures (r=0.14, p=.58) and high positive correlation between the *Metabolic Screening* and *Metabolic Monitoring* measures (r=0.72, p<0.001).

Ranking

Among MAX states and one state's Medicaid plans, we found good consistency in the states and plans, respectively, with the best and worst performance.

MAX State Performance Rankings: General Population

State	**Metabolic Monitoring**	**Metabolic Screening**	**Follow-Up Visit**
1	14.2	**2.6**	**60.2**

2	19.4	4.5	**68.4**
3	20.6	5.5	75.0
4	**6.5**	3.8	71.2
5	**4.8**	**0.4**	74.9
6	18.7	4.8	76.4
7	20.0	6.3	69.0
8	14.8	5.3	N/A
9	**29.1**	**10.7**	N/A
10	19.6	8.3	**81.3**
11	**36.2**	**14.0**	**78.8**
Mean	**18.5**	**6.0**	**72.8**

Medicaid Health Plan Performance Rankings for One State

Plan	Metabolic Monitoring	Metabolic Screening	Follow-Up Visit
3	**2.3**	**0.2**	**71.0**
9	30.8	**4.9**	81.8
6	34.0	12.3	83.5
17	**39.7**	14.8	**86.7**
2	38.8	**15.4**	80.5
8	35.0	12.6	81.1
4	28.4	9.3	78.7
5	33.8	10.6	80.0
1	36.0	12.8	82.1
11	29.1	6.1	74.4
16	31.2	10.6	78.8
15	30.4	10.8	80.9
12	34.7	13.3	77.2
13	32.5	**17.8**	**70.4**
7	**20.3**	5.1	85.3
14	27.9	7.1	**98.7**
10	**40.0**	10.6	78.9
Mean	**30.9**	**10.3**	**80.6**

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Face Validity

The results indicate the expert panels showed good agreement that the measure as specified will accurately differentiate quality across states and health plans. Our interpretation of these results is that this measure has sufficient face validity.

Empirical Validity

Correlations

Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. The results indicate that plans that perform well on initial metabolic screening for those youth newly on antipsychotics also perform well on ongoing metabolic monitoring for those who continue on antipsychotics.

Ranking

The results show that plans and states can be approximately ranked based on profiles of performance across multiple measures. The consistent performance across measures suggest the measures are assessing a dimension of quality.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the*

information provided related to performance gap in 1b)
To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR). The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Variation in Performance Rates across MAX States (2008 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
18.5	6.5	14.2	19.4	20.6	29.1	6.4

IQR: Interquartile range

Variation in Performance Rates across Medicaid Plans from one State (2010 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
30.9	24.9	28.8	32.5	35.5	39.2	6.7

IQR: Interquartile range

Variation in Performance Rates across Commercial Plans Nationwide (2012 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
7.7	2.8	4.5	7.2	10.9	13.2	6.4

IQR: Interquartile range

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
The results indicate that there is 6.7% gap in performance between Medicaid plans at the 25th and 75th percentiles, a 6.4% gap in performance among commercial plans and a 6.4% gap in performance among states at the 25th and 75th percentiles. This means that states at the 25th percentile have on average 866 less children and adolescents getting recommended metabolic monitoring than states at the 75th percentile. For Medicaid plans, those at the 25th percentile have on average 56 less children and adolescents getting recommended metabolic monitoring than plans at the 75th percentile. For commercial plans, those at the 25th percentile have on average 19 less children and adolescents getting recommended metabolic monitoring than plans at the 75th percentile.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than***

one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
States and plans collect this measure using all administrative data sources, for all intents and purposes, there are no missing data in administrative data. We have done no assessment to look for the distribution of missing data. For plans reporting on this measure for HEDIS, NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
N/A

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Field testing results, more fully described in the Scientific Acceptability section, showed the measure is feasible to be collected by health plans and states using administrative claims data. Further, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the managed care organization´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Through our Policy Clarification Support System, NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of measures. Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Quality Compass http://www.ncqa.org/tabid/177/Default.aspx http://www.ncqa.org/tabid/836/Default.aspx The State of Health Care Quality Report Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx http://www.ncqa.org/tabid/836/Default.aspx The State of Health Care Quality Report

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

QUALITY COMPASS: This measure has just been approved for use in Quality Compass, a tool that displays health plan-level performance rates for HEDIS measures. It is used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans,

measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks. The Quality Compass 2015 Commercial tool includes data for 400 public reporting commercial health plan products, serving approximately 103.5 million covered lives. Benchmarks are calculated from a total pool of 420 public and non-public reporting health plan products, serving approximately 104 million covered lives. The Quality Compass 2015 Medicaid tool includes data for 182 public reporting Medicaid health plan products, serving approximately 20 million covered lives. Benchmarks are calculated from a total pool of 244 public and non-public reporting health plan products, serving approximately 25 million covered lives.

THE STATE OF HEALTH CARE QUALITY REPORT: HEDIS measures are publicly reported nationally and by geographic regions in the State of Health Care Quality Report. This report published by NCQA summarizes findings on quality of care. In 2015 the report included measures on 15.4 million Medicare Advantage beneficiaries in 507 Medicare Advantage health plans, 103.9 million members in 413 commercial health plans, and 25.4 million Medicaid beneficiaries in 237 plans across 50 states.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No negative consequences have been reported since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are

compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
1932 : Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)
2337 : Antipsychotic Use in Children Under 5 Years Old

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
This new measure assesses metabolic monitoring during the measurement year among children and adolescents who are prescribed antipsychotics. Below we detail how this measure is related to measures 2337 and 1932 but how it addresses a different target population and measure focus. Measure 2337 assesses whether children under 5 are prescribed an antipsychotic at some point during the measurement year. Similar to the Metabolic Monitoring for Children and Adolescents on Antipsychotics measure, this measure is specified for the health plan level and uses administrative claims as the data source. Measure 2337 focuses on all children under 5 years of age; our measure focuses on a broader range of children (up to age 18) who have been prescribed antipsychotics in order to assess whether they are receiving recommended testing. Measure 1932 assesses whether adults with schizophrenia or bipolar disorder who were prescribed antipsychotics are screened for diabetes. Similar to the Metabolic Monitoring for Children and Adolescents on Antipsychotics measure, this measure is specified for the health plan level and uses administrative claims as the data source. The measures have different target populations but a similar measure focus. Measure 1932 focuses on adults 18 to 64 years of age who have schizophrenia or bipolar disorder and who are prescribed antipsychotics. The Metabolic Monitoring for Children and Adolescents on Antipsychotics measure includes all children and adolescents up to 18 years of age who are prescribed antipsychotics and does not focus on any specific conditions. Measure 1932 is focused on diabetes screening by receipt of a glucose test. While the Metabolic Monitoring for Children and Adolescents on Antipsychotics measure also includes assessing whether a glucose test was received, it additionally assesses whether a cholesterol test was received since the focus is not just diabetes screening. The two measures are aligned in the way glucose testing is identified and measured.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee on Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-
Co.3 **Measure Developer if different from Measure Steward:** National Committee on Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
NCQA Behavioral Health Measurement Advisory Panel
Bruce Bobbitt, PhD, LP, Optum
Peter Delany, PhD, LCSW-C, Substance Abuse and Mental Health Services Administration
Ben Druss, MD, MPH, Emory University
Frank A. Ghinassi, PhD, ABPP, Western Psychiatric Institute and University of Pittsburgh Medical Center
Rick Hermann, MD, Tufts Medical Center and UpToDate, Inc.
Connie Horgan, ScD, Brandeis University
Neil Korsen, MD, Maine Health
Charlotte Mullican, BSW, MPH, Agency for Healthcare Research and Quality
Harold Pincus, MD, Columbia University and RAND Corporation
Bruce L. Rollman, MD, MPH, University of Pittsburgh School of Medicine
Michael Schoenbaum, PhD, National Institute of Mental Health
John H. Straus, MD, Massachusetts Behavioral Health Partnership and Beacon Health Options

NCQA Committee on Performance Measurement (CPM)
Bruce Bagley, MD, American Academy of Family Physicians
Andrew Baskin, MD, Aetna
Patrick Conway, MD, MMSc, Center for Medicare & Medicaid Services
Jonathan D. Darer, MD, Geisinger Health System
Helen Darling, National Business Group on Health
Rebekah Gee, MD, MPH, FACOG, LSU School of Medicine and Public Health
Foster Gesten, MD, NYSDOH Office of Managed Care
David Grossman, MD, MPH, Group Health Physicians
Christine Hunter, MD (Co-Chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Bernadette Loftus, MD, The Permanente Medical Group
J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, RN, PhD, AARP
Eric C. Schneider, MD, MSc (Co-Chair), RAND Corporation
Marcus Thygeson, MD, MPH, Blue Shield of Californa

NCQA Technical Measurement Advisory Panel
Andy Amster, MSPH, Kaiser Permanente
Kathryn Coltin, MPH, Independent Consultant
Lekisha Daniel-Robinson, Centers for Medicare and Medicaid Services

Marissa Finn, MBA, Cigna HealthCare
Scott Fox, MS, MEd, Independence Blue Cross
Carlos Hernandez, CenCalHealth
Kelly Isom, MA, RN, Aetna
Harmon Jordan, ScD, RTI International
Ernest Moy, MD, MPH, Agency for Healthcare Research and Quality
Patrick Roohan, New York State Department of Health
Lynne Rothney-Kozlak, MPH, Rothney-KozlakConsulting, LLC
Natan Szapiro, Independence Blue Cross

National Collaborative for Innovation in Quality Measurement (NCINQ) Measurement Advisory Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Katie Brookler, Colorado Department of Health Care Policy and Financing
Cathy Caldwell, MPH, Alabama Department of Public Health
Jennifer Havens, MD, NYU School of Medicine
Ted Ganiats, MD, University of California, San Diego
Darcy Gruttadaro, JD, National Allegiance on Mental Illness
Virginia Moyer, MD, MPH, FAAP, Baylor College of Medicine, USPSTF
Edward Schor, MD, Lucile Packard Foundation for Children's Health
Xavier Sevilla, MD, FAAP, Whole Child Pediatrics
Gwen Smith, Illinois Department of Healthcare and Family Services/Health Management Associates
Janet (Jessie) Sullivan, MD, Hudson Health Plan
Kalahn Taylor-Clark, PhD, MPH, George Mason University
Craig Thiele, MD, CareSource
Charles Wibbelsman, MD, Kaiser Permanente Medical Group, Inc.
Jeb Weisman, PhD, Children's Health Fund

NCINQ Consumer Panel
Joan Alker, MPhil, Georgetown Center for Children and Families
Roni Christopher, MEd, OTR/L, PCMH-CCE, The Greater Cincinnati Health Collaborative
Daniel Coury, MD, Nationwide Children's Hospital
Eileen Forlenza, Colorado Medical Home Initiative, Children and Youth with Special Health Care Needs Unit
Michaelle Gady, JD, Families USA
Janis Guerney, JD, Family Voices
Jocelyn Guyer, MPA, Georgetown Center for Children and Families
Catherine Hess, MSW, National Academy for State Health Policy
Carolyn Muller, RN, Montgomery County Health Department
Cindy Pellegrini, March of Dimes
Judith Shaw, EdD, MPH, RN, VCHIP
Stuart Spielman, JD, LLM, Autism Speaks
Michelle Sternthal, PhD, March of Dimes

NCINQ Foster Care Panel
Kamala Allen, MHS, Center for Health Care Strategies
Mary Applegate, MD, Ohio Department of Job and Family Services
Samantha Jo Broderick, Foster Care Alumni of America
Mary Greiner, MD, Cincinnati Children's Hospital Medical Center
David Harmon, MD, FAAP, Superior HealthPlan
Patricia Hunt, Magellan Health Services
Audrey LaFrenier, MSW, Parsons Child and Family Center
Bryan Samuels, MPP, Chapin Hall
Phil Scribano, DO, MSCE, The Children's Hospital of Philadelphia
Lesley Siegel, MD, State of Connecticut Department of Children and Families
Chauncey Strong, MSW, LGSW, Fairfax County Department of Family Services/Foster Care and Adoption
Janet (Jessie) Sullivan, MD, Hudson Health Plan
Nora Wells, MS, National Center for Family/Professional Partnerships

NCINQ Mental Health Panel
Francisca Azocar, PhD, Optum Health Behavioral Solutions
Frank Ghinassi, PhD, Western Psychiatric Institute and Clinic of UPMC Presbyterian Shadyside
Jennifer Havens, MD, NYU Langone Medical Center
Danielle Laraque, MD, FAAP, Maimonides Infants and Children's Hospital of Brooklyn

NCINQ State Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Sharon Carte, MHS, State of West Virginia Children's Health Insurance Program
Susan Castellano, Minnesota Department of Human Services
Catherine Hess, MSW, National Academy for State Health Policy
Michael Hogan, PhD, New York State office of Mental Health
Barbara Lantz, MN, RN, State of Washington Department of Social and Health Services, Medicaid Purchasing Administration
Judy Mohr Peterson, PhD, Oregon Health Authority
Tracy Plouck, MPA, Ohio Department of Mental Health
Gina Robinson, Colorado Department of Health Care Policy and Financing
Janet Stover, Illinois Association of Rehabilitation Facilities
Eric Trupin, PhD, University of Washington

NCINQ Measure Development Partners
Shahla Amin, MS, Rutgers University
Scott Bilder, PhD, Center for Health Services Research, Rutgers University
Stephen Crystal, PhD, Institute for Health, Health Care Policy and Aging Research, Rutgers University
Molly Finnerty, PhD, NY State Office of Mental Health
Emily Leckman-Westin, PhD, NY State Office of Mental Health
Sheree Neese-Todd, MA, Institute for Health, Health Care Policy and Aging Research, Rutgers University

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2014
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2016

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.
These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance or otherwise modify the measures, and shall not disassemble, recompile or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA. © 2012 by the National Committee for Quality Assurance

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2801
Measure Title: Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics
Measure Steward: National Committee on Quality Assurance
Brief Description of Measure: Percentage of children and adolescents 1–17 years of age with a new prescription for an antipsychotic, but no indication for antipsychotics, who had documentation of psychosocial care as first-line treatment.
Developer Rationale: This measure addresses use of first-line psychosocial care as one facet of safe and judicious use of antipsychotics in children and adolescents. Antipsychotic prescribing for youth has increased rapidly in recent decades. Although antipsychotic medications may serve as effective treatment for a narrowly defined set of psychiatric disorders in youth, they are often being prescribed for nonpsychotic conditions for which psychosocial interventions are considered first-line treatment. Thus, clinicians may be underutilizing safer first-line psychosocial interventions, and youth may be unnecessarily incurring the risks associated with antipsychotic medications and experiencing poorer mental and physical health outcomes.

Numerator Statement: Children and adolescents from the denominator who had psychosocial care as first-line treatment prior to (or immediately following) a new prescription of an antipsychotic.
Denominator Statement: Children and adolescents who had a new prescription of an antipsychotic medication for which they do not have a U.S Food and Drug Administration primary indication.
Denominator Exclusions: Exclude children and adolescents with a diagnosis of a condition for which antipsychotic medications have a U.S. Food and Drug Administration indication and are thus clinically appropriate: schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders.

Measure Type: Process
Data Source: Administrative claims
Level of Analysis: Health Plan, Integrated Delivery System, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this process measure:

- This measure encourages the use of psychosocial care prior to or immediately following administration of antipsychotics if the child does not have a U.S. Food and Drug Administration (FDA) indication for antipsychotics (schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders). If psychosocial care is successful,

antipsychotic use may be halted or avoided altogether. The path envisioned is as follows. Child does NOT have a primary indication for antipsychotic use >>> Health care provider utilizes psychosocial care intervention >>> Child avoids unnecessary antipsychotic use >>> Child avoids adverse side effects associated with antipsychotic medications >>> Child experiences improvement in mental and physical outcomes (desired outcome).

- This measure is based on 24 clinical practice guidelines, standards, and recommendations from three organizations, including the American Academy of Child and Adolescent Psychiatry (AACAP), the Center for Education and Research on Mental Health Therapeutics, and the Center for the Advancement of Children's Mental Health. The developer reports that clinical standards are based on rigorous and/or substantial empirical evidence, and/or overwhelming clinical consensus and that the clinical guidelines are based on strong empirical evidence and/or strong clinical consensus. The recommendations are based on varying levels of evidence, from very strong to expert consensus, with strong levels of agreement. Overall the quality of the evidence is moderate to high, according to the developer.
- The developer focuses on the guidelines from AACAP as it is most relevant to the measure. The evidence review encompassed 147 clinical trials, meta-analyses, practice guidelines, randomized controlled trials (RCTs), systematic literature reviews, and case reports and series. The measure is based on two recommendations, which are graded by AACAP:
    - **Recommendation 1:** "Prior to the initiation of and during treatment with an AAA, the general guidelines that pertain to the prescription of psychotropic medications should be followed… including education and psychotherapeutic interventions for the treatment and monitoring of improvement."
    - **Recommendation 2:** "In the absence of specific FDA indications or substantial evidence for effectiveness, physicians should consider other medication or psychosocial treatments before initiating antipsychotic treatment."
- According to the developer, the guidelines from AACAP do not include an exact estimate of the benefits of psychosocial care, or the potential harms of treating children with psychosocial care prior to initiating antipsychotics. However, the evidence has established that antipsychotic use is associated with adverse short-term metabolic and other side effects in youth and with negative long-term health outcomes throughout the lifespan.

Questions for the Committee

- *Is the relationship between the process to patient outcomes clear and reasonable, and what is the strength of evidence for the relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer reports the following:

- Clinicians may be underutilizing safer first-line psychosocial interventions, and youth may be unnecessarily incurring the risks associated with antipsychotic medications and experiencing poorer mental and physical health outcomes.
- Even as the use of psychopharmacological interventions has increased, the proportion of children and adolescents receiving outpatient psychotherapy declined from 2.95 percent in 1998 to 2.72 percent in 2007. One study of Medicaid-enrolled children and youth starting an antipsychotic medication found that nearly one-third did not receive concurrent psychosocial therapy (Harris et al., 2012). This study also found that youth 12-17 years who are prescribed antipsychotics are less like to receive concurrent psychotherapy than children 6-11 years. A second study of privately insured children 2-5 years found that only 40 percent prescribed an antipsychotic also had one or more therapy visits in the measurement year (Olfson et al., 2010).
- During testing, the developer used MAX data from 11 states. It assessed the proportion of Medicaid children age 0-20 years on antipsychotic medications who had documented psychosocial care. Administrative claims data from eight states found a range of 35.8-64.1 percent of children prescribed antipsychotic medication who had

documented psychosocial care; the average was 48.2 percent.

- Psychosocial care rates were lower among White (43.4 percent) children than Black (49.3 percent) and Hispanic (46.6 percent) children. In the foster care population, rates were lowest in the Hispanic (53.7 percent) population compared to Black (57.2 percent) and White (57.5 percent) children.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- No concerns.
- I have several concerns with this measure:
  - 1) antipsychotics are also used in kids with severe reactive behaviors associated with non-approved conditions (genetic and neurodevelopmental conditions which may be related to severe reactivity, for example), where therapy and psychosocial interventions may not be effective unless appropriate medications are on board.
  - 2) there is risk for not treating where clinically indicated (school failure/suspension, social and family functioning...).
  - 3) this measure doesn't address important role for early intervention services (0-3 yo) in preventing the need for pharm intervention.
- This process measure is based on guidelines of professional societies, that themselves were developed through a combination of evidence review and consensus. It is unclear whether the specific recommendation underlying this measure was directly supported by data in some way, or by expert opinion. For example the designation as a clinical standard is "based on rigorous/substantial empirical evidence and/or overwhelming clinical consensus." It would be good to know which.
- I also wonder whether the evidence was really directly about behavioral approaches prior to prescribing AAAs or rather merely about the effectiveness of those approaches for various conditions. The developers state that "there have been no studies comparing the short-term cost-effectiveness" of the two approaches.
- The AACAP recommendation is really quite general… "general guidelines that pertain to psycotroptic medications should be followed…". In the second one, it is that physicians "should consider" other medication or psychosocial treatment.
- There is no evidence of harms presented, and I agree harm is unlikely. But, it is possible that in some cases there are risks (or self harm etc.) to delay in treatment with medication. These may be very rare.
- In reviewing the algorithm for evidence, the determination hinges on whether the SR indicates high certainty. There is not enough information presented to know this. Rather, given the professional consensus documented, this may be "Insufficient Evidence with Exception."
- Process measure - The developers again cite AACAP recommendations (and other organizations) that encourage "psychotherapeutic interventions" prior to initiation of and during treatment with an antipsychotic. While I agree it is the right thing to do, I don't see any evidence cited tying receipt of psychotherapy to improved outcomes nor absence of psychotherapy tied to poorer outcomes. The relationship between the process measure and patient outcomes does seem reasonable, but I think the strength of the evidence (other than expert consensus recommendations) for the relationship is poor.
- Evidence supports the use of psychosocial prior to initiating antipsychotic medications, especially when antipsychotic medications are not indicated. This is supported by 24 clinical practice guidelines according to the author of the measure.
- The evidence relates directly, to the process outcome. The process, the quantification by health care entities that report the use of psychosocial interventions as first line of treatment for the targeted groups of patients before or at the same time antipsychotic drugs are prescribed, has the goal of ensuring that these none drug therapies are used

first. It is noted that the effectiveness of psychosocial therapy may not be quantified, but it useful in many situations and it is preferred to the use of these stated drug therapies with known short term and long term consequences.

1b. Performance Gap.

- Overall there is a lower than optimal performance and therefore, performance gap.
- Performance gap exists, but also likely related to lack of access to appropriate services in a timely manner.
- The developers provide data that less than half of patients receiving these medications have evidence (in claims) of psychosocial intervention. The increase in use of antipsychotics without FDA indication does make this a significant national concern.
- A significant issue is whether behavioral interventions always appear as distinct claims for health plans. They certainly appear frequently, but given the number of professionals (social workers, other therapists) and settings (clinics, community settings, schools), it would be good to see some evidence of the proportion that appear in claims. A validation with chart review (of a sample) would be one way to assess this.
- Differences exist by race/ethnicity but are not as dramatic as in some other measures.
- Performance data on the measure is provided, both from published studies and the developers' analysis of 11 state MAX data. The data suggests use of psychotherapeutic interventions are on the decline. The developers cite disparities between different populations, although it isn't clear if these disparities are statistically significant.
- Use of antipsychotic medication is on the rise, while use of psychosocial therapies in children and adolescents is declining. Those receiving medications are also often not receiving concurrent therapy.
- Several gaps in performance were noted from studies. A review of MAX data in 8 states indicated an average of 48% of 0 - 20 year old received psychosocial care. Older children (11-17) has less psychosocial therapy than young children. Over recent years, the use of psychosocial therapy has decreased somewhat while use of antipsychotic drugs has markedly increased.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The numerator is children and adolescents from the denominator who had psychosocial care as first-line treatment prior to (or immediately following) a new prescription of an antipsychotic.
- The denominator is children and adolescents who had a new prescription of an antipsychotic medication for which they do not have a U.S Food and Drug Administration primary indication.
- Applicable coding for the measure is included in "S2b. Data Dictionary Code Table" of the measure submission form; the value set is provided as an attachment. There are ICD-9 and ICD-10 codes used for this measure and they are included in the excel file. The ICD-10 conversion information also is provided.
- The algorithm logic is straightforward.
- This process measure is not risk adjusted, but, in addition to reporting a single rate, it is stratified by age.

Questions for the Committee

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is

precise enough to distinguish differences in performance across providers.

The developer reports that:

- The measure was tested with 2008 MAX data from 11 states, 2010 claims data from 17 Medicaid health plans from one state, and 2012 claims data from 73 commercial plans nationwide.
- This measure was tested at the performance measure score level using a beta-binomial signal-to-noise analysis. For this type of testing, a score of zero implies that all the variability in a measure is attributable to measurement error. A score of 1.0 implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one reporting entity from another. A score of 0.7 or higher indicates adequate reliability to distinguish performance between two entities and is considered acceptable.
    - The average reliability at the state level was 0.99, the Medicaid plan level was 0.97, and the commercial plan level was 0.77, suggesting a very high level of reliability for the measure, particularly for states and Medicaid plans.
- Per the NQF algorithm, reliability testing at the computed performance measure score may be rated HIGH, MODERATE, or LOW depending on the testing results.

Questions for the Committee

- *Does the Committee concur with the developer's conclusion that the results demonstrate sufficient reliability so that differences in performance can be identified?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The specifications appear consistent with the evidence. The goal of the measure is to encourage psychosocial interventions prior to beginning use of antipsychotic treatments for nonprimary indications, in order to reduce harmful metabolic and physical health complications caused by the medications. The evidence provided supports this goal.

Question for the Committee

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer provides the following information:

- The measure was tested at the performance measure score level using both empirical testing and face validity.
- For the empirical testing, the developer assessed construct validity with two types of analyses: correlations among measures and rankings of health plans and states on measures on the three antipsychotic medication measures. The developer reports the following results:
    - Correlations were tested using health plans, as there were not enough entities to test among states.
    - Among national commercial plans, there was moderate positive correlation between the Follow-Up Visit and Psychosocial Care measures (r=0.59, p=0.03) and very slight positive correlation between the Metabolic Screening and Psychosocial Care measures (r=0.18, p=.55). Among MAX states and one state's Medicaid plans, testing found good consistency in the states and plans, respectively, with the best and worst performance.
- Per the NQF algorithm, validity testing at the computed performance measure score may be rated HIGH,

MODERATE, or LOW depending on the testing results.

- The developer used its standardized HEDIS process to test face validity at the health plan level, but does not explicitly call out face validity of the **computed performance score**, as required by NQF.
  - The developer worked with five expert panels to identify the most appropriate method for assessing the use of multiple concurrent antipsychotics among this patient population. All of the panels concluded this measure was specified to assess multiple concurrent use of antipsychotics.
  - The draft measure was put out for public comment and brought to the developer's Committee on Performance Measurement.
  - The developer states that the measure has sufficient face validity.

Questions for the Committee

- *Do the results of the empiric testing demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you believe that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The measure excludes youth with conditions for which there is a U.S. Food and Drug Administration indication for antipsychotics (schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders).
- The developer reports that on average 25% of children age 0-5 years with a new start of an antipsychotic met the exclusion criteria for having a primary indication for antipsychotic use (i.e., schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders); 29% of children age 6-11 years met the exclusion criteria; 25% of adolescents age 12-17 uear met the exclusion criteria. The application of the exclusion to the measure reduced rates on average across plans by less than 2% for those age 0-5 years, increased rates by less than 2% for those age 6-11 uears and increased rates by just over 2% for those age 12-17 years.
- The developer states that because the data is collected administratively, the exclusions do not pose a burden and they do not adversely affect the denominator.

Questions for the Committee

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment:

- This measure is not risk-adjusted.

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer calculates an inter-quartile range (IQR) to demonstrate meaningful differences. The IQR provides a measure of the dispersion of performance, and can be interpreted as the difference between the 25th and 75th percentile on a measure.
- The developer states that the results indicate there is a 23.7% gap in performance between Medicaid plans at the 25th and 75th percentiles, a 16.5 % gap in performance among commercial plans, and a 20.7% gap in performance among states at the 25th and 75th percentiles.

Question for the Committee:

- *Does this measure identify meaningful differences in quality?*

2b6. Comparability of data sources/methods:

- This is not needed.

2b7. Missing Data

- The measure is collected using all administrative data sources. According to the developer, there are no missing data from the data used for the measure.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- Focus placed on potential harms of medication, but no focus placed on potential harms of not placing on medication, especially if referring to therapy and services cannot occur in a timely manner.
- Data elements are clearly defined. The definition of behavioral treatments is broad (which is a strength). My major concern is that given that without some look back period, it would not be clear that the dispensing in the measurement year was a "new" treatment. It is not possible that the patient could have been on the medication for years, and that psychosocial treatment had been attempted in the prior year. In short, I don't see how the specifications ensure that this is a "new" prescription.
- The developers reference a "psychosocial core value set" - the services within this set seem to cover a wide spectrum of services from psychotherapy for patients and families to dance/music/art therapy. Although they are identified by claims data and therefore reliably identified, I would assume that the potential benefits from the various services aren't equal. The other data elements are clearly defined and inclusive from a coding perspective. The measure appears that it could be consistently implemented.
- The data elements appear to be clearly defined. The logic algorithm appears to be sound.

2a2. Reliability testing

- Reliability testing was performed using beta-binomial models. It showed good reliability at the state level and for large health plans. However 6 of 19 commercial plans had insufficient denominators.
- The reliability results put the reliability in the Moderate range, assuming that data capture is complete.
- I concur with the developer that the results demonstrate sufficient reliability.
- Reliability testing included: 2008 MAX - 11 state data, 2010 Medicaid Health plan from one state and 2012 data from 73 commercial plans nationwide. Performance measure scores were evaluation by Beta-binomial signal to noise analysis. With highest score of 1.0 and score of 0.7 consistence with adequate reliability scores were: 0.99 for state date, 9.7 for Medicaid Plan data and 0.77 for Commercial group data. Developers felt that this was consistent with High Reliability. Sample size for all groups appears adequate, with the commercial group having fewer subjects.
- I believe this was at the data element level.

2b1. Validity Specifications

- The evidence is limited and general. The specifications are consistent with the consensus opinion with the caveat above regarding whether the specifications identify "new" prescriptions as the measure title implies.
- Again - I'm fully supportive of psychosocial interventions as first-line for certain indications. I just don't know that I'm convinced that the evidence connecting kids who receive psychosocial interventions as first line, to children not receiving as many antipsychotics, to improved outcomes is clear (although it is reasonable).
- The test results suggest that the measure have adequate reliability for states and health plans, with very high reliability for Medicaid plans and states.
- The goal of this measure: To encourage psychosocial intervention prior to beginning use of antipsychotic therapy for non-primary indications. The goal would support decreased known harmful metabolic and physical consequences of drug use.
- The evidence supports this goal.

2b2. Validity Testing

- The face validity is high because of the process using expert panels and public comment that NCQA undertakes.
- The validity testing has adequate scope in terms of geography and both Medicaid and commercial health plans.
- The correlations with other measures are fairly weak. The ranking analysis is quite general, and there is no quantification of whether this level of correlation in ranking is adequate for this type of measure.

- I would rate the validity of this measure as Low, based on the data presented.
- Similar to previous measures - assessed construct validity with multiple analyses as well as used a standardized process to demonstrate face validity with good outcomes.
- Expert panels suggest good face validity of measure.
- Health plans that performed well on follow up visit and metabolic screening also consistently perform well on the Psychosocial Care measure.
- States and plans can be ranked based on their profiles of performance.
- Validity testing appears to have been largely in face value with consensus of 5 expert Panels, though there was not too much information presented about this. The developer also looked at correlations with other related types of care in this population and found the correlations supported this measure.

2b3-2b7. Threats to Validity

- While it is not considered "missing data" has the developer conducted a cross-walk with the measure and the AAP health care standards for treating foster care children to ensure they are not conflicting.
- Missing data may be an issue if psychosocial treatment does not appear in claims.
- Developer states this measure is collected using administrative data sources and therefore there are no missing data.
- Exclusions of persons where an antipsychotic medication is a first-line intervention appear appropriate for this measure.
- No missing data reported by author.
- Exclusions were well defined and included patients who have USDA approved indications for antipsychotic use as in these cases psychosocial therapy is not a first one treatment.
- The exclusions are consistent with the evidence.
- No risk adjustment noted.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer notes:

- All elements are generated through normal process of care, and are in defined fields in electronic claims.
- The measure is a part of HEDIS, which has a standardized collection and calculation process, as well as a system to collect real-time feedback from measure users.
- Testing results showed the measure is feasible to be collected by health plans and states using administrative claims data.
- As part of HEDIS, the data elements are subject to that program's data collection and audit requirements.
- This is not an eMeasure.

Questions for the Committee

- *Are the required data elements routinely generated and used during care delivery?*
- *Does the testing data collection strategy indicate the measure is ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Balancing concerns of overprescribing antipsychotics and associated safety issues, there should also be discussion on potentially withholding effective treatment for targeted symptoms. (altered developmental trajectory, global social and family functioning, sense of well being, etc...).
- Feasibility will not be a problem given that this is a claims-based measure. Currently used in NCQA's Quality

Compass.

- The required data elements are routinely generated in the normal course of patient care. The data collection strategy places the measure ready to be put into operational use.
- All data elements can be obtained through electronic health records or claims data.
- The required at a elements ARE routinely generated during care delivery.
- The measure is already a part of HEDIS measures and in 2015 part of Quality Compass for Medicaid.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

The developer notes:

- The measure is currently in use in for both public reporting and quality improvement with benchmarking.
- The measure is included in Quality Compass for Medicaid 2015, a tool that displays health plan-level performance rates for HEDIS measures. It is used for selecting a health plan, conducting competitor analysis, examining quality improvement, and benchmarking plan performance.
- The measure also is reported on in The State of Health Care Quality Report, a national report produced by the developer that reports the results from HEDIS measures.
- This is a new measure and improvement results are not yet available.
- No unintended consequences have been reported thus far.

Question for the Committee

- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Can the developer provide the analysis from the use of this measure for HEDIS 2015. Any lessons learned and/or proposed modifications to the technical specifications being considered? Feedback from health plans?
- The measure is currently in use for multiple public reporting opportunities. If I had to identify potential unintended consequences - one may be that children instead of beginning pharmacotherapy instead are referred for BH services. Resources for these services will be scarce in many areas. There will also be children and families who may present for initial evaluation but then don't consistently follow through with the recommended services. In these populations there will be potential for bad outcomes if the child is receiving neither psychotherapy nor pharmacotherapy. How you balance that with adverse effects from pharmacotherapy alone, I don't know.
- Measure has been approved for Quality Compass (reporting of HEDIS measures). This measure can be used for public reporting and quality improvement efforts to improve the quality of healthcare provided to children and adolescents receiving behavioral health interventions.

Criterion 5: Related and Competing Measures

- This measure is related to the NQF-endorsed 2337: Antipsychotic Use in Children Under 5 Years Old. However, this new measure has a broader age population and different focus (i.e., focus on new diagnosis and use of psychosocial care).

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** N/A

Measure Title: Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 10/9/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Psychosocial care provided before or immediately following a new start of an antipsychotic medication

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

This measure encourages the use of psychosocial care prior to or immediately following administration of antipsychotics if the child does not have a U.S. Food and Drug Administration (FDA) indication for antipsychotics (schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders). If psychosocial care is successful, antipsychotic use may be halted or avoided altogether. The path envisioned is as follows.

Child does NOT have a primary indication for antipsychotic use >>> Health care provider utilizes psychosocial care intervention >>> Child avoids unnecessary antipsychotic use >>> Child avoids adverse side effects associated with antipsychotic medications >>> Child experiences improvement in mental and physical outcomes (desired outcome).

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

- American Academy of Child and Adolescent Psychiatry. 2012. Practice Parameter for the Use of Atypical Antipsychotic Medications in Children and Adolescents. http://www.aacap.org/App_Themes/AACAP/docs/practice_parameters/Atypical_Antipsychotic_Medications_Web.pdf (July 12, 2012)
- McClellan J., R. Kowatch, R.L. Findling. January 2007. Practice parameter for the assessment and treatment of children and adolescents with bipolar disorder. *J Am Acad Child Adolesc Psychiatry.* 46(1):107–25.
- Gleason, M.M., H.L. Egger, G.J. Emslie, et al. December 2007. Psychopharmacological treatment for very young children: contexts and guidelines. J Am Acad Child Adolesc Psychiatry. 46(12):1532–72.
- Scotto, Rosato N., C.U. Correll, E. Pappadopulos, A. Chait, S. Crystal, P.S. Jensen. June 2012. Treatment of maladaptive aggression in youth: CERT guidelines II. Treatments and ongoing management. Pediatrics. 129(6):e1577–86.
- Steiner H, Remsing L. 2007. Practice parameter for the assessment and treatment of children and adolescents with oppositional defiant disorder. J Am Acad Child Adolesc Psychiatry. 46:126–141.
- Pappadopulos, E., Ii J.C. Macintyre, M.L. Crismon, et al. February 2003. Treatment recommendations for the use of antipsychotics for aggressive youth (TRAAY). Part II. J Am Acad Child Adolesc Psychiatry. 42(2):145–61.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Guideline (Date)	Population	Recommendation or Statement	Type/Grade
AACAP-AAA (2011) Practice parameter for the use of atypical antipsychotic medications in children and adolescents.	5-18 years	"Prior to the initiation of and during treatment with an AAA, the general guidelines that pertain to the prescription of psychotropic medications should be followed… *including education and psychotherapeutic interventions for the treatment and monitoring of improvement*" (Recommendation 1)	Clinical Standard
		"*In the absence of specific FDA indications or substantial evidence for effectiveness, physicians should consider other medication or psychosocial treatments before initiating antipsychotic treatment*." (under Recommendation 2)	Clinical Standard
AACAP-BP (2007) Practice parameter for the assessment and treatment of children and adolescents with bipolar disorder.	≤18 years	"Psychotherapeutic interventions are an important component of a comprehensive treatment plan for early-onset bipolar disorder".(Recommendation 10)	Minimal Standard

Guideline (Date)	Population	Recommendation or Statement	Type/Grade
AACAP-ODD (2007) Practice parameter for the assessment and treatment of children and adolescents with oppositional defiant disorder.	≤18 years	"The clinician should develop an individualized treatment plan based on the specific clinical situation…..*The two types of evidence-based treatments for youth with ODD are individual approaches in the form of problem solving skills and family interventions in the form of parent management training*" (Recommendation 7)	Minimal Standard
		"The clinician should consider parent intervention based on one of the empirically tested interventions" (Recommendation 8)	Minimal Standard
		"Medications may be helpful as adjuncts to treatment packages, for symptomatic treatment and to treat comorbid conditions" (Recommendation 9) Supporting notes recommend that if medications are initiated, it should be after psychosocial interventions are in place, and that medications should not be the only treatment. *"Several open and double-blind placebo controlled studies show that typical and atypical antipsychotics are helpful in treating aggression after appropriate psychosocial interventions have been applied in the context of mental retardation and PDD"* (under Recommendation 9)	Clinical Guideline
AACAP-SZ (2001) Practice parameter for the assessment and treatment of children and adolescents with schizophrenia.	≤18 years	"Adequate treatment requires the combination of psychopharmacological agents plus psychosocial interventions" (Recommendations – Treatment)	Minimal Standard
		"The following psychosocial interventions are recommended: 1. Psychoeducational therapy for the patient, including ongoing education about the illness, treatment options, social skills training, relapse prevention, basic life skills training, and problem-solving skills and strategies, 2. Psychoeducational therapy for the family to increase their understanding of the illness, treatment options, and prognosis and for developing strategies to cope with the patients symptoms." (Recommendations—Psychosocial Interventions)	Minimal Standard
		"Specialized educational programs and/or vocational training programs may be indicated for some children or adolescents to address the cognitive and functional deficits with the illness." (Recommendations—Psychosocial Interventions)	Clinical Guidelines

Guideline (Date)	Population	Recommendation or Statement	Type/Grade
PPWG (2007) The AACAP-sponsored Preschool Psycho-pharmacology Working Group—Psychopharma-cological treatment for very young children: Contexts and guidelines.	<6 years	"Universal guidelines are provided to encourage careful and planful clinical practice: Avoid medications when therapy is likely to produce good results Generally, an adequate trial of psychotherapy precedes consideration of medication, and psychotherapy continues if medications are used…"	(See diagnostic specific ratings)
		ADHD: Parent Management Training or other behavioral intervention x 8 weeks minimum, is first line for preschoolers	A (preschool)
		Disruptive behavioral disorders: Psychotherapy (e.g., Parent management training, parent child interaction therapy) x 10-20 weeks	A (preschool)
		MDD: Psychotherapy is first line (e.g., dyadic psychotherapy, target emotional regulation) x 3-6 months	C (preschool) A (6-18yrs)
		BP: Psychotherapy is first line (e.g., dyadic psychotherapy, target emotional regulation) x 8-12 sessions	C (preschool) A (6-18yrs)
		Anxiety (GAD, SAD, SM, SP): CBT is first line, x 12 weeks	C (preschool) A (6-18yrs)
		PTSD: Psychotherapy is first line (Child Parent Psychotherapy x 6 months minimum; or CBT x 12 weeks minimum, or if unavailable then Play therapy x months	A (Preschool CPP, CBT) B (Preschool; Play therapy) A (6-18yrs, CBT)
		OCD: CBT with parent involvement, behavioral therapy x 12 weeks minimum	C (Preschool) A (6-18 yrs)
		PPD: Behavioral, developmental, psychoeducational intervention is first line	A (Preschool and 0-18 yrs)
		Sleep: Parent education and sleep hygiene	C (Preschool) A (6-18yrs)
TMAY (2012) Center for Education and Research on Mental Health Therapeutics—Treatment of maladaptive aggression in youth.	≤18 years	"Provide or assist the family in obtaining evidence-based parent and child skills training during all phases of care" (Recommendation 10)	Grade of evidence= A Strength of recommendation = Very Strong
		"Engage the child and family in taking an active role in implementing psychosocial strategies and help them to maintain consistency" (Recommendation 11)	Grade of evidence= B Strength of

Guideline (Date)	Population	Recommendation or Statement	Type/Grade
			recommendation= Very Strong
		"Recommendations 10 and 11 pertain to psychosocial interventions, which should be the first line of treatment because of its lower risk, preceding the use of medication to address aggression except in emergency circumstances…" (Under Treatment Recommendations – unrated explanatory comment)	Not specified
TRAAY (2003) Center for the Advancement of Children's Mental Health: Treatment recommendations for the use of antipsychotics for aggressive youth.[7]	≤18 years	Psychosocial and educational interventions should continue after medication treatment begins.	Not specified*

*TRAAY (2003) did not specify the use of a rating system.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

Guideline Developer	Definition
AACAP	*Minimal Standard/Clinical Standard:* Rigorous/ substantial empirical evidence (meta-analyses, systematic reviews, RCTs) and/or overwhelming clinical consensus; expected to apply more than 95 percent of the time
	Clinical guidelines: Strong empirical evidence (non-randomized controlled trials, cohort or case-control studies), and/or strong clinical consensus; expect to apply in most cases (75% of the time)
PPWG	*A:* Well controlled RCTs, large meta-analyses, or overwhelming clinical consensus
	B: Empirical evidence (open trials, case series) or strong clinical consensus
	C: Single case reports or no published reports, recommendation developed by expert consensus (informal)
TMAY Ratings	Oxford Centre for Evidence-Based Medicine grade of evidence (A-D)
	A: Consistent level 1 studies
	B: Consistent level 2 or 3 studies or extrapolations from level 1 studies
	Strength of Recommendation: Very strong (≥90% agreement)

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Guideline Developer	Definition
AACAP	*Options:* Acceptable but not required; there may be insufficient evidence to support higher recommendation (uncontrolled trials, case/series reports).

Guideline Developer	Definition
	Not endorsed: Ineffective or contraindicated.
AACAP endorsed best-practice principles	Best practice principles that underlie medication prescribing, to promote the appropriate and safe use of psychotropic medications
TMAY Ratings	Oxford Centre for Evidence-Based Medicine grade of evidence (A-D) C: Level 4 studies or extrapolations from level 2 or 3 studies
	D: Level 5 evidence or troublingly inconsistent or inconclusive studies of any level
	Strength of Recommendation: Strong (70-89% agreement)
	Strength of Recommendation: Fair (50-69% agreement)
	Strength of Recommendation: Weak (<50% agreement)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Several guidelines address the use of antipsychotic medications in children and adolescents; each guideline recommends use of psychosocial services prior to antipsychotics initiation.

The American Academy of Child and Adolescent Psychiatry (AACAP) guideline recommends use of psychosocial services prior to antipsychotics initiation, particularly in the absence of an FDA indication. These recommendations are based on established metabolic and other health risks of antipsychotics as well as evidence of efficacy of psychosocial treatments. While we list the full range of guidelines in sections 1a.4.2 and 1a.4.3 above, we focus on and describe in more detail the AACAP Guideline in the remaining sections, as it is most closely relevant to the specified measure, which assesses use of first-line psychosocial care in a general population of children given antipsychotics.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

AACAP rates the strength of the empirical evidence in descending order as follows:

- (rct) Randomized, controlled trial is applied to studies in which subjects are randomly assigned to two or more treatment conditions
- (ct) Controlled trial is applied to studies in which subjects are non-randomly assigned to two or more treatment conditions
- (ut) Uncontrolled trial is applied to studies in which subjects are assigned to one treatment condition
- (cs) Case series/report is applied to a case series or a case report

The other supporting guidelines recommend use of psychosocial services for specific mental health conditions. See tables in section 1a.4.2 for the level-of-evidence grade given to the remaining guidelines and section 1a.4.3 for the definition of each grade.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See table under 1a.4.4 for definitions.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).

Date range: 1990-2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

The AACAP-AAA guideline is rated a ***Clinical Standard***, indicating it is based on rigorous/substantial empirical evidence and/or overwhelming clinical consensus. AACAP includes several condition-specific guidelines around use of psychosocial services; we focus on the general AACAP-AAA antipsychotics guideline here and describe the body of evidence for each relevant recommendation below.

When developing their guideline, AACAP limited its evidence review to clinical trials, meta-analysis, practice guidelines, randomized controlled trials (RCTs), systematic literature reviews, and case reports and series. AACAP selected a total of 147 publications for careful examination based on their weight in the hierarchy of evidence attending to the quality of individual studies, relevance to clinical practice and the strength of the entire body of evidence. AACAP did not provide a breakdown of specific numbers of each publication type. We have identified where there are certain publication types available to support each guideline.

Recommendation 1: "Prior to the initiation of and during treatment with an AAA, the general guidelines that pertain to the prescription of psychotropic medications should be followed… including education and psychotherapeutic interventions for the treatment and monitoring of improvement."

This recommendation is based on a literature review conducted by a medical professional society on the established metabolic impacts of antipsychotics and other health risks and evidence of efficacy of psychosocial treatments. The literature review contained a total of 147 publications that included clinical trials, meta-analysis, practice guidelines, RCTs, systematic literature reviews, and case reports and series.
American Academy of Child and Adolescent Psychiatry. Practice parameter on the use of psychotropic medications in children and adolescents. *J Am Acad Child Adolesc Psychiatry.* 2009;48:961-973.

Recommendation 2: "In the absence of specific FDA indications or substantial evidence for effectiveness, physicians should consider other medication or psychosocial treatments before initiating antipsychotic treatment."

This recommendation is based on literature regarding the use of antipsychotics in specific clinical populations and the current FDA indications, which include only schizophrenia, bipolar disorder, tic disorders and specific symptoms of autistic disorder. In the absence of substantial empirical support for antipsychotics for other specific problems or specific FDA indications, AACAP recommends health care providers implement other pharmacological or psychosocial treatment modalities with more established efficacy and safety profiles prior to the onset of antipsychotics use.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The evidence review used by AACAP prioritized study designs less subject to bias and studies that represent the best scientific evidence. The evidence review included a large number of studies with large numbers of patients from various populations. Overall, the quality of the evidence regarding use of first-line psychosocial care for children and adolescents on antipsychotics is moderate to high.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

The AACAP-AAA review did not include an exact estimate of benefits of psychosocial care. However, the evidence has established that antipsychotic use is associated with adverse short-term metabolic and other side effects in youth and with negative long-term health outcomes throughout the lifespan.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The AACAP review did not examine the potential harms of treating children with psychosocial care prior to initiating antipsychotics. However, the harms of unnecessary antipsychotic use in kids has been well established (Andrade et al. 2011; Bobo et al., 2013; Correll, 2008; Correll et al., 2009; Crystal et al., 2009; Daniels, 2006; Lean and Pajonk, 2003; Srinivasan et al. 2002).

Citations

Andrade, S.E., J.C. Lo, D. Roblin, et al. December 2011. Antipsychotic medication use among children and risk of diabetes mellitus. *Pediatrics*. 128(6):1135–41.

Bobo, W.V., W.O. Cooper, C.M. Stein, et al. October 1, 2013. Antipsychotics and the risk of type 2 diabetes mellitus in children and youth. *JAMA Psychiatry.* 70(10):1067–75.

Correll, C.U. 2008. Antipsychotic use in children and adolescents: minimizing adverse effects to maximize outcomes. *FOCUS: The Journal of Lifelong Learning in Psychiatry.* 6(3):368–78.

Correll, C. U., Manu, P., Olshanskiy, V., Napolitano, B., Kane, J. M., & Malhotra, A. K. 2009. Cardiometabolic risk of second-generation antipsychotic medications during first-time use in children and adolescents. *Journal of the American Medical Association*. 302(16):1765-1773.

Crystal, S., M. Olfson, C. Huang, H. Pincus and T. Gerhard. 2009. Broadened use of atypical antipsychotics: Safety, effectiveness, and policy challenges. *Health Affairs.* 28:w770–81.

Daniels, S.R. 2006. The consequences of childhood overweight and obesity. *The future of children.* 16(1):47–67.

Lean, M.E., and F.G. Pajonk. 2003. Patients on Atypical Antipsychotic Drugs Another high-risk group for type 2 diabetes. *Diabetes Care.* 26(5), 1597–605.

Srinivasan, S. R., Myers, L., & Berenson, G. S. 2002. Predictability of childhood adiposity and Insulin for developing insulin resistance syndrome (syndrome X) in young adulthood the Bogalusa heart study. *Diabetes.* 51(1):204-209.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

To our knowledge, there have been no new studies that contradict the current body of evidence.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Psychosocial_Care_Evidence.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure addresses use of first-line psychosocial care as one facet of safe and judicious use of antipsychotics in children and adolescents. Antipsychotic prescribing for youth has increased rapidly in recent decades. Although antipsychotic medications may serve as effective treatment for a narrowly defined set of psychiatric disorders in youth, they are often being prescribed for nonpsychotic conditions for which psychosocial interventions are considered first-line treatment. Thus, clinicians may be underutilizing safer first-line psychosocial interventions, and youth may be unnecessarily incurring the risks associated with antipsychotic medications and experiencing poorer mental and physical health outcomes.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
New measure: not applicable

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Multiple studies have found that antipsychotics are increasingly being prescribed for children who have conditions which are not primary indications for the use of antipsychotics (Cooper et al., 2004; Olfson et al., 2006). Good evidence supports that psychosocial interventions are associated with positive outcomes for children and youth diagnosed with those conditions (Ollendick et al., 2006; Pelham and Fabiano, 2008; Weisz et al., 2005; Kutcher et al., 2004).

Even as the use of psychopharmacological interventions has increased, the proportion of children and adolescents receiving outpatient psychotherapy declined from 2.95 percent in 1998 to 2.72 percent in 2007 (Olfson et al., 2010). One study of Medicaid-enrolled children and youth starting an antipsychotic medication found that almost one-third did not receive concurrent psychosocial therapy (Harris et al., 2012). This study also found that youth 12–17 years who are prescribed antipsychotics are less like to receive concurrent psychotherapy than children 6–11. A second study of privately insured children 2–5 years found that only 40 percent prescribed an antipsychotic also had one or more therapy visits in the measurement year (Olfson et al., 2010).

As part of the measure's field-testing, using the Medicaid Analytic eXtract (MAX) data files, we assessed the proportion of Medicaid children age 0-20 years on antipsychotic medications who had documented psychosocial care. Analysis of administrative claims data from eight states demonstrated that the average percentage of children prescribed antipsychotic medication who had documented psychosocial care was 48.2 percent, with a range of 35.8–64.1 percent. Additional field-testing using data from one state's Medicaid plans found the average percentage of children prescribed antipsychotic medication who had documented psychosocial care to be 44.7 percent, with a range of 26.4–67.7 percent. These results suggest gaps in care and much room for improvement.

Citations

Cooper, W.O., G.B. Hickson, C. Fuchs, P.G. Arbogast, W.A. Ray. 2004. New Users of Antipsychotic Medications Among Children Enrolled in TennCare. Archives of Pediatric Adolescent Medicine. 158(8):753–9. DOI:10.1001/archpedi.158.8.753.

Harris, E., M. Sorbero, J.N. Kogan, J. Schuster, B.D. Stein. April 2012. Concurrent mental health therapy among Medicaid-enrolled youths starting antipsychotic medications. Psychiatric Services. 63(4):351–6.

Kutcher, S., M. Aman, S.J. Brooks, J. Buitelaar, E. van Daalen, J. Fegert, and S. Tyano. 2004. International consensus statement on attention-deficit/hyperactivity disorder (ADHD) and disruptive behaviour disorders (DBDs): clinical implications and treatment practice suggestions. European Neuropsychopharmacology. 14(1):11–28.

Olfson, M., C. Blanco, L. Liu, C. Moreno, G. Laje. 2006. National Trends in the Outpatient Treatment of Children and Adolescents with Antipsychotic Drugs. Archives of General Psychiatry. 63(6):679–85. DOI:10.1001/archpsyc.63.6.679.

Olfson, M., S.C. Marcus. December 2010. National trends in outpatient psychotherapy. Am J Psychiatry. 67(12):1456–63.

Ollendick, T.H., N.J. King, and B.F. Chorpita. 2006. Empirically supported treatments for children and adolescents in Child and adolescent therapy: Cognitive-behavioral procedures (3rd ed.) (pp. 492–520). New York, NY, US: Guilford Press.

Pelham, Jr., W.E., and G.A. Fabiano. 2008. Evidence-based psychosocial treatments for attention-deficit/hyperactivity disorder. Journal of Clinical Child & Adolescent Psychology. 37(1):184–214.

Weisz, J.R., A.J. Doss, and K.M. Hawley. 2005. Youth psychotherapy outcome research: A review and critique of the evidence base. Annual Review of Psychology. 56:337–63.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We tested the measure and evaluated disparities in its performance using Medicaid Analytic eXtract (MAX)data. We assessed performance by race/ethnicity as well as foster-care status.

MAX DATA DESCRIPTION AND RESULTS
Our MAX dataset was composed of 2008 service data from eight states. The analysis population included all Medicaid enrolled youth aged 0-20 on December 31, 2008 in the eight states. Both fee-for-service and managed care enrollees were included. Data files included person summary, outpatient claims, inpatient claims and prescription claims. States were chosen due to completeness of their data for managed care enrolled beneficiaries.

We found that rates of psychosocial care visits were slightly lower among White Non-Hispanic (43.4 percent) children compared to Black Non-Hispanic (49.3 percent) and Hispanic (46.6 percent) children. In the foster care population, rates of psychosocial care visits were slightly lower among Hispanic (53.7 percent) than Black Non-Hispanic (57.2 percent) and White Non-Hispanic (57.5 percent) children.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
Affects large numbers, A leading cause of morbidity/mortality, Patient/societal consequences of poor quality
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare.

List citations in 1c.4.

Antipsychotic prescribing for children and adolescents has increased rapidly in recent decades, driven by new prescriptions and by longer duration of use (Patten et al., 2012, Cooper et al., 2006). Although antipsychotic medications may serve as effective treatment for a narrowly defined set of psychiatric disorders in children, they are often being prescribed for nonpsychotic conditions such as attention-deficit disorder and disruptive behaviors (McKinney and Renk, 2011; Cooper et al., 2004; Olfson et al., 2006), conditions for which psychosocial interventions are considered first-line treatment (Kutcher et al., 2004; Pappadopulos et al., 2003; Scotto Rosato et al., 2012). Thus, clinicians may be underutilizing safer first-line psychosocial interventions and using antipsychotics for nonprimary indications in children and adolescents.

Use of antipsychotics in children can increase a child's risk for developing serious health issues such as metabolic and physical health complications (Crystal et al., 2009), which are of particular concern given the potential for adversely affecting development. Antipsychotic medications are associated with a number of potential adverse impacts, including weight gain (Correll et al., 2009) and diabetes (Andrade et al. 2011; Bobo et al., 2013); both can have serious implications for future health outcomes. For example, metabolic problems in childhood and adolescence are associated with poor cardio-metabolic outcomes in adulthood (Srinivasan et al., 2002). Obesity and dyslipidemias in childhood carry increased long-term health risk into adulthood, including heart disease, cancer and shortened life span (Daniels, 2006). Other serious risks associated with antipsychotic medications in children include extrapyramidal side effects, sedation and somnolence, liver toxicity and cardiac arrhythmias (Correll, 2008).

Children without a primary indication for an antipsychotic, who are not given the benefit of a trial of psychosocial treatment first, may unnecessarily incur the risks associated with antipsychotic medications. To the extent that psychosocial interventions are associated with better outcomes (Jensen et al., 2001; Eyberg et al., 2008; Schimmelmann et al., 2013), underuse of these therapies may lead to poorer mental and physical health outcomes.

There have been no studies comparing the short-term cost-effectiveness of antipsychotic treatment with psychosocial interventions, but psychosocial treatment is not known or proposed to have any ongoing costs or negative effects after termination, while antipsychotics have the potential to cause lasting health impacts and associated treatment costs. Children without a primary indication for an antipsychotic who are not given the benefit of a trial of psychosocial treatment may unnecessarily incur the costs/harms associated with antipsychotics, one of the most costly medication classes (Crystal et al., 2009), and substantial long-term costs of treating the health impacts associated with antipsychotic medications, including treatment of obesity, diabetes and dyslipidemias. There is some evidence that these health conditions, such as new onset diabetes, may not resolve after discontinuation of the antipsychotic (Lean and Pajonk, 2003). Although this is an understudied area, it is reasonable to assume that unresolved health impacts of antipsychotics could be associated with long-term increases in health costs established for obesity and diabetes.

1c.4. Citations for data demonstrating high priority provided in 1a.3

Andrade, S.E., J.C. Lo, D. Roblin, et al. December 2011. Antipsychotic medication use among children and risk of diabetes mellitus. Pediatrics. 128(6):1135–41.

Bobo, W.V., W.O. Cooper, C.M. Stein, et al. October 1, 2013. Antipsychotics and the risk of type 2 diabetes mellitus in children and youth. JAMA Psychiatry. 70(10):1067–75.

Cooper, W.O., P.G. Arbogast, H. Ding, G.B. Hickson, D.C. Fuchs, and W.A. Ray. 2006. Trends in prescribing of antipsychotic medications for US children. Ambulatory Pediatrics. 6(2):79-83.

Cooper, W.O., G.B. Hickson, C. Fuchs, P.G. Arbogast, W.A. Ray. 2004. New Users of Antipsychotic Medications Among Children Enrolled in TennCare. Archives of Pediatric Adolescent Medicine. 158(8):753–9. DOI:10.1001/archpedi.158.8.753.

Correll, C.U. 2008. Antipsychotic use in children and adolescents: minimizing adverse effects to maximize outcomes. FOCUS: The Journal of Lifelong Learning in Psychiatry. 6(3):368–78.

Correll, C. U., Manu, P., Olshanskiy, V., Napolitano, B., Kane, J. M., & Malhotra, A. K. 2009. Cardiometabolic risk of second-generation antipsychotic medications during first-time use in children and adolescents. Journal of the American Medical Association. 302(16):1765-1773.

Crystal, S., M. Olfson, C. Huang, H. Pincus and T. Gerhard. 2009. Broadened use of atypical antipsychotics: Safety, effectiveness, and policy challenges. Health Affairs. 28:w770–81.

Daniels, S.R. 2006. The consequences of childhood overweight and obesity. The future of children. 16(1):47–67.

Eyberg, S.M., M.M. Nelson, S.R. Boggs. January 2008. Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. Journal of Clinical Child and Adolescent Psychology. 37(1):
215–37.

Jensen, P.S., S.P. Hinshaw, J.M. Swanson, et al. February 2001. Findings from the NIMH Multimodal Treatment Study of ADHD (MTA): implications and applications for primary care providers. Journal of Developmental and Behavioral Pediatrics. 22(1):60–73.

Kutcher, S., M. Aman, S.J. Brooks, J. Buitelaar, E. van Daalen, J. Fegert, and S. Tyano. 2004. International consensus statement on attention-deficit/hyperactivity disorder (ADHD) and disruptive behaviour disorders (DBDs): clinical implications and treatment practice suggestions. European Neuropsychopharmacology. 14(1):11–28.

Lean, M.E., and F.G. Pajonk. 2003. Patients on Atypical Antipsychotic Drugs Another high-risk group for type 2 diabetes. Diabetes Care. 26(5), 1597–605.

McKinney, C., and K. Renk. 2011. Atypical antipsychotic medications in the management of disruptive behaviors in children: safety guidelines and recommendations. Clinical psychology review. 31(3):465–71.

Olfson, M., C. Blanco, L. Liu, C. Moreno, G. Laje. 2006. National Trends in the Outpatient Treatment of Children and Adolescents with Antipsychotic Drugs. Archives of General Psychiatry. 63(6):679–85. DOI:10.1001/archpsyc.63.6.679.

Pappadopulos, E., N.S. Rosato, C.U. Correll, et al. December 2011. Experts´ recommendations for treating maladaptive aggression in youth. Journal of Child and Adolescent Psychopharmacology. 21(6):505-515.

Patten, S.B., W. Waheed, L. Bresee. 2012. A review of pharmacoepidemiologic studies of antipsychotic use in children and adolescents. Canadian Journal of Psychiatry. 57:717–21.

Schimmelmann, B.G., S.J. Schmidt, M. Carbon, C.U. Correll. March 2013. Treatment of adolescents with early-onset schizophrenia spectrum disorders: in search of a rational, evidence-informed approach. Current Opinion in Psychiatry. 26(2):219–30.

Scotto, Rosato N., C.U. Correll, E. Pappadopulos, A. Chait, S. Crystal, P.S. Jensen. June 2012. Treatment of maladaptive aggression in youth: CERT guidelines II. Treatments and ongoing management. Pediatrics. 129(6):e1577–86.

Srinivasan, S. R., Myers, L., & Berenson, G. S. 2002. Predictability of childhood adiposity and Insulin for developing insulin resistance syndrome (syndrome X) in young adulthood the Bogalusa heart study. Diabetes. 51(1):204-209.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Mental Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access, Safety, Safety : Medication Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
None

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** XXXX_APP_Value_Sets.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Children and adolescents from the denominator who had psychosocial care as first-line treatment prior to (or immediately following) a new prescription of an antipsychotic.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
January 1 – December 31

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Children and adolescents who had documentation of psychosocial care (Psychosocial Care Value Set) in the 121-day period from 90 days prior to the index prescription start date (IPSD) through 30 days after the IPSD during the measurement year (January 1 – December 31). See attachment for all value sets (S.2b).

The Psychosocial Care Value Set contains claims codes for behavioral health acute inpatient and outpatient encounters, including psychotherapy for patients, families, and/or groups; psychophysiological therapy; hypnotherapy; activity therapy, such as music, dance, or art; training and educational services related to the care and treatment of mental health issues; community and rehabilitations programs; and crisis interventions. These services align with a recent Institute of Medicine (IOM) report*, which defined psychosocial interventions for mental health and substance use disorders as "interpersonal or informational activities, techniques, or strategies that target biological, behavioral, cognitive, emotional, interpersonal, social, or environmental factors with the aim of reducing symptoms of these disorders and improving functioning or well-being." The IOM notes these interventions include psychotherapies, vocational rehabilitation and peer support services, and that they can utilize different formats, including individual, family, or group therapy.

DEFINITIONS
IPSD: The earliest prescription dispensing date for an antipsychotic medication where the date is in the Intake Period and there is a Negative Medication History.
Negative Medication History: A period of 120 days (4 months) prior to the IPSD when the member had no antipsychotic medications dispensed for either new or refill prescriptions.

*Intitute of Medicine. Committee on Developing Evidence-Based Standards for Psychosocial Interventions for Mental Disorders,

Board on Health Sciences Policy. England MJ, Butler AS and Gonazlez ML, eds. Psychosocial Interventions for Mental and Substance Use Disorders: a Framework for Establishing Evidence-Based Standards. 2015. National Academies Press; Washington, DC (Prepublication copy).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Children and adolescents who had a new prescription of an antipsychotic medication for which they do not have a U.S Food and Drug Administration primary indication.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Children and adolescents age 1-17 as of December 31 of the measurement year (January 1 – December 31) who had a new prescription for an antipsychotic medication (Table APP-A) during the intake period (January 1 through December 1 of the measurement year).

Table APP-A: Antipsychotic Medications
First-generation antipsychotic medications: Chlorpromazine HCL; Fluphenazine HCL; Fluphenazine decanoate; Fluphenazine enanthate; Haloperidol; Haloperidol decanoate; Molindone HCL; Perphenazine; Pimozide; Haloperidol lactate; Loxapine HCL; Loxapine succinate; Promazine HCL; Thioridazine HCL; Thiothixene; Thiothixene HCL; Trifluoperazine HCL; Triflupromazine HCL
Second-generation antipsychotic medications: Aripiprazole; Asenapine; Clozapine; Iloperidone; Lurasidone; Olanzapine; Olanzapine pamoate; Paliperidone; Paliperidone palmitate; Quetiapine fumarate; Risperidone; Risperidone microspheres; Ziprasidone HCL; Ziprasidone mesylate
Combinations: Olanzapine-fluoxetine HCL (Symbyax); Perphenazine-amitriptyline HCL (Etrafon, Triavil [various])

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude children and adolescents with a diagnosis of a condition for which antipsychotic medications have a U.S. Food and Drug Administration indication and are thus clinically appropriate: schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Exclude children and adolescents for whom first-line antipsychotic medications may be clinically appropriate. Any of the following during the measurement year (January 1 – December 31) meet criteria:

Children and adolescents who have at least one acute inpatient encounter with a diagnosis of schizophrenia, bipolar disorder or other psychotic disorder during the measurement year. Any of the following code combinations meet criteria:
-BH Stand Alone Acute Inpatient Value Set with Schizophrenia Value Set.
-BH Stand Alone Acute Inpatient Value Set with Bipolar Disorder Value Set.
-BH Stand Alone Acute Inpatient Value Set with Other Psychotic Disorders Value Set.
-BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set and Schizophrenia Value Set.
-BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set and Bipolar Disorder Value Set.
-BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set and Other Psychotic Disorders Value Set.

Children and adolescents who have at least two visits in an outpatient, intensive outpatient or partial hospitalization setting, on different dates of service, with a diagnosis of schizophrenia, bipolar disorder or other psychotic disorder during the measurement year. Any of the following code combinations meet criteria:
-BH Stand Alone Outpatient/PH/IOP Value Set with Schizophrenia Value Set.
-BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set and Schizophrenia Value Set.
-BH Stand Alone Outpatient/PH/IOP Value Set with Bipolar Disorder Value Set.
-BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set and Bipolar Disorder Value Set.
-BH Stand Alone Outpatient/PH/IOP Value Set with Other Psychotic Disorders Value Set.
-BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set and Other Psychotic Disorders Value Set.

See attachment for all value sets (S.2b).

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Report three age stratifications and a total rate:
1–5 years.
6–11 years.
12–17 years.
Total (sum of the age stratifications).

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Determine the eligible population, or the denominator, by identifying the number of children and adolescents in the specified age range who were dispensed an antipsychotic medication (Table APP-A) during the intake period (January 1 – December 1).
Step 2: Exclude those who did not have a negative medication history and who have a diagnosis for which antipsychotic medications are clinically appropriate (see S.10).
Step 3: Determine the numerator by identifying the number of children and adolescents in the eligible population who had documentation of psychosocial care in the 121-day period from 90 days prior through 30 days after the new prescription of an antipsychotic.
Step 4: Divide the numerator by the denominator to calculate the rate.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.

N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is part of the Healthcare Effectiveness Data and Information Set (HEDIS). As part of HEDIS, the measure pulls from administrative claims collected in the course of providing care to health plan members. NCQA collects the HEDIS data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

The measure has also been tested at the state level and could be reported by states if added to a relevant program.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic, Behavioral Health/Psychiatric : Inpatient, Behavioral Health/Psychiatric : Outpatient
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Psychosocial_Care_Testing_10-12-15.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 2801 (New Measure)
Measure Title: Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics
Date of Submission: 10/9/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

This measure was tested using administrative claims data from the following sources.

- State analyses
  - Medicaid Analytic eXtract (MAX)
- Health plan analyses
  - Medicaid health plans from one state
  - Sample of Commercial health plans nationwide

For more information about MAX, refer to http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Data-and-Systems/MAX/MAX-General-Information.html.

1.3. What are the dates of the data used in testing? Click here to enter date range

MAX data 2008, 2010 Medicaid health plan data for 17 plans, and 2012 commercial health plan data for 73 plans.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency

☒ health plan	☒ health plan
☒ other: State; Integrated Delivery System	☒ other: State; Integrated Delivery System

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

As part of the Pediatric Quality Measures Program (PQMP), NCINQ had access to the Medicaid Analytic eXtract (MAX) for conducting state analyses. In addition, NCINQ was able to test this measure in Medicaid health plan data from one large mid-Atlantic state. In order to assess the measure's use for HEDIS, we conducted an additional analysis in commercial data from a large administrative database. Our samples were as follows.

- State analyses
  - 2008 claims data from the MAX for 11 states
- Health plan analyses
  - 2010 claims data from 17 Medicaid health plans from one mid-Atlantic state
  - 2012 claims data from 19 commercial health plans nationwide

These administrative data sources included claims for all of the data elements needed to capture this measure, including claims for health care system encounters, laboratory codes, and pharmacy codes.

For our MAX analysis, the 11 states were chosen on the basis of Mathematica Policy Research reports that suggested that they provided adequate encounter/managed care data (Byrd & Dodd, 2012; Byrd & Dodd, 2013). Of these 11 states, three were excluded in the testing for this measure due to lack of completeness of data.

Citations

Byrd VLH, Dodd AH. *Assessing the usability of encounter data for enrollees in comprehensive managed care across MAX 2007-2009.* December 2012 2012.

Byrd VLH, Dodd AH. Assessing the Usability of MAX 2008 Encounter Data for Comprehensive Managed Care. *Medicare & Medicaid Research Review.* 2013;3(1).

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

We tested a set of several measures related to antipsychotic use in three datasets described above. Our analyses included enrollees who met continuous enrollment and measure-specific criteria. Our commercial health plan analyses included enrollees age 0-17 years during the measurement year. All other analyses included enrollees ages 0 to 20 during the measurement year. The age ranges varied slightly as our draft concepts were refined and in order to make the measures relevant to states (children/adolescents typically defined as age up to 21 years) and health plans (children/adolescents typically defined as age up to 18 years). We excluded enrollees who were dually eligible for Medicaid and Medicare. In the MAX data, a total of 14,598 children and adolescents met the denominator criteria and were included in the sample for this measure. Across the 17 Medicaid plans, the total number of children and adolescents who met denominator criteria was 8,525, and across 13 commercial plans the total was 1,472.

Below are descriptions of the patient samples in terms of denominator sizes across the entities measured. They include the mean denominator, minimum denominator, maximum denominator, and the 25th, 50th (or median), and 75th percentiles.

Denominator Size Distribution Across Eight States (MAX) (2008)

Mean	832
Minimum	269
25th	371
Median	1,350
75th	1,990
Maximum	3,376

Denominator Size Distribution Across 17 State Medicaid Health Plans from One State (2010)

Mean	501
Minimum	53
25th	133
Median	426
75th	749
Maximum	1,384

Denominator Size Distribution Across 13* Commerical Health Plans Nationwide (2012)

Mean	113
Minimum	37
25th	48
Median	70
75th	112
Maximum	387

*Of the 19 plans included in the testing of this measure, 13 had sufficient denominators (>30)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Reliability of the measure score was tested using a beta-binomial calculation and this analysis included the entire data samples described in the sections above.

Validity was demonstrated through a systematic assessment of face validity. Per NQF instructions we have described the composition of the technical expert panels which assessed face validity in the data sample questions above. In addition, validity was demonstrated through two types of analyses: correlations among measures using Spearman Correlation Coefficients (using commercial health plan data sample) and rankings of health plans and states on measures (using MAX state data sample and Medicaid health plan data sample). This analysis is described further in section 2b2.3.

For testing the impact of exclusions, the commercial health plan data sample was used.

For identifying statistically significant & meaningful differences in performance, all three data samples were used (MAX state data, Medicaid heath plan, commercial health plan).

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
We assessed differences across multiple age strata (0-5, 6-11, 12-17, and total [0-17]), race/ethnicity (Hispanic; White, non-Hispanic; Black, non-Hispanic), and foster care status.

2a2. RELIABILITY TESTING
***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
Reliability Testing of Performance Measure Score: The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the "signal." The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents "noise." The reliability of the measure is represented as the ratio of signal to noise.

- A score of 0 indicates none of the variation (signal) is attributable to the plan
- A score of 1.0 indicates all of the variation (signal) is attributable to the plan
- A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY
The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan's variation for the average plan's variation. Thus, the reliability for some plans may be more or less than the overall reliability across plans.

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)
This measure achieved a reliability score above 0.7 for both state- and plan-level reliability.

Data Source	Average Reliability	Minimum Reliability
MAX States	.99	.91
Medicaid Health Plans	.97	.77
Commercial Health Plans	.77	.53

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
As stated in 2a2.2, we estimated reliability with a beta-binomial model (Adams, 2009). A score of zero implies that all the variability in a measure is attributable to measurement error. A score of 1.0 implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one reporting entity from another. A score of 0.7 or higher indicates adequate reliability to distinguish performance between two entities and is considered acceptable. The testing results suggest that this measure has adequate reliability for states and health plans, with very high reliability for Medicaid health plans and states in particular.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☒ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
Face Validity
The health-plan level of this measure was assessed for use in the HEDIS Health Plan Measure Set. As part of this process, NCQA assessed the face validity of the measure using its HEDIS process. NCQA staff shared the measure concepts, supporting evidence and field test results with its standing Behavioral Health Measurement Advisory Panel, Technical Measurement Advisory Panel and additional panels. We posted the measures for Public Comment, a 30-day period of review that allowed interested parties to offer feedback about the measure. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle. This measure has undergone the following steps associated with that cycle.

Step 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA's Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

Step 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

Step 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQAs Board of Directors will be included in the next HEDIS year and reported as first-year measures.

Step 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

Empirical Validity

As part of field testing, we assessed construct validity, which considers whether measures are capturing important aspects of a quality concept. We conducted two types of analyses: correlations among measures and rankings of health plans and states on measures.

We tested for construct validity by exploring whether this measure was correlated with two related measures: *Metabolic Screening for Children and Adolescents on Antipsychotics* and *Follow-Up Visit for Children and Adolescents on Antipsychotics* measures. The *Metabolic Screening* measure assesses the percentage of youth who undergo metabolic testing prior to or immediately after the start of a new antipsychotic prescription, and the *Follow-Up Visit* measure assesses the percentage of youth who have one or more visits with a prescriber within 30 days after the start of a new antipsychotic prescription. A higher rate indicates better performance for all three measures.

We hypothesized that organizations that perform well on this measure should perform well on the other measures. We calculated correlations using the Spearman correlation coefficient, which estimates the strength of the linear association between two continuous variables. The magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable are associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable are associated with decreasing values of the second variable.

We then explored whether entities that manage use of first-line psychosocial care well also manage other aspects of antipsychotic-related care well. We looked to see if plans and states can be approximately ranked based on profiles of performance across multiple measures. Consistency of performance across measures suggests that the measures are assessing a dimension of quality.

ICD-10 Conversion

Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

Steps in ICD-9 to ICD-10 Conversion Process

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use GEM to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

Tools Used to Identify/Map to ICD-10
All tools used for mapping/code identification from CMS ICD-10 website (http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html).
GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

Expert Participation
The NCQA HEDIS Expert Coding Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under **Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development**.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Face Validity Results
Step 1: This measure was developed to address the need for first-line psychosocial care for those youth started on antipsychotics who do not have a primary indication for an antipsychotic. NCQA and five expert panels worked together in 2013 and 2014 to identify the most appropriate method for assessing first-line psychosocial care among this patient population. Across the multiple expert panels that reviewed this measure, all panels concluded this measure was specified to assess the use of psychosocial care as first-line treatment for children without a primary indication for antipsychotics.

Step 2: The measure was written and field-tested in 2013 and 2014. After reviewing field test results, the CPM recommended to send the measure to public comment with a majority vote in January 2014.

Step 3: The measure was released for Public Comment in 2014 prior to publication in HEDIS. Of 73 comments received, the vast majority (80 percent) supported it as-is or with suggested modifications. The CPM recommended moving this measure to first year data collection by a majority vote in May 2014.

Step 4: The measure was introduced in HEDIS 2015. Organizations voluntarily reported this measure in the first year (2014) and the results were analyzed for public reporting in the following year (2015). The measure was approved in September 2015 by the CPM for public reporting in HEDIS 2016 for Medicaid and commercial plans.

Empirical Validity Results
Correlations
When determining correlations among measures, we focused on health plans, as there were not enough entities to measure correlations with the state data.

Among national commercial plans, there was moderate positive correlation between the *Follow-Up Visit* and *Psychosocial Care* measures (r=0.59, p=0.03) and very slight positive correlation between the *Metabolic Screening* and *Psychosocial Care* measures (r=0.18, p=.55).

Measure	**Pearson Correlation Coefficients**		
	First-Line Psychosocial Care	Follow-Up Visit	Metabolic Screening
First-Line Psychosocial Care	1	0.59	0.18
Follow-Up Visit		1	0.06
Metabolic Screening			1

Ranking

Among MAX states and one state's Medicaid plans, we found good consistency in the states and plans, respectively, with the best and worst performance.

MAX State Performance Rankings

State	**First-Line Psychosocial Care**	**Follow-Up Visit**	**Metabolic Screening**
1	**36.7**	**60.2**	**2.6**
2	**35.8**	**68.4**	4.5
3	60.3	75.0	5.5
4	48.9	71.2	3.8
5	45.0	74.9	**0.4**
6	64.1	76.4	4.8
7	41.5	69.0	6.3
8	N/A*	N/A*	5.3
9	N/A*	N/A*	**10.7**
10	**53.3**	**81.3**	8.3
11	N/A*	**78.8**	**14.0**
Mean	**48.2**	**72.8**	**6.0**

*State was excluded from analysis due to incomplete data

Medicaid Health Plan Performance Rankings for One State

Plan	**First-Line Psychosocial Care**	**Follow-Up Visit**	**Metabolic Screening**
3	41.7	**71.0**	**0.2**

9	48.6	81.8	**4.9**
6	30.1	83.5	12.3
17	**26.4**	**86.7**	14.8
2	**27.4**	80.5	**15.4**
8	43.5	81.1	12.6
4	46.9	78.7	9.3
5	42.4	80.0	10.6
1	51.6	82.1	12.8
11	43.8	74.4	6.1
16	56.6	78.8	10.6
15	28.0	80.9	10.8
12	43.3	77.2	13.3
13	30.7	**70.4**	**17.8**
7	**67.7**	85.3	5.1
14	64.3	**98.7**	7.1
10	**67.0**	78.9	10.6
Mean	**44.7**	**80.6**	**10.3**

ICD-10 Conversion Results

Summary of Stakeholder Comments Received

NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. NCQA received comments from four organizations:

- Support recommendations.
- Questions about select codes.
- Recommended additional codes for consideration.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e.*, *what do the results mean and what are the norms for the test conducted?*)

Face Validity

The expert panels consulted showed good agreement that the measure as specified will accurately differentiate quality across states and health plans. Our interpretation of these results is that this measure has sufficient face validity.

Empirical Validity

Correlations

Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. The results indicate that commercial plans that performed well on providing follow-up visits for those newly prescribed antipsychotics also performed well on providing first-line psychosocial care to those newly on antipsychotics. There was also a very slight positive correlation between

the *Psychosocial Care* measure and *Metabolic Screening* measures, indicating that plans that perform well on providing first-line psychosocial care also perform well on providing baseline metabolic screening for those newly prescribed antipsychotics.

Ranking

The results show that plans and states can be approximately ranked based on profiles of performance across multiple measures. The consistent performance across these measures suggest the measures are assessing a common dimension of quality.

2b3. EXCLUSIONS ANALYSIS

☐ **no exclusions — *skip to section 2b4***

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

The measure excludes youth with conditions for which there is a U.S. Food and Drug Administration indication for antipsychotics (schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders). We tested the impact of exclusions using the commercial health plan data. The aim of testing exclusions in the field test data was to determine how common exclusions are in the eligible patient population and the impact of these exclusions on denominator sizes and performance rates. Our results (detailed below) show differences in performance rates with and without exclusions.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

On average 25% of children age 0-5 with a new start of an antipsychotic met the exclusion criteria for having a primary indication for antipsychotic use (i.e., schizophrenia, bipolar disorder, psychotic disorder, autism, tic disorders); 29% of children age 6-11 met the exclusion criteria; 25% of adolescents age 12-17 met the exclusion criteria. The application of the exclusion to the measure reduced rates on average across plans by less than 2% for those age 0-5, increased rates by less than 2% for those age 6-11 and increased rates by just over 2% for those age 12-17 (see Table below).

Exclusion for Diagnosis during Measurement Year that has FDA Indication for Antipsychotics

	Age 0-5			Age 6-11			Age 12-17		
	# with exclusion diagnosis	Rate with exclusion	Rate without exclusion	# with exclusion diagnosis	Rate with exclusion	Rate without exclusion	# with exclusion diagnosis	Rate with exclusion	Rate without exclusion
Plan 1	0	25.0%	25.0%	53	63.2%	64.9%	102	73.6%	73.8%
Plan 2	3	75.0%	42.9%	8	76.9%	70.6%	23	77.8%	79.1%
Plan 3	4	33.3%	23.1%	26	35.2%	42.1%	59	54.2%	55.9%
Plan 4	0	NA	NA	4	76.5%	76.2%	13	64.0%	65.8%
Plan 5	0	NA	NA	12	75.0%	66.7%	23	65.8%	69.6%
Plan 6	2	25.0%	50.0%	9	59.4%	58.5%	21	51.2%	55.2%

Plan 7	0	100.0%	100.0%	7	64.3%	71.4%	23	57.4%	68.6%
Plan 8	0	0.0%	0.0%	4	61.5%	70.6%	20	67.8%	72.2%
Plan 9	0	50.0%	50.0%	6	18.2%	35.3%	15	50.0%	52.7%
Plan 10	0	0.0%	0.0%	10	66.7%	60.0%	10	69.8%	71.4%
Plan 11	1	60.0%	66.7%	4	81.8%	66.7%	17	70.7%	70.7%
Plan 12	1	33.3%	25.0%	4	54.5%	66.7%	18	57.1%	60.9%
Plan 13	1	NA	100.0%	5	71.4%	58.3%	10	50.0%	47.5%
Total	12	40.0%	38.3%	152	57.7%	59.5%	354	63.7	66.0%

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
The exclusions in this measure are designed to focus the measure on children in whom psychosocial care is recommended as first-line treatment. The exclusions did not adversely impact the denominator of the measure. Because the exclusions can be collected administratively, they do not pose an undue burden.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR). The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Variation in Performance Rates across 8 MAX States (2008 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
48.2	36.4	37.9	46.9	58.6	61.5	20.7

IQR: Interquartile range

Variation in Performance Rates across 17 Medicaid Plans from one State (2010 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
44.7	27.8	30.4	43.5	54.1	65.4	23.7

IQR: Interquartile range

Variation in Performance Rates across 13 Commercial Plans Nationwide (2012 data)

Mean Rate	10th	25th	50th	75th	90th	IQR
61.8	49.4	53.3	65.8	69.8	71.7	16.5

IQR: Interquartile range

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
The results show that there is a 23.7% gap in performance between Medicaid plans at the 25th and 75th percentiles, a 16.5% gap in performance among commercial plans and a 20.7% gap in performance among states at the 25th and 75th percentiles. This means that states at the 25th percentile have on average 172 less children and adolescents getting recommended first-line psychosocial care than states at the 75th percentile. For Medicaid plans, those at the 25th percentile have on average 119 less children and adolescents getting recommended first-line psychosocial care than plans at the 75th percentile. For commercial plans, those at the 25th percentile have on average 18 less children and adolescents getting recommended first-line psychosocial care than plans at the 75th percentile.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
States and plans collect this measure using all administrative data sources, for all intents and purposes, there are no missing data in administrative data. We have done no assessment to look for the distribution of missing data. For plans reporting on this measure for HEDIS, NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

N/A

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Field testing results, more fully described in the Scientific Acceptability section, showed the measure is feasible to be collected by health plans and states using administrative claims data. Further, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the managed care organization´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications

5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures. Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Quality Compass http://www.ncqa.org/tabid/177/Default.aspx The State of Health Care Quality Report http://www.ncqa.org/tabid/836/Default.aspx Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx The State of Health Care Quality Report http://www.ncqa.org/tabid/836/Default.aspx

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

QUALITY COMPASS: This measure has just been approved for use in Quality Compass, a tool that displays health plan-level

performance rates for HEDIS measures. It is used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. The Quality Compass 2015 Commercial tool includes data for 400 public reporting commercial health plan products, serving approximately 103.5 million covered lives. Benchmarks are calculated from a total pool of 420 public and non-public reporting health plan products, serving approximately 104 million covered lives. The Quality Compass 2015 Medicaid tool includes data for 182 public reporting Medicaid health plan products, serving approximately 20 million covered lives. Benchmarks are calculated from a total pool of 244 public and non-public reporting health plan products, serving approximately 25 million covered lives.

STATE OF HEALTH CARE QUALITY REPORT: HEDIS measures are reported nationally and by geographic regions in the State of Health Care Quality Report, published by NCQA and summarizing findings on quality of care. In 2015, the report included measures on 15.4 million Medicare Advantage beneficiaries in 507 Medicare Advantage health plans, 103.9 million members in 413 commercial health plans, and 25.4 million Medicaid beneficiaries in 237 plans across 50 states.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No negative consequences have been reported since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same

target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
2337 : Antipsychotic Use in Children Under 5 Years Old

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
This new measure assesses receipt of psychosocial care among children and adolescents who are prescribed antipsychotics without a primary indication. Both measures address use of antipsychotics. However, 2337 assesses if children under 5 are prescribed an antipsychotic. Our Psychosocial Care measure assesses children of a broader age range (up to age 18) who are currently on antipsychotics but do not have a primary indication. Our measure also addresses a different focus: whether these children received first-line psychosocial care.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee on Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-

Co.3 **Measure Developer if different from Measure Steward:** National Committee on Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
NCQA Behavioral Health Measurement Advisory Panel
Bruce Bobbitt, PhD, LP, Optum
Peter Delany, PhD, LCSW-C, Substance Abuse and Mental Health Services Administration
Ben Druss, MD, MPH, Emory University
Frank A. Ghinassi, PhD, ABPP, Western Psychiatric Institute and University of Pittsburgh Medical Center
Rick Hermann, MD, Tufts Medical Center and UpToDate, Inc.
Connie Horgan, ScD, Brandeis University
Neil Korsen, MD, Maine Health
Charlotte Mullican, BSW, MPH, Agency for Healthcare Research and Quality
Harold Pincus, MD, Columbia University and RAND Corporation
Bruce L. Rollman, MD, MPH, University of Pittsburgh School of Medicine
Michael Schoenbaum, PhD, National Institute of Mental Health
John H. Straus, MD, Massachusetts Behavioral Health Partnership and Beacon Health Options

NCQA Committee on Performance Measurement (CPM)
Bruce Bagley, MD, American Academy of Family Physicians
Andrew Baskin, MD, Aetna
Patrick Conway, MD, MMSc, Center for Medicare & Medicaid Services
Jonathan D. Darer, MD, Geisinger Health System
Helen Darling, National Business Group on Health
Rebekah Gee, MD, MPH, FACOG, LSU School of Medicine and Public Health
Foster Gesten, MD, NYSDOH Office of Managed Care
David Grossman, MD, MPH, Group Health Physicians
Christine Hunter, MD (Co-Chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Bernadette Loftus, MD, The Permanente Medical Group
J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, RN, PhD, AARP
Eric C. Schneider, MD, MSc (Co-Chair), RAND Corporation
Marcus Thygeson, MD, MPH, Blue Shield of Califorina

NCQA HEDIS Expert Coding Panel
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Health Services Advisory Group
Nelly Leon-Chisen, RHIA, American Hospital Association
Tammy Marshall, LVN, Aetna
Alec McLure, RHIA, CCS-P, Verisk Health
Michele Mouradian, RN, BSN, McKesson Health Solutions
Craig Thacker, RN, CIGNA HealthCare
Mary Jane F. Toomey, RN CPC, Aetna Better Health

NCQA Technical Measurement Advisory Panel
Andy Amster, MSPH, Kaiser Permanente
Kathryn Coltin, MPH, Independent Consultant
Lekisha Daniel-Robinson, Centers for Medicare and Medicaid Services
Marissa Finn, MBA, Cigna HealthCare

Scott Fox, MS, MEd, Independence Blue Cross
Carlos Hernandez, CenCalHealth
Kelly Isom, MA, RN, Aetna
Harmon Jordan, ScD, RTI International
Ernest Moy, MD, MPH, Agency for Healthcare Research and Quality
Patrick Roohan, New York State Department of Health
Lynne Rothney-Kozlak, MPH, Rothney-KozlakConsulting, LLC
Natan Szapiro, Independence Blue Cross

National Collaborative for Innovation in Quality Measurement (NCINQ) Measurement Advisory Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Katie Brookler, Colorado Department of Health Care Policy and Financing
Cathy Caldwell, MPH, Alabama Department of Public Health
Jennifer Havens, MD, NYU School of Medicine
Ted Ganiats, MD, University of California, San Diego
Darcy Gruttadaro, JD, National Allegiance on Mental Illness
Virginia Moyer, MD, MPH, FAAP, Baylor College of Medicine, USPSTF
Edward Schor, MD, Lucile Packard Foundation for Children's Health
Xavier Sevilla, MD, FAAP, Whole Child Pediatrics
Gwen Smith, Illinois Department of Healthcare and Family Services/Health Management Associates
Janet (Jessie) Sullivan, MD, Hudson Health Plan
Kalahn Taylor-Clark, PhD, MPH, George Mason University
Craig Thiele, MD, CareSource
Charles Wibbelsman, MD, Kaiser Permanente Medical Group, Inc.
Jeb Weisman, PhD, Children's Health Fund

NCINQ Consumer Panel
Joan Alker, MPhil, Georgetown Center for Children and Families
Roni Christopher, MEd, OTR/L, PCMH-CCE, The Greater Cincinnati Health Collaborative
Daniel Coury, MD, Nationwide Children's Hospital
Eileen Forlenza, Colorado Medical Home Initiative, Children and Youth with Special Health Care Needs Unit
Michaelle Gady, JD, Families USA
Janis Guerney, JD, Family Voices
Jocelyn Guyer, MPA, Georgetown Center for Children and Families
Catherine Hess, MSW, National Academy for State Health Policy
Carolyn Muller, RN, Montgomery County Health Department
Cindy Pellegrini, March of Dimes
Judith Shaw, EdD, MPH, RN, VCHIP
Stuart Spielman, JD, LLM, Autism Speaks
Michelle Sternthal, PhD, March of Dimes

NCINQ Foster Care Panel
Kamala Allen, MHS, Center for Health Care Strategies
Mary Applegate, MD, Ohio Department of Job and Family Services
Samantha Jo Broderick, Foster Care Alumni of America
Mary Greiner, MD, Cincinnati Children's Hospital Medical Center
David Harmon, MD, FAAP, Superior HealthPlan
Patricia Hunt, Magellan Health Services
Audrey LaFrenier, MSW, Parsons Child and Family Center
Bryan Samuels, MPP, Chapin Hall
Phil Scribano, DO, MSCE, The Children's Hospital of Philadelphia
Lesley Siegel, MD, State of Connecticut Department of Children and Families
Chauncey Strong, MSW, LGSW, Fairfax County Department of Family Services/Foster Care and Adoption
Janet (Jessie) Sullivan, MD, Hudson Health Plan
Nora Wells, MS, National Center for Family/Professional Partnerships

NCINQ Mental Health Panel

Francisca Azocar, PhD, Optum Health Behavioral Solutions
Frank Ghinassi, PhD, Western Psychiatric Institute and Clinic of UPMC Presbyterian Shadyside
Jennifer Havens, MD, NYU Langone Medical Center
Danielle Laraque, MD, FAAP, Maimonides Infants and Children's Hospital of Brooklyn

NCINQ State Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Sharon Carte, MHS, State of West Virginia Children's Health Insurance Program
Susan Castellano, Minnesota Department of Human Services
Catherine Hess, MSW, National Academy for State Health Policy
Michael Hogan, PhD, New York State office of Mental Health
Barbara Lantz, MN, RN, State of Washington Department of Social and Health Services, Medicaid Purchasing Administration
Judy Mohr Peterson, PhD, Oregon Health Authority
Tracy Plouck, MPA, Ohio Department of Mental Health
Gina Robinson, Colorado Department of Health Care Policy and Financing
Janet Stover, Illinois Association of Rehabilitation Facilities
Eric Trupin, PhD, University of Washington

NCINQ Measure Development Partners
Shahla Amin, MS, Rutgers University
Scott Bilder, PhD, Center for Health Services Research, Rutgers University
Stephen Crystal, PhD, Institute for Health, Health Care Policy and Aging Research, Rutgers University
Molly Finnerty, PhD, NY State Office of Mental Health
Emily Leckman-Westin, PhD, NY State Office of Mental Health
Sheree Neese-Todd, MA, Institute for Health, Health Care Policy and Aging Research, Rutgers University

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2014

Ad.3 **Month and Year of most recent revision:**

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines change significantly.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2016

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2802
Measure Title: Overuse of Imaging for the Evaluation of Children with Post-Traumatic Headache
Measure Steward: Q-METRIC – The University of Michigan
Brief Description of Measure: Percentage of children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the emergency department (ED) within 24 hours after an injury, and imaging of the head (computed tomography [CT] or magnetic resonance imaging [MRI]) was obtained in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or basilar skull fracture.
Developer Rationale: Post-traumatic headaches in children are a common clinical presentation in the setting of concussion and mild traumatic brain injury. In the United States, it has been estimated that more than 500,000 children younger than 15 years of age were evaluated in an ED following mild traumatic brain injury each year from 1998 to 2000 (Bazarian et al., 2005). Over the past decade, ED visits for traumatic brain injuries have increased substantially (Coronado et al., 2015).

Well-established evidence shows that neuroimaging to evaluate children with post-traumatic headache in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or skull fracture is rarely clinically indicated and is potentially harmful (Kuppermann et al., 2009; Lateef et al., 2009; Lateef et al., 2012; ACR Expert Panel on Pediatric Imaging, Ryan et al., 2014). The American Academy of Pediatrics Choosing Wisely initiative includes guidance to discourage the unnecessary use of CT scans for the immediate evaluation of minor head injuries and encourage reliance on clinical observation/PECARN criteria to determine whether imaging is indicated (AAP Choosing Wisely, 2013; Kuppermann et al., 2009).

CT use has increased in the past 20 years. In a cross-sectional analysis of data from the National Hospital Ambulatory Medical Care Survey, Blackwell et al. (2007) found the use of CT scans for the evaluation of children with head injury nearly doubled from 1995 to 2003 (13% to 22%); Zonfrillo et al. (2015) found evidence to suggest continued increases in CT use for ED patients with concussion from 2006 to 2011. Some research suggests that rates of imaging following head injury appear to have declined in free-standing children's hospitals (Menoch et al., 2012; Mannix et al., 2012; Parker et al., 2015) and general EDs (Marin et al., 2014). Also, CT rates for children with mild head trauma vary widely between hospitals. CT rates ranged from 19% to 69% across 25 EDs (Stanley et al., 2014). Similarly, CT rates ranged from 19% to 58% for patients with minor head injury in a retrospective analysis of 5 years of hospital administrative data from 40 free-standing children's hospitals (Mannix et al., 2012).

Overuse has been defined as any patient who undergoes a procedure or test for an inappropriate indication (Lawson et al., 2012). Imaging overuse for the evaluation of children with post-traumatic headaches without signs or symptoms of intracranial injury subjects children to a number of risks (Malviya et al., 2000; Mathews et al., 2013; Pearce et al., 2012; Wachtel et al., 2009). Individuals who undergo CT scans in early childhood tend to be at greater risk for developing leukemia, primary brain tumors, and other malignancies later in life (Mathews et al., 2013; Pearce et al., 2012). Children are also at risk for complications from sedation or anesthesia, which are often required for longer CT imaging sequences and for MRI, and from intravenous contrast media (Zo'o et al., 2011). Cost is also an issue (Callaghan et al., 2014) that burdens the patient, as well as payers.

Citations:
American Academy of Pediatrics (AAP). Choosing Wisely: An initiative of the ABIM Foundation. Ten Things Physicians and Patients Should Question. 2013. Available at: http://www.choosingwisely.org/doctor-patient-lists/american-academy-of-pediatrics/; accessed: February 24, 2015.

American College of Radiology Expert Panel on Pediatric Imaging: Ryan ME, Palasis S, Saigal G, et al. ACR Appropriateness Criteria: Head trauma — child. American College of Radiology, 2014. Available at: https://acsearch.acr.org/docs/3083021/Narrative/;

accessed July 1, 2015.

Bazarian JJ, McClung J, Shah MN, Cheung YT, Flesher W, Kraus J. Mild traumatic brain injury in the United States, 1998-2000. Brain Inj 2005; 19(2):85-91.

Blackwell CD, Gorelick M, Holmes JF, Bandyopadhyay S, Kuppermann N. Pediatric head trauma: Changes in use of computed tomography in emergency departments in the United States over time. Ann Emerg Med 2007; 49(3):320-324.

Callaghan BC, Kerber KA, Pace RJ, Skolarus LE, Burke JF. Headaches and neuroimaging: High utilization and costs despite guidelines. JAMA Intern Med 2014; 174(5):819-821.

Coronado VG, Haileyesus T, Cheng TA, et al. Trends in sports- and recreation-related traumatic brain injuries treated in US emergency departments: The National Electronic Injury Surveillance System-All Injury Program (NEISS_AIP) 2001-2012. J Head Trauma Rehabil 2015; 30(3): 185-197.

Kuppermann N, Holmes JF, Dayan PS, et al., Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study. Lancet 2009; 374: 1160–1170.

Lateef TM, Grewal M, McClintock W, Chamberlain J, Kaulas H, Nelson KB. Headache in young children in the emergency department: Use of computed tomography. Pediatrics 2009; 124:1 e12-e17.

Lateef TM, Kriss R, Carpenter K, Nelson KB. Neurologic complaints in young children in the ED: When is cranial computed tomography helpful? Am J Emerg Med 2012; 30(8):1507-1514.

Lawson EH, Gibbons MM, Ko CY, Shekelle PG. The appropriateness method has acceptable reliability and validity for assessing overuse and underuse of surgical procedures. J Clin Epidemiol 2012; 65(11):1133-1143.

Malviya S, Voepel-Lewis T, Eldevik OP, Rockwell DT, Wong JH, Tait AR. Sedation and general anesthesia in children undergoing MRI and CT: Adverse events and outcomes. Br J Anaesth 2000; 84(6):743-748.

Mannix R, Meehan WP, Monuteaux MC, Bachur RG. Computed tomography for minor head injury: Variation and trends in major United States emergency departments. J Pediatr 2012; 160:136-139.

Marin JR, Weaver MD, Barnato AE, Yabes JG, Yealy DM, Roberts MS. Variation in emergency department head computed tomography use for pediatric head trauma. Acad Emerg Med 2014; 21(9):987-995.

Mathews JD, Forsythe AV, Brady Z, et al. Cancer risk in 680,000 people exposed to computed tomography scans in childhood or adolescence: Data linkage study of 11 million Australians. BMJ 2013; 346:f2360.

Menoch MJ, Hirsh DA, Khan NS, Simon HK, Sturm JJ. Trends in computed tomography utilization in the pediatric emergency department. Pediatrics 2012; 129(3):e690-e697.

Parker MW, Shah SS, Hall M, Fieldston ES, Coley BD, Morse RB. Computed tomography and shifts to alternate imaging modalities in hospitalized children. Pediatrics 2015; 136(3):e573-e581.

Pearce MS, Salotti JA, Little MP. Radiation exposure from CT scans in childhood and subsequent risk of leukemia and brain tumours: A retrospective cohort study. Lancet 2012; 380(9840): 499–505.

Stanley RM, Hoyle JD Jr, Dayan PS, et al. Emergency department practice variation in computed tomography use for children with minor blunt head trauma. J Pediatr 2014; 165(6):1201-1206.

Wachtel RE, Dexter F, Dow AJ. Growth rates in pediatric diagnostic imaging and sedation. Anesth Analg 2009; 108(5):1616-1621.

Zonfrillo MR, Kim KH, Arbogast KB. Emergency department visits and head computed tomography utilization for concussion patients from 2006 to 2011. Acad Emerg Med 2015; 22(7):872-877.

Zo'o M, Hoermann M, Balassy C, et al. Renal safety in pediatric imaging: Randomized, double blind phase IV clinical trial of iobitridol

300 versus iodixanol 270 in multidetector CT. Pediatr Radiol 2011; 41(11); 1393-1400.

Numerator Statement: The number of numerator eligible children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or basilar skull fracture.
Denominator Statement: The number of children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained in the absence of suspected child abuse or neglect or a history of a medical condition that would otherwise warrant neuroimaging.
Denominator Exclusions: Children under evaluation for child abuse or neglect and children with a history of a medical condition that could otherwise warrant neuroimaging (e.g., bleeding disorder, intracranial tumor, hydrocephalus) for the evaluation of a post-traumatic headache were excluded from this overuse measure.

Children with a diagnosis of headache without a documented history of trauma and children with a diagnosis of concussion without documentation of headache as a symptom were excluded because post-traumatic headache is the focus of this measure.

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

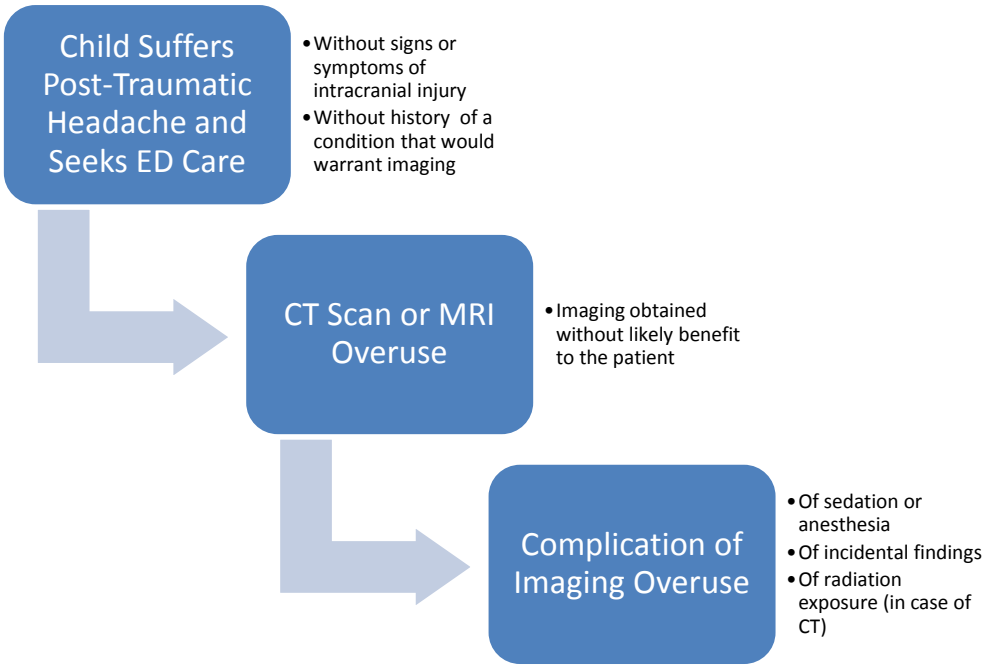
The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this process measure:



- The American College of Radiology Appropriateness Criteria states that CT scans and MRI of the head in children older than two years with minor head injury and without neurologic signs or high risk factors has been rated Category 3 or lower for appropriateness. Categories 1, 2 and 3 are considered "usually not appropriate," where the harms of the procedure outweigh the benefits.
    - The evidence is based on 47 Review/Other-Diagnostic Studies and 21 Observational-Diagnostic Studies. Of the 68 studies, evidence ranges from 5 studies in category 2 (moderately well-designed study that accounts for most common biases), 14 studies in Category 3 (study that has important design limitations) and 49 studies in Category 4 **(**not useful as primary evidence, may not be a clinical study or the study design is invalid, or conclusions are based on expert consensus)
    - The largest study, the Pediatric Emergency Care Applied Research Network (PECARN) head imaging clinical decision rule for children with mild traumatic brain injury has 99.9% negative predictive value and 96.8% sensitivity for predicting clinically important injury. The PECARN study provides evidence that imaging was overused in approximately 20% of the study population 2 years and older who demonstrated none of the six predictors comprising the decision rule.
    - The developers state that the main benefit of reducing neuroimaging among children with post-traumatic headache relates to the avoidance of harms. The potential risks and harms associated with imaging include radiation exposure (Pearce et al., 2012; Mathews et al., 2013); complications from sedation and/or anesthesia (Malviya et al., 2000; Wachtel et al., 2009); incidental findings leading to potentially invasive and costly follow-up testing (Lumbreras et al., 2010; Rogers et al., 2013); and excess costs to the healthcare system, which are passed on to families (Callaghan et al., 2014).

Questions for the Committee:

- *Is the evidence directly applicable to the process of care being measured?*
- *Has the developer provided sufficient evidence between the relationship of this measure to patient outcomes?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following information related to performance gap and disparities:

- In the United States, it has been estimated that more than 500,000 children younger than 15 years were evaluated in an emergency department (ED) following mild traumatic brain injury each year from 1998 to 2000 (Bazarian et al., 2005). Over the past decade, ED visits for traumatic brain injuries have increased substantially (Coronado et al., 2015). CT rates for children with mild head injury ranged from 19% to 69% across the 25 EDs that collected data for the PECARN study.
- The developer states that evidence shows that neuroimaging to evaluate children with post-traumatic headache in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or skull fracture is rarely clinically indicated and is potentially harmful.
- Overuse has been defined as any patient who undergoes a procedure or test for an inappropriate indication (Lawson et al., 2012).
- The developer reports 8 of 57 children (14%) were imaged in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrahage or skull fracture, indicating an opportunity to reduce overuse of neuroimaging among children with post-traumatic headache. Specifically:
    - 204 charts were reviewed, and 57 (27.9%) met denominator criteria: children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained.
    - The developer was unable to assess plan, hospital emergency department, and provider level variations based on the limited number of eligible medical records that were available for calculation of this measure following chart review.
- The small numbers of eligible numerator and denominator cases (n=8 and n=57, respectively) did not allow for

meaningful comparisons of overuse of neuroimaging among children with post-traumatic headache evaluated in EDs across different socio-demographic groups. The developer notes however:
    - On average, children with post-traumatic headache who obtained neuroimaging resided in ZIP codes reporting primarily white race (80.2%) and modest levels of Hispanic ethnicity (9.8%); the median household income for the ZIP-codes in which these children resided was substantially higher than the median household income of the population of the entire United States
    - Children with post-traumatic headache who obtained neuroimaging primarily reside in urban ZIP codes.
- PECARN found that children of black non-Hispanic or Hispanic race/ethnicity had lower odds of undergoing head CT than white non-Hispanic children. Parental anxiety and parental request were cited as reasons for ordering head CT in children of white, non-Hispanic race/ethnicity.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- The data linking overuse with significant non-monetary outcomes is largely theoretical. While one might accept later life-malignancies or sedation complications as important harms, the actual direct evidence remains somewhat scanty. Need more detail on underlying decision rule and better discussion of varied performance rates in different populations. What are factors influencing NPV, underlying prevalence of important neurologic diagnoses?
- The evidence provided for this process measure is based on a systematic review with grading of the empirical evidence included. The process measure is aimed at the avoidance of harm by reducing neuroimaging which is related to the process of care being measured. Although there is significant evidence to support reducing/limiting neuroimaging in children with minor head injury without neurologic signs, the relationship of neuroimaging to potential risks could be strengthened.
- The measure attempts to reduce risk of radiation exposure that can increase the risk of malignancy. The measure also attempts to reduce health care costs by reducing unneeded imaging (CT and MRI) The population comprises children who have had traumatic brain injury presenting to and ED with headache within 24 hours. The numerator is those in this group that are not otherwise excluded that had either a CT or MRI.
- Measure 2802 is a process measure with Moderate level of clinical evidence. Based on Algorithm 1, question 1 is scored as no, question 3 as yes based on ACR appropriateness criteria as a systematic review and the systematic review only scores imaging as a "3". In addition, Measure 2802 is not well aligned with the largest study (PECARN) on this topic. This lack of alignment likely reflects the difficulty of transforming a clinical decision rule into an operational measure.
- Addressing questions posed to the committee:
    - Re: evidence applicable to the process of care: The measure developer's provide compelling evidence that imaging is overused and that such overuse has consequences.
    - Re: relationship of this measure to patient outcomes: As detailed below, the current measure includes criteria (age >2, post-traumatic headache, absence of certain signs) that are not fully supported by the available evidence. These criteria become problematic when attempting to fully understand the measure, explain it to frontline teams and explain it to patients and their families.

1b. Performance Gap.

- Certainly is a gap/variance in performance. The data concerning disparities is interesting and worth exploring. Perhaps in this case access to advanced imaging is worse for certain groups thereby paradoxically improving outcomes (overuse).
- There is a significant number of children receiving care in Emergency Departments for mild traumatic brain injury, potentially warranting a national performance measure. The data reviewed by the developer did not yield sufficient numbers to determine differences in treatment across socio-demographic groups.
- Based on the PECARN data there is high confidence that performance gaps currently persist. However, based on the developer's measure testing using the HealthCore Integrated Research Database (HIRD), the small test sample did not demonstrate convincing evidence of a performance gap.
- Addressing questions posed to the committee:

- Re: warrants a national performance measure. Yes - In this reviewer's opinion, a measure is warranted but as detailed below, the proposed set of specifications and calculation algorithm is likely not suitable for a national performance measure.
- Re: evidence of disparities. This question can best be answered as more data on this topic is collected. Given evidence such as the PECARN study and Dartmouth Atlas showing that head CT rates vary substantially amongst different groups and regions, it is likely that disparities exist.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The numerator is the number of eligible children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or basilar skull fracture. The denominator is the number of children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained in the absence of suspected child abuse or neglect or a history of a medical condition that would otherwise warrant neuroimaging.
- The developer includes the ICD-9 and ICD-10 codes.
- The calculation algorithm is included. Data for the denominator population are derived from administrative data, followed by chart review and a straightforward calculation to achieve the performance rate.
- There is no risk adjustment.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

- Validity testing was performed at the data element level. Per NQF guidance, separate reliability testing of the data elements is not required if validity testing is conducted on the data elements.

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The measure is intended to assess overuse of CT or MRI in children in the ED. The evidence supports limiting use of CT/MRI in the absence of certain documented signs and symptoms. The specifications are consistent with the evidence.

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- Empirical validity testing was performed at the critical data element level, and the developer used face validity at the performance score level.
- The developer reports inter-rater reliability for the data abstracted from the medical record was very high. Of the 204 abstracted medical records, 30 (15%) were reviewed for IRR; percent agreement and kappa were calculated. Overall, abstractor agreement was 99.3% (kappa 0.98). The sensitivity of the abstractors to identify chart-based exclusions compared with the senior abstractor was 100% (95% CI; 94.6, 100); specificity was 99.5% (95% CI; 98.1, 99.9); positive predictive value was 97.1% (95% CI; 89.9, 99.7) and negative predictive value was 100% (95% CI; 99.0, 100.0).
- Face validity at the performance score level was performed by an expert panel of reviewers. The panel rated this a 7.0 (with 9.0 as highest) for relative importance and concluded that this measure would be able to distinguish good from poor quality care and could reduce unnecessary imaging.
- The developer states that it could exclude nearly all children with conditions that would require imaging through ICD-9 codes in administrative claims. Medical record abstraction is required, however, to appropriately document the exclusions and therefore to collect and report on the measure. Administrative claims alone are insufficient.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients meeting certain clinical conditions that require imaging and patients with suspected child abuse were excluded from the measure.
- The developer notes several exclusion criteria that should be applied to administrative claims to narrow the population eligible for chart review—i.e., there are claims-based exclusions (suspected abuse/neglect, history of a medical condition that could warrant neuroimaging, loss of consciousness, skull fracture, and intracranial hemorrhage) that reduced the denominator for chart review by 18.5% during testing.
- Chart review is required to assess additional appropriate numerator exclusions. The developer notes, for example, that no ICD-9-CM or ICD-10-CM codes exist for the major inclusion criteria of having an injury within 24 hours of the ED visit.
- The developer notes that without the chart review, overuse is over-estimated by 56 to 80 percentage points.

Questions for the Committee:

- *Are the denominator exclusions consistent with the evidence and codes as specified?*
- *Are the numerator exclusions consistent with the evidence and specifications?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed and outweigh the data collection burden of manual chart review?*

2b4. Risk adjustment:

- This process measure is not risk adjusted.

Questions for the Committee:

- *Should this measure be risk adjusted?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance*

measure scores can be identified):

The developer provides the following information:

- The small numbers of eligible numerator and denominator cases (n=8 and n=57, respectively) did not allow for meaningful comparisons of overuse of neuroimaging among children evaluated in EDs with post-traumatic headache across different socio-demographic groups.
- Due to the small sample size of charts eligible for inclusion, the developers were unable to compare health plans or hospital EDs.
- A two-sided two-proportion z-test was conducted to determine if the observed overuse percentage in the sample was statistically different than the observed rate of overuse within the 2009 PECARN study and a one-sided one-proportion z-test to determine if the observed overuse percentage in the sample was greater than a theoretical target overuse percentage of 5%.
    - The developer reports the following: "When comparing our overuse percentage (14.0%) with the rate of imaging in children lacking all of the six predictors of clinically important traumatic brain injury in the PECARN derivation sample (25.4%), a z-score of -2.0704 was obtained, corresponding to a two-sided p-value of 0.038. When comparing our overuse percentage (14.0%) with the rate of imaging in children lacking all of the six predictors of clinically important traumatic brain injury in the PECARN validation sample (24.1%), a z-score of -1.8339 was obtained, corresponding to a two-sided p-value of 0.067. When comparing our observed overuse percentage with the theoretical target overuse percentage of 5%, a z-score of 3.017 was obtained, corresponding to a one-sided p-value of 0.0013."
    - According to the developer, even with the small sample size, the measure was able to successfully distinguish statistically significant differences between this sample and the PECARN sample.
    - The developer states that "A minimum of 196 charts included in the denominator after chart review would be recommended to obtain a 95% confidence interval with a 5% half-width around an expected overuse percentage of 15%. A per group minimum of 335 charts included in the denominator after chart review would be recommended for a two-sample proportion test to detect a 10% difference from a control proportion of 15% with power of 0.90 and alpha of 0.05."

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable.

2b7. Missing Data

- No information was provided on missing data.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- As detailed above, both the numerator and denominator seem inconsistent with the prior evidence.
- Addressing questions posed to the committee:
    - Re: Are specifications consistent with the evidence? As stated above, the key issue is that the denominator examines children who underwent imaging when it seems a more appropriate indicator of imaging overuse would use a denominator of children presenting with comparable symptoms who did or did not undergo imaging.
- How do you clearly operationalize abuse/neglect?
- See above. I think that this measure could be re-specified to reduce the need for chart abstraction as well as to include others who might be inadvertently excluded.

2a2. Reliability testing

- Being able to specify a risk of child abuse and exclusion is subject to much bias.
- Data elements are clearly defined and appropriate codes ICD-9-CM and ICD-10-CM) included. The algorithm is clear lending to consistent implementation. The evidence provided supports the limitation of neuroimaging for minor

head trauma in the absence of documented signs and symptoms so the specifications are consistent with the evidence.
- Significant concerns about the fact that this measure will require chart review as many of the elements for both inclusion and exclusion will not necessarily be coded. In my clinical experience, children presenting with post-traumatic headache may not be coded as such.. they may simply be given diagnosis of head trauma. The headache and the associated clinical findings will more often be included in the progress note. If the measure limits the population to those with a coded diagnosis of post-traumatic headache it will exclude many.

2b1. Validity Specifications
- Validity - Pretty consistent but I would like to see a more robust sensitivity analysis.
- Addressing questions posed to the committee:
    - Re: Test sample adequate to generalize? No, out of a dataset of 60 million lives, the filtering algorithm found only 5912 children presenting to the ER with post-traumatic headache. This figure seems lower than would be predicted from CDC data on ER visits for head trauma in children. (http://www.cdc.gov/traumaticbraininjury/data/rates_ed_byage.html).
    - Further, of the 5912, 50% had CT or MR imaging, but only 2419 were eligible for chart review and only 1714 were considered for chart review. Of the 204 charts obtained, the majority (147) were had to be excluded at some point during the chart review process. Even if the entire 1714 charts were reviewed, the data suggest the majority would have been excluded at some point leaving only 478 available for review out of a population that started with 60 million lives. This would suggest a nationwide denominator of 2390 and a numerator of only 336.
    - Re: Sufficient validity so that conclusions about quality can be made? No, the small numbers and the difficulty linking this measure to the prior data on imaging in children with minor head trauma creates concerns about whether this measure can be used to assess quality of care.
    - Re: Score from this measure as an indicator of quality? No, as above.

2b2. Validity Testing
- I would want to see more information on the chart abstraction agreement.
- Empirical validity testing was performed at the critical data element level and showed high inter-rater reliability for data abstraction of 204 medical records. Face validity was performed at the performance score level using an expert panel of reviewers. Validity testing seems adequate to generalize for widespread implementation.

2b3-2b7. Threats to Validity
- Using Algorithm #3, the measure is rated as insufficient, since some potential threats to validity persist. The proposed measures fails the test of "ability to identify statistically significant and meaningful differences in performance".
- Addressing questions posed to the committee:
    - Re: Denominator definition and exclusions - These do not match those used in prior studies (eg PECARN)
    - Re: Numerator exclusions - decision to focus on children >2 and only those with post-traumatic headache reflect exclusions that were not used in the most important prior studies (eg PECARN)
    - Re: Inappropriate exclusions - the reliability and validity of the exclusions that occurred during chart reviews is uncertain.
    - Re: Frequency and variability of exclusions during chart review - the difficulty of standardizing chart review and the burden of performing these reviews also lower enthusiasm for this measure
- Addressing questions posed to the committee:
    - Re: Does this measure identify meaningful differences about quality? This reviewer believes the numbers from the test sample are simply too small and the measure algorithm too different from prior studies to draw meaningful conclusions.
- Clearly chart review is needed for this to be a valid measure. Risk adjustment is difficult to assess on the basis of the given data. I have not had an opportunity to review in detail the two large trials. There does seem to be the potential to identify true variation in population rates of imaging based on statistical analysis provided assuming the underlying construct is useful and valid.
- Exclusion criteria for both the numerator and denominator are appropriate, however the chart review required to assess numerator exclusions are labor intensive and challenging to implement.
- The exclusions are clinically relevant

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Data elements are in electronic sources (administrative data) to define the initial denominator population, some of the numerator, and some exclusions. Some information, in particular numerator exclusions, may be recorded in problem lists or in provider notes, requiring chart review.
  - During testing, the developer found that the majority of numerator exclusions (i.e., symptoms of intracranial injury that represent a clinical indication for neuroimaging) were not adequately captured in administrative claims.
  - The developer concludes that using administrative data alone would result in a substantial overestimate of the degree to which neuroimaging is overused in the evaluation of children with post-traumatic headache.
- Regarding sample size, the developer reports the following: To detect differences between two health plans, hospital emergency departments, or providers with overuse percentages of 20% and 10% would require a sample size of at least 199 denominator eligible cases per group with a p-value of 0.05 and 80% power.
- According to the developer, continuing advances in the development and implementation of EHRs may prompt providers to document key elements needed for application of inclusion and exclusion criteria necessary for this measure. Advances would further allow for electronic capture of structured clinical information needed to determine if and when neuroimaging has been overused in the evaluation of children experiencing a post-traumatic headache.
- This is not an eMeasure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Not all of the required data elements are available in electronic form at this time. How does this affect the feasibility of the measure?*
- *Is the data collection strategy ready to be put into operational use?*
- *Is the proposed sampling approach reasonable?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Rated as low
- Addressing questions posed to the committee:
  - Re: Data elements generated and used during care delivery: No and this markedly diminishes the feasibility of the measure
  - Re: Data collection strategy ready for operational use. No
  - Re: Is the proposed sampling approach reasonable. No - based on expanding the results found with a test sample to the entire nation, many ERs would have 0 or 1 children eligible for further assessment. Makes little sense to draw a sample from such small populations and it also would jeopardize the ability to obtain meaningful data from chart reviews since reviewers at any one site would have little experience and reviewers overseeing multiple sites would be viewing data embedded within each hospital's different recording structure/culture.
- Clearly, this would require significant resources in the form of chart review. While EHRS might eventually capture/codify data elements like suspicion of abuse, this is not a standard regularly implemented.
- Feasibility is limited as not all data elements are available in electronic form. This will make data collection burdensome and potential limit the operational use of the measure.
- This measure would require extensive chart review.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- The measure is not currently in use. The developer is currently submitting it to the National Quality Measures Clearinghouse.

- The developer identifies a three-year plan for implementation (public reporting) of the measure, with an additional three year-plan for updating and refining the measure.
- The developer did not identify any unintended consequences during measure testing.

***Questions for the Committee**:*

- *Is the implementation plan feasible?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Not implemented, but field testing could yield a lot of information. Are there untoward consequences/harms? Probably ok, but not well delineated.
- The measure is not currently in use and therefore not being publically reported. The performance results could be used to improve quality and efficiency of healthcare by limiting unnecessary imaging and avoiding potentially harmful consequences.
- Concerns that public reporting of overuse may be somewhat confusing to the general population. If the measure focused solely on CT, the overuse can be described as reducing unnecessary exposure to radiation. Overuse of MRI is primarily related to cost.

Criterion 5: Related and Competing Measures

- The measure is related to 0668: Appropriate Head CT Imaging in Adults with Mild Traumatic Brain Injury. This measure focuses on children 2-18 years; 0668 includes adolescents 16-18 years.
- The developer indicates that the measures are harmonized in terms of basic clinical criteria, but differ in a number of ways, including the inclusion of MRIs in the pediatric measure, different evidence on the needs of pediatric imaging, and the use administrative claims to narrow the population eligible for chart review.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Overuse of Imaging for the Evaluation of Children with Post-Traumatic Headache

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Imaging (CT or MRI) of children with post-traumatic headache who are evaluated in the emergency department (ED) within 24 hours after an injury, in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or skull fracture.

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

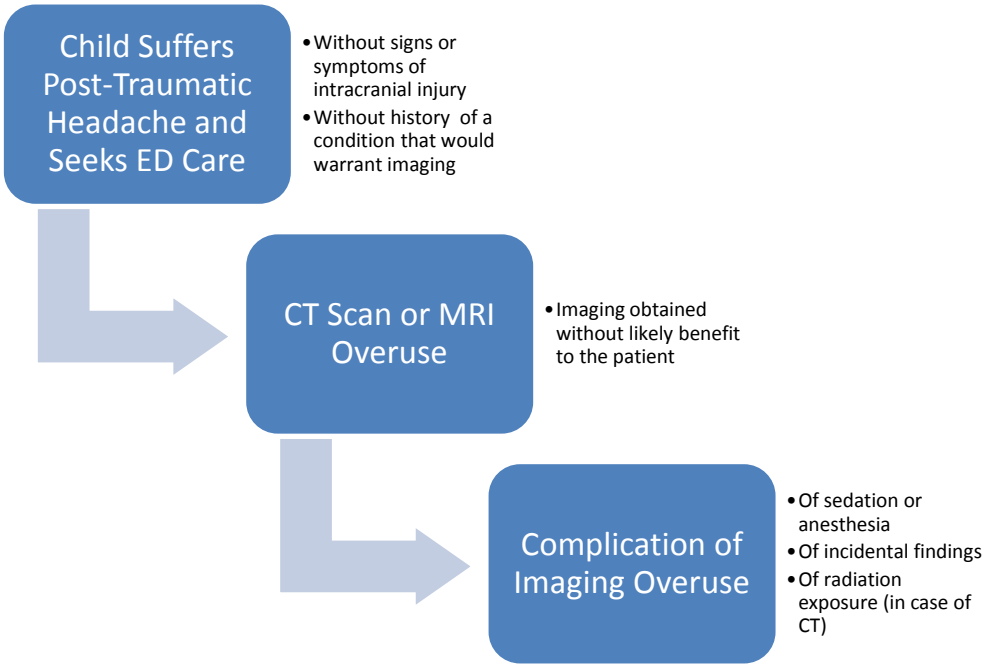
1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.



CT and MRI of the brain are the neuroimaging modalities at the center of this overuse measure. Both are radiologic modalities used to create images of internal structures in a slice-by-slice manner. CT uses X-ray radiation (hereafter simply called radiation), and MRI uses magnetic fields and radio waves.

Currently, professional guidelines do not support neuroimaging in children 2 years and older with minor head injury in the absence of neurologic signs or high risk factors indicative of intracranial injury (ACR Expert Panel on Pediatric Imaging, Ryan et al., 2014). Potential consequences of imaging overuse include complications of sedation or anesthesia, incidental findings, and radiation exposure. Therefore, measurement of overuse of neuroimaging with CT and MRI is an important quality indicator among children with post-traumatic headache following minor head injury.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

American College of Radiology Expert Panel on Pediatric Imaging: Ryan ME, Palasis S, Saigal G, et al. ACR Appropriateness Criteria: Head Trauma — Child. American College of Radiology, 2014.

URL: https://acsearch.acr.org/docs/3083021/Narrative/

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

The American College of Radiology (ACR) Appropriateness Criteria® (AC) are evidence-based guidelines to assist referring physicians and other providers in making the most appropriate imaging or treatment decision for a specific clinical condition. The AC assess the benefits and harms of recommended medical care or advanced diagnostic imaging options, using scientific evidence, to the extent possible, and clinical judgment and expert consensus, as necessary. The guidelines are developed by experts in diagnostic imaging, interventional radiology, and radiation oncology with participation from over 20 medical societies.

American College of Radiology
ACR Appropriateness Criteria®

Clinical Condition: **Head Trauma — Child**

Variant 1: **Minor head injury (GCS >13) ≥2 years of age without neurologic signs or high risk factors (eg, altered mental status, clinical evidence of basilar skull fracture). Excluding nonaccidental trauma.**

Radiologic Procedure	Rating	Comments	RRL*
CT head without contrast	3	This is a known low-yield procedure.	☢☢☢
MRI head without contrast	2		O
X-ray head	1		☢
CT head without and with contrast	1		☢☢☢☢
CT head with contrast	1		☢☢☢
CTA head with contrast	1		☢☢☢☢
MRI head without and with contrast	1		O
MRA head without contrast	1		O
MRA head without and with contrast	1		O
Arteriography cerebral	1		☢☢☢☢
US head	1		O
FDG-PET/CT head	1		☢☢☢☢
Tc-99m HMPAO SPECT head	1		☢☢☢☢
Rating Scale: 1,2,3 Usually not appropriate; 4,5,6 May be appropriate; 7,8,9 Usually appropriate			*Relative Radiation Level

Relative Radiation Level Designations		
Relative Radiation Level*	**Adult Effective Dose Estimate Range**	**Pediatric Effective Dose Estimate Range**
O	0 mSv	0 mSv
☢	<0.1 mSv	<0.03 mSv
☢☢	0.1-1 mSv	0.03-0.3 mSv
☢☢☢	1-10 mSv	0.3-3 mSv
☢☢☢☢	10-30 mSv	3-10 mSv
☢☢☢☢☢	30-100 mSv	10-30 mSv
*RRL assignments for some of the examinations cannot be made, because the actual patient doses in these procedures vary as a function of a number of factors (eg, region of the body exposed to ionizing radiation, the imaging guidance that is used). The RRLs for these examinations are designated as "Varies".		

Note: These tables are reproduced from pages 1 and 10 of the American College of Radiology (ACR) Expert Panel on Pediatric Imaging: Ryan ME, Palasis S, Saigal G, et al. ACR Appropriateness Criteria: Head Trauma — Child. American College of Radiology, 2014. Available at: https://acsearch.acr.org/docs/3083021/Narrative/ ; accessed June 30, 2015.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

CT scans and MRI of the head in children older than 2 years with minor head injury and without neurologic signs or high risk factors has been rated Category 3 or lower for appropriateness. Categories 1, 2 and 3 are considered "usually not appropriate," where the harms of the procedure outweigh the benefits.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

NA

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

The ACR panel members rate appropriateness based on the RAND Appropriateness Method (Fitch K. The Rand/UCLA appropriateness method user's manual. Santa Monica: Rand;2001).

URL: http://www.rand.org/content/dam/rand/pubs/monograph_reports/2011/MR1269.pdf

Each panel member assigns a rating; these are then are then presented to the group with the frequency distribution and the median group rating. Final ratings are determined using a modified Delphi method.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The specific service addressed was imaging of children with head trauma.

The body of evidence summarized in these responses is from the American College of Radiology Expert Panel on Pediatric Imaging: Ryan ME, Palasis S, Saigal G, et al. ACR Appropriateness Criteria: Head Trauma — Child. American College of Radiology, 2014. https://acsearch.acr.org/docs/3083021/Narrative/; accessed July 1, 2015.

Evidence Table URL: http://www.acr.org/~/media/6F3EEA65C42E47E7BCC529CDDCC77DB7.pdf

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

ACR staff determine if the following Study Quality Elements are described in each article included in the Evidence Tables that are presented with the Appropriateness Criteria:

1) Uncertainty measure
2) Prospective study
3) Systematic recruitment or recruitment of a consecutive series of patients
4) Standard of reference or comparison of two imaging tests
5) Reference standard applied
6) Independent readers of the imaging test
7) Index test results interpreted in a blinded fashion

The staff then counts the number of quality elements recorded as present in each article and assigns a Study Quality Category from 1 to 4. **Category 1** (well-designed study that accounts for common biases) must have all eight study quality elements present; **Category 2** (moderately well-designed study that accounts for most common biases) has six to seven quality elements present; **Category 3** (study that has important design limitations) has three, four or five quality elements present, and **Category 4 (**not useful as primary evidence, may not be a clinical study or the study design is invalid, or conclusions are based on expert consensus) has two or fewer quality elements present.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See answer in 1a.7.2.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1984-2013

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

47 Review/Other-Diagnostic Studies

21 Observational-Diagnostic Studies

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Of the 68 studies referenced in the evidence table, the assigned evidence quality grades range from Category 2 to Category 4. No studies were rated as Category 1; five studies were rated as Category 2 (three related to decision rules for imaging children with minor trauma, one related to severe head injury, and one related to imaging modalities for the evaluation of head injury); 14 studies were rated as Category 3 (eight related to epidemiology of head injury and imaging decision rules); and 49 studies were rated as Category 4.

The evidence supports that computed tomography is the primary imaging modality for children with acute traumatic brain injury and is overused for the evaluation of children. Numerous clinical decision rules have been put forth to reduce neuroimaging in children who have a low likelihood of intracranial injury requiring intervention.

The Pediatric Emergency Care Applied Research Network (PECARN) conducted the largest prospective study of children presenting to the ED within 24 hours of head injury and confirmed numerous prior lower quality studies that have documented low yield of neuroimaging of children with head injuries in the absence of signs or symptoms to suggest intracranial injury, as summarized in the Evidence Table published by the ACR Expert Panel on Pediatric Imaging, Head Trauma — Child (Ryan et al., 2014).

The PECARN head imaging clinical decision rule for children with mild traumatic brain injury has 99.9% negative predictive value and 96.8% sensitivity for predicting clinically important injury. The PECARN study provides evidence that imaging was overused in approximately 20% of the study population 2 years and older who demonstrated none of the six predictors comprising the decision rule.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

The main benefit of reducing neuroimaging among children with post-traumatic headache relates to the avoidance of harms. Schachar et al. (2011) tested the sensitivity and specificity of three clinical decision rules (New Orleans Criteria, Canadian CT Head Rule, and NEXUS II) in a population of 2,101 children with head injuries. The authors found sensitivities ranging from 65.2% (95% CI 69.9-86.7) for the Canadian CT Head Rule to 96.7% (95%CI: 93.1-100) for the New Orleans Criteria and negative predictive values above 97%.

Specificity ranged from 11.2% (95% CI: 9.8-12.6) for the New Orleans Criteria to 64.2% for the Canadian CT Head Rule.

The evidence related to the need for neuroimaging in the evaluation of children within 24 hours of mild traumatic brain injury was greatly strengthened by research conducted by PECARN investigators. Their research found that CT scans were obtained for 14,969 (35%) of 42,412 children evaluated in participating EDs within 24 hours of head injury; however, clinically important traumatic brain injuries were present in just 376 (<1%) (Kuppermann et al., 2009). This study generated a clinical decision rule that can guide the decision to order CT imaging for children with mild head trauma and no findings that suggest clinically important traumatic brain injury.

The PECARN head imaging clinical decision rule for children with mild traumatic brain injury has 99.9% negative predictive value and 96.8% sensitivity for predicting clinically important injury. The PECARN study provides evidence that imaging was overused in approximately 20% of the study population 2 years and older who demonstrated none of the six predictors comprising the decision rule.

Kuppermann N, Holmes JF, Dayan PS, et al., Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study. *Lancet* 2009; 374: 1160–1170.

Schachar JL, Zampolin RL, Miller TS, Farinhas JM, Freeman K, Taragin BH. External validation of the New Orleans Criteria (NOC), the Canadian CT Head Rule (CCHR) and the National Emergency X-Radiography Utilization Study II (NEXUS II) for CT scanning in pediatric patients with minor head injury in a non-trauma center. *Pediatr Radiol* 2011; 41(8):971-979.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

CT use has increased in the past 20 years without an increase in the yield of imaging studies. CT rates for children with mild head trauma vary widely between hospitals. CT rates ranged from 19% to 58% for patients with minor head injury in a retrospective analysis of 5 years of hospital administrative data from 40 free-standing children's hospitals (Mannix et al., 2012). This research also suggests that rates of imaging following head injury may be declining in free-standing children's hospitals in recent years.

The harms of neuroimaging among children with post-traumatic headache have not been directly studied but can be implied from the literature that describes the potential harm associated with radiation exposure (Pearce et al., 2012). The absolute incidence of induced lethal malignancy is estimated at 1/1000-1/5000 per cranial CT (Brenner et al., 2007).

Citations:

Brenner DJ, Hall EJ. Computed tomography — an increasing source of radiation exposure. *N Engl J Med* 2007; 357(22):2277-2284.

Mannix R, Meehan WP, Monuteaux MC, Bachur RG. Computed tomography for minor head injury: Variation and trends in major United States emergency departments. *J Pediatr* 2012; 160:136-139.

Pearce MS, Salotti JA, Little MP. Radiation exposure from CT scans in childhood and subsequent risk of leukemia and brain tumours: A retrospective cohort study. *Lancet* 2012; 380(9840): 499–505.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

The systematic review of the body of evidence primarily focuses around the yield of imaging and does not directly address the risks /harms associated with imaging. Studies related to trends in imaging use and the risks/harms associated with imaging are briefly described below.

CT use has increased in the past 20 years without an increase in the yield of imaging studies. In a cross-sectional analysis of data from the National Hospital Ambulatory Medical Care Survey, Blackwell et al. (2007) found the use of CT scans for the evaluation of children with head injury nearly doubled from 1995 to 2003 (13% to 22%); Zonfrillo et al. (2015) found evidence to suggest continued increases in CT use for ED patients with concussion from 2006 to 2011. Some research suggests that rates of imaging following head injury have declined in free-standing children's hospitals (Menoch et al., 2012; Mannix et al., 2012; Parker et al., 2015) and general EDs (Marin et al., 2014). CT rates for children with mild head injury ranged from 19% to 69% across the 25 EDs that collected data for the PECARN study (Stanley et al., 2014). Similarly, CT rates ranged from 19% to 58% for patients with minor head injury in a retrospective analysis of 5 years of hospital administrative data from 40 free-standing children's hospitals (Mannix et al., 2012).

The potential risks and harms associated with imaging include radiation exposure (Pearce et al., 2012; Mathews et al., 2013); complications from sedation and/or anesthesia (Malviya et al., 2000; Wachtel et al., 2009); incidental findings leading to potentially invasive and costly follow-up testing (Lumbreras et al., 2010; Rogers et al., 2013); and excess costs to the healthcare system, which are passed on to families (Callaghan et al., 2014).

Citations Not Included in Systematic Review (Evidence Table):

Blackwell CD, Gorelick M, Holmes JF, Bandyopadhyay S, Kuppermann N. Pediatric head trauma: Changes in use of computed tomography in emergency departments in the United States over time. *Ann Emerg Med* 2007; 49(3):320-324.

Callaghan BC, Kerber KA, Pace RJ, Skolarus LE, Burke JF. Headaches and neuroimaging: High utilization and costs despite guidelines. *JAMA Intern Med* 2014; 174(5):819-821.

Lumbreras B, Donat L, Hernández-Aquado I. Incidental findings in imaging diagnostic tests: A systematic review. *Br J Radiol* 2010; 83(988):276-289.

Malviya S, Voepel-Lewis T, Eldevik OP, Rockwell DT, Wong JH, Tait AR. Sedation and general anesthesia in children undergoing MRI and CT: Adverse events and outcomes. *Br J Anaesth* 2000; 84(6):743-748.

Mannix R, Meehan WP, Monuteaux MC, Bachur RG. Computed tomography for minor head injury: Variation and trends in major United States emergency departments. *J Pediatr* 2012; 160:136-139.

Marin JR, Weaver MD, Barnato AE, Yabes JG, Yealy DM, Roberts MS. Variation in emergency department head computed tomography use for pediatric head trauma. *Acad Emerg Med* 2014; 21(9):987-995.

Mathews JD, Forsythe AV, Brady Z, et al. Cancer risk in 680,000 people exposed to computed tomography scans in childhood or adolescence: Data linkage study of 11 million Australians. *BMJ* 2013; 346:f2360.

Menoch MJ, Hirsh DA, Khan NS, Simon HK, Sturm JJ. Trends in computed tomography utilization in the pediatric emergency department. *Pediatrics* 2012; 129(3):e690-e697.

Parker MW, Shah SS, Hall M, Fieldston ES, Coley BD, Morse RB. Computed tomography and shifts to alternate imaging modalities in hospitalized children. *Pediatrics* 2015; 136(3):e573-e581.

Pearce MS, Salotti JA, Little MP. Radiation exposure from CT scans in childhood and subsequent risk of leukemia and brain tumours: A retrospective cohort study. *Lancet* 2012; 380(9840): 499–505.

Rogers AJ, Maher CO, Schunk JE, et al. Incidental findings in children with blunt head trauma evaluated with cranial CT scans. *Pediatrics* 2013; 132(2):e356-e363.

Stanley RM, Hoyle JD Jr, Dayan PS, et al. Emergency department practice variation in computed tomography use for children with minor blunt head trauma. *J Pediatr* 2014; 165(6):1201-1206.

Wachtel RE, Dexter F, Dow AJ. Growth rates in pediatric diagnostic imaging and sedation. *Anesth Analg* 2009; 108(5):1616-1621.

Zonfrillo MR, Kim KH, Arbogast KB. Emergency department visits and head computed tomography utilization for concussion patients from 2006 to 2011. *Acad Emerg Med* 2015; 22(7):872-877.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Q-METRIC_IMG_Post-TraumaHD_NQF_EvidenceAttachment.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Post-traumatic headaches in children are a common clinical presentation in the setting of concussion and mild traumatic brain injury. In the United States, it has been estimated that more than 500,000 children younger than 15 years of age were evaluated in an ED following mild traumatic brain injury each year from 1998 to 2000 (Bazarian et al., 2005). Over the past decade, ED visits for traumatic brain injuries have increased substantially (Coronado et al., 2015).

Well-established evidence shows that neuroimaging to evaluate children with post-traumatic headache in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or skull fracture is rarely clinically indicated and is potentially harmful (Kuppermann et al., 2009; Lateef et al., 2009; Lateef et al., 2012; ACR Expert Panel on Pediatric Imaging, Ryan et al., 2014). The American Academy of Pediatrics Choosing Wisely initiative includes guidance to discourage the unnecessary use of CT scans for the immediate evaluation of minor head injuries and encourage reliance on clinical observation/PECARN criteria to determine whether imaging is indicated (AAP Choosing Wisely, 2013; Kuppermann et al., 2009).

CT use has increased in the past 20 years. In a cross-sectional analysis of data from the National Hospital Ambulatory Medical Care Survey, Blackwell et al. (2007) found the use of CT scans for the evaluation of children with head injury nearly doubled from 1995 to 2003 (13% to 22%); Zonfrillo et al. (2015) found evidence to suggest continued increases in CT use for ED patients with concussion from 2006 to 2011. Some research suggests that rates of imaging following head injury appear to have declined in free-standing children's hospitals (Menoch et al., 2012; Mannix et al., 2012; Parker et al., 2015) and general EDs (Marin et al., 2014). Also, CT rates for children with mild head trauma vary widely between hospitals. CT rates ranged from 19% to 69% across 25 EDs (Stanley et al., 2014). Similarly, CT rates ranged from 19% to 58% for patients with minor head injury in a retrospective analysis of 5 years of hospital administrative data from 40 free-standing children's hospitals (Mannix et al., 2012).

Overuse has been defined as any patient who undergoes a procedure or test for an inappropriate indication (Lawson et al., 2012). Imaging overuse for the evaluation of children with post-traumatic headaches without signs or symptoms of intracranial injury subjects children to a number of risks (Malviya et al., 2000; Mathews et al., 2013; Pearce et al., 2012; Wachtel et al., 2009). Individuals who undergo CT scans in early childhood tend to be at greater risk for developing leukemia, primary brain tumors, and other malignancies later in life (Mathews et al., 2013; Pearce et al., 2012). Children are also at risk for complications from sedation or anesthesia, which are often required for longer CT imaging sequences and for MRI, and from intravenous contrast media (Zo'o et al., 2011). Cost is also an issue (Callaghan et al., 2014) that burdens the patient, as well as payers.

Citations:
American Academy of Pediatrics (AAP). Choosing Wisely: An initiative of the ABIM Foundation. Ten Things Physicians and Patients Should Question. 2013. Available at: http://www.choosingwisely.org/doctor-patient-lists/american-academy-of-pediatrics/; accessed: February 24, 2015.

American College of Radiology Expert Panel on Pediatric Imaging: Ryan ME, Palasis S, Saigal G, et al. ACR Appropriateness Criteria: Head trauma — child. American College of Radiology, 2014. Available at: https://acsearch.acr.org/docs/3083021/Narrative/; accessed July 1, 2015.

Bazarian JJ, McClung J, Shah MN, Cheung YT, Flesher W, Kraus J. Mild traumatic brain injury in the United States, 1998-2000. Brain Inj 2005; 19(2):85-91.

Blackwell CD, Gorelick M, Holmes JF, Bandyopadhyay S, Kuppermann N. Pediatric head trauma: Changes in use of computed tomography in emergency departments in the United States over time. Ann Emerg Med 2007; 49(3):320-324.

Callaghan BC, Kerber KA, Pace RJ, Skolarus LE, Burke JF. Headaches and neuroimaging: High utilization and costs despite guidelines. JAMA Intern Med 2014; 174(5):819-821.

Coronado VG, Haileyesus T, Cheng TA, et al. Trends in sports- and recreation-related traumatic brain injuries treated in US emergency departments: The National Electronic Injury Surveillance System-All Injury Program (NEISS_AIP) 2001-2012. J Head Trauma Rehabil 2015; 30(3): 185-197.

Kuppermann N, Holmes JF, Dayan PS, et al., Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study. Lancet 2009; 374: 1160–1170.

Lateef TM, Grewal M, McClintock W, Chamberlain J, Kaulas H, Nelson KB. Headache in young children in the emergency department: Use of computed tomography. Pediatrics 2009; 124:1 e12-e17.

Lateef TM, Kriss R, Carpenter K, Nelson KB. Neurologic complaints in young children in the ED: When is cranial computed tomography helpful? Am J Emerg Med 2012; 30(8):1507-1514.

Lawson EH, Gibbons MM, Ko CY, Shekelle PG. The appropriateness method has acceptable reliability and validity for assessing overuse and underuse of surgical procedures. J Clin Epidemiol 2012; 65(11):1133-1143.

Malviya S, Voepel-Lewis T, Eldevik OP, Rockwell DT, Wong JH, Tait AR. Sedation and general anesthesia in children undergoing MRI and CT: Adverse events and outcomes. Br J Anaesth 2000; 84(6):743-748.

Mannix R, Meehan WP, Monuteaux MC, Bachur RG. Computed tomography for minor head injury: Variation and trends in major United States emergency departments. J Pediatr 2012; 160:136-139.

Marin JR, Weaver MD, Barnato AE, Yabes JG, Yealy DM, Roberts MS. Variation in emergency department head computed tomography use for pediatric head trauma. Acad Emerg Med 2014; 21(9):987-995.

Mathews JD, Forsythe AV, Brady Z, et al. Cancer risk in 680,000 people exposed to computed tomography scans in childhood or adolescence: Data linkage study of 11 million Australians. BMJ 2013; 346:f2360.

Menoch MJ, Hirsh DA, Khan NS, Simon HK, Sturm JJ. Trends in computed tomography utilization in the pediatric emergency department. Pediatrics 2012; 129(3):e690-e697.

Parker MW, Shah SS, Hall M, Fieldston ES, Coley BD, Morse RB. Computed tomography and shifts to alternate imaging modalities in hospitalized children. Pediatrics 2015; 136(3):e573-e581.

Pearce MS, Salotti JA, Little MP. Radiation exposure from CT scans in childhood and subsequent risk of leukemia and brain tumours: A retrospective cohort study. Lancet 2012; 380(9840): 499–505.

Stanley RM, Hoyle JD Jr, Dayan PS, et al. Emergency department practice variation in computed tomography use for children with minor blunt head trauma. J Pediatr 2014; 165(6):1201-1206.

Wachtel RE, Dexter F, Dow AJ. Growth rates in pediatric diagnostic imaging and sedation. Anesth Analg 2009; 108(5):1616-1621.

Zonfrillo MR, Kim KH, Arbogast KB. Emergency department visits and head computed tomography utilization for concussion patients from 2006 to 2011. Acad Emerg Med 2015; 22(7):872-877.

Zo'o M, Hoermann M, Balassy C, et al. Renal safety in pediatric imaging: Randomized, double blind phase IV clinical trial of iobitridol 300 versus iodixanol 270 in multidetector CT. Pediatr Radiol 2011; 41(11); 1393-1400.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

We determined the neuroimaging overuse percentage among children evaluated with head CT or MRI in an emergency department for post-traumatic headache sampled from the HealthCore Integrated Research Database (HIRD). Of the 204 reviewed charts, 57 (27.9%) met denominator criteria: children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained. Among these, 8 children (14.0%) were imaged in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or skull fracture. Overall, our results indicate there is an opportunity to reduce the overuse of neuroimaging among children with post-traumatic headache. However, we were unable to assess plan, hospital emergency department, and provider level variations based on the limited number of eligible medical records that were available for calculation of this measure following chart review.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Patient-level demographic and socioeconomic characteristics were generally unavailable from the medial records reviewed for measure testing. Therefore, we used ZIP-code level race and ethnicity, median household income, and urbanicity, collected for the 2010 United States Census and the 2011 American Community Survey (ACS), as proxy variables to characterize the population. The small numbers of eligible numerator and denominator cases (n=8 and n=57, respectively) do not allow for meaningful comparisons of overuse of neuroimaging among children with post-traumatic headache evaluated in EDs across different socio-demographic groups.

Race and Ethnicity - Census Characteristics

On average, children with post-traumatic headache who obtained neuroimaging resided in ZIP codes reporting primarily white race (80.2%) and modest levels of Hispanic ethnicity (9.8%). The children included in the denominator group resided in ZIP codes reporting a higher proportion of white residents (81.8%) and a similar proportion of Hispanic ethnicity (10.0%). The children included in the numerator group resided in ZIP codes reporting a still higher proportion of white residents (84.9%) and a slightly lower proportion of residents of Hispanic ethnicity (6.6%). These demographic characteristics differ from the population of the United States as a whole, as the 2010 US Census data indicates that approximately 72.4% of the population was white, 13.2% of the population was black, and 16.3% of the population was of Hispanic ethnicity in 2010. The summary statistics for race and ethnicity within ZIP code across the sampled subgroups of children with valid ZIP codes are reported in the Appendix – Tables 1 and 2.

Socioeconomic Status – Census Characteristics

On average, the ZIP code-level median household income for children with post-traumatic headache who obtained neuroimaging was $69,540. The children in the denominator group resided in ZIP codes with higher median household incomes (mean $81,430) and those included in the numerator group resided in ZIP codes with lower median household incomes (mean $64,401). The median household income for the ZIP-codes in which these children resided was substantially higher than the median household income of the population of the entire United States as reported in the American Community Survey in 2011, which is $50,502. The summary statistics for distribution of the ZIP-code level median household income for sampled groups of children with valid ZIP codes and complete census data are reported in the Appendix – Table 3.

Urbanicity – Census Characteristics

Children with post-traumatic headache who obtained neuroimaging primarily reside in urban ZIP codes (75.4%). The subset of children meeting denominator criteria resided in ZIP codes that were slightly more urban (77.9%), and those children meeting numerator criteria resided in substantially less urban ZIP codes (52.8%). The proportion of children in this sample who resided in urban ZIP codes is similar to the rest of the United States, where approximately 79% of the population resides in an urban area. The summary statistics for urbanicity within ZIP code for sampled groups of children with valid ZIP codes are reported in the Appendix 1 – Table 4.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from**

the literature that addresses disparities in care on the specific focus of measurement. Include citations.
In a cross-sectional study of 50,835 pediatric emergency visits for head injury captured in the National Hospital Ambulatory Medical Care Survey 2002-2006, white race was associated with higher odds of neuroimaging (OR 1.5, 95% CI: 1.02-2.1) (Mannix et al., 2010). Natale and colleagues (2012) conducted a secondary analysis of data prospectively collected for the Pediatric Emergency Care Applied Research Network (PECARN) head imaging decision rule (Kuppermann et al., 2009) to test for associations between race/ethnicity and the ordering of CT among children with blunt head injury. They found that children of black non-Hispanic or Hispanic race/ethnicity had lower odds of undergoing head CT than white non-Hispanic children. Parental anxiety and parental request were cited as reasons for ordering head CT in children of white, non-Hispanic race/ethnicity. Their findings suggest that overuse of CT imaging may disproportionately affect white, non-Hispanic children. Similarly, Morrison and colleagues (2015) found that minority race was associated with less radiologic testing in the children of parents with low health literacy in a cross-sectional study of 504 caregivers accompanying their child to a pediatric ED. When associated with race/ethnicity, overuse of health care, in general, is greater among white patients (Kressin and Groeneveld, 2015).

Citations:
Kressin NR, Groeneveld PW. Race/ethnicity and overuse of care: A systematic review. Milbank Q 2015; 93(1):112-138.

Kuppermann N, Holmes JF, Dayan PS, et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study. Lancet 2009; 374: 1160–1170.

Mannix R, Bourgeois FT, Schutzman SA, Bernstein A, Lee LK. Neuroimaging for pediatric head trauma: Do patient and hospital characteristics influence who gets imaged? Acad Emerg Med 2010; 17(7):694-700.

Morrison AK, Brousseau DC, Brazauskas R, Levas MN. Health literacy affects likelihood of radiology testing in the pediatric emergency department. J Pediatr 2015; 166(4):1037-1041.

Natale JE, Joseph JG, Rogers AJ, et al. Cranial computed tomography use among children with minor blunt head trauma. Association with race/ethnicity. Arch Pediatr Adolesc Med 2012; 166(8):732-737.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Per NQF pre-review: Not currently an evaluation criterion.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
See Citations in 1b.1.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Neurology : Brain Injury

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse, Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.chear.org/sites/default/files/stories/pdfs/img3_speconly.pdf

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Q-METRIC_IMG_Post-TraumaHD_NQF_Code_Tables.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The number of numerator eligible children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained in the absence of documented neurologic signs or symptoms that suggest intracranial hemorrhage or basilar skull fracture.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The time period for data include the measurement year (January 1 through December 31) (for imaging of the head for the evaluation of a post-traumatic headache) and the year (365 days) prior to the imaging event (for the purpose of identifying a claims-based denominator exclusion).

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Numerator exclusions are based on chart review; they are briefly summarized here and identified in the measure specification.

- Severe mechanism of injury (e.g., penetrating trauma, fall from more than 5 feet, struck by vehicle)
- History of seizure or convulsions associated with trauma
- History of loss of consciousness associated with trauma
- Repeated vomiting
- Documented basilar skull fracture or signs of suspected basilar skull fracture, including “Raccoon eyes”, Battle’s sign, and hemotympanum

- Absence of documented neurologic examination
- Abnormal neurologic examination or signs or symptoms of intracranial hemorrhage or increased intracranial pressure (e.g., decreased alertness, altered mental status, Glasgow Coma Scale Score <14, diplopia, abnormal face or eye movements, gait disturbance)

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The number of children, ages 2 through 17 years old, with post-traumatic headache who were evaluated in the ED within 24 hours after an injury, and imaging of the head (CT or MRI) was obtained in the absence of suspected child abuse or neglect or a history of a medical condition that would otherwise warrant neuroimaging.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Eligible children must be ages 2 through 17 years old during the measurement year for which imaging of the head is obtained and must be continuously enrolled in their insurance plan during both the measurement year and the year prior. Eligible children must also receive head imaging in association with an ED visit for post-traumatic headache within 24 hours of the time of injury.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Children under evaluation for child abuse or neglect and children with a history of a medical condition that could otherwise warrant neuroimaging (e.g., bleeding disorder, intracranial tumor, hydrocephalus) for the evaluation of a post-traumatic headache were excluded from this overuse measure.

Children with a diagnosis of headache without a documented history of trauma and children with a diagnosis of concussion without documentation of headache as a symptom were excluded because post-traumatic headache is the focus of this measure.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The complete list of ICD-9-CM code-based exclusions (with conversion to ICD-10-CM codes) that can be applied to administrative claims data are provided in the Data Code Tables identified in S.2b. Denominator exclusions are also applied to chart review and identified in the measure specification.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify children in the denominator:
a. Using administrative claims, identify the population eligible for the denominator. The eligible population consists of all individuals who satisfy specified criteria, including age, enrollment, diagnosis, and imaging requirements within the measurement year.
b. Using administrative claims, exclude individuals with ICD-9-CM codes/ICD-10-CM codes associated with child abuse/neglect or a history of a medical condition that could otherwise warrant neuroimaging for the evaluation of a post-traumatic headache.
c. Select a random sample of those still eligible for the denominator for chart abstraction.
d. Among those who have a chart abstracted, exclude individuals with no documented time of injury or a time of injury greater than 24 hours prior to the ED visit, a diagnosis of headache without documentation of trauma, a diagnosis of concussion without documentation of headache as a symptom, concern for child abuse/neglect, or a history of a medical condition that could otherwise warrant neuroimaging to obtain the population included within the final denominator.

2. Identify children in the numerator:
a. Among children included within the final denominator, exclude from the numerator individuals who have documented within the medical chart the following: severe mechanism of injury, seizure associated with trauma, loss of consciousness associate with trauma, repeated vomiting, documented or suspected basilar skull fracture, no documentation of a neurologic examination, or abnormal neurologic examination including altered mental status, Glasgow Coma Scale score <14, abnormal face or extremity movements, or gait disturbance.

3. Calculate the percentage overuse (numerator / denominator multiplied by 100%).

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Sampling: Administrative claims are used to identify the eligible population for the denominator and to identify claims-based denominator exclusions. From those still eligible for the denominator, a random sample is selected for chart abstraction. The final denominator population is determined using medical record data to identify remaining denominator exclusions. Medical record data are then used to identify numerator exclusions among children meeting eligibility for inclusion in the denominator.

Availability of medical records meeting inclusion criteria will vary by the entity using this measure. This measure was tested using a target sample of 200 abstracted charts for eligible children during the measurement year. Of the 204 charts abstracted for testing, 75 children had a headache diagnosis code and 67 of those charts were excluded because there was no clinical documentation of trauma occurring within 24 hours of the ED visit. Overall, we found 57 charts (27.9% of the sample obtained for chart review) met denominator criteria and were eligible for evaluation of measure numerator exclusions. A sample of 55 charts included in the denominator would yield a 95% confidence interval (CI) with a half-width of 8% for an expected overuse percentage of 10%. A sample size of 554 would be needed to achieve a 95% CI with a half-width of 2.5% for an expected overuse percentage of 10%. Larger numbers of abstracted charts will be required to ensure sufficient sample size; this will allow greater confidence in overuse percentage estimates and enable testing for differences between providers, hospital EDs, or health plans.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Use of this measure requires administrative claims associated with emergency department visits during which neuroimaging was obtained for the evaluation of a child with post-traumatic headache. The clinical documentation from that emergency department visit, in paper or electronic medical record format, is required to determine if a case is eligible for inclusion in the measure denominator and numerator. Data could be obtained and analyzed at the hospital or health plan level.

Testing this measure using medical record data required the development of an abstraction tool and the use of qualified nurse abstractors. We provide an example data abstraction tool for chart review (see URL identified in S.1. above).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Q-METRIC_IMG_Post-TraumaHD_NQF_TestingAttachment.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Overuse of Imaging for the Evaluation of Children with Post-Traumatic Headache
Date of Submission: 9/30/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☒ abstracted from paper record – N and D	☒ abstracted from paper record – N and D
☒ administrative claims – D only	☒ administrative claims – D only
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record – N and D	☒ abstracted from electronic health record -N and D
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
Data used for testing were obtained from HealthCore, Inc., an independent subsidiary of Anthem, Inc., which is the largest health benefits company/insurer in the United States. HealthCore owns and operates the HealthCore Integrated Research Database (HIRD), a longitudinal database of medical and pharmacy claims and enrollment information.

1.3. What are the dates of the data used in testing? January 1, 2011 through December 31, 2012

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the*

analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)
This measure was tested using data contained in the HIRD. The HIRD includes automated computerized claims data and enrollment information for members from 14 geographically diverse Blue Cross and/or Blue Shield (BC/BS) Health Plans in the Northeast, South, West, and Central regions of the United States, with members living in all 50 states. The HIRD represents data from approximately 60 million lives with medical enrollment, over 37 million lives with combined medical and pharmacy enrollment information, and 16 million with outpatient laboratory data.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
This measure belongs to the Q-METRIC Overuse of Imaging for the Evaluation of Children with Headache or Seizures measures collection. As part of the initial sampling strategy for testing multiple measures in this collection, approximately 2.1 million children, ages 6 months through 17 years old, were identified in the HIRD for the study's 2012 measurement year. Of these, a cohort of children with diagnosis codes for headaches and seizures were identified (57,748). Members who did not have continuous eligibility during the 2011 and 2012 calendar years were excluded, narrowing the group to 36,985 (64.0%).

Specifically for this measure, administrative claims were used to identify children, ages 2 through 17 years old, who had ICD-9-CM codes that indicated a post-traumatic headache, concussion, or general symptoms of headache evaluated in the emergency department (ED; 5,912, 16.0%). From this group, 2,967 children (50.2%) were identified as having either CT or MR imaging. After applying claims-based exclusions (suspected abuse/neglect, history of a medical condition that could warrant neuroimaging, loss of consciousness, skull fracture, and intracranial hemorrhage), 2,419 children (81.5%) were eligible to sample for chart review.

Once the population eligible for chart review was determined using administrative claims, providers associated with visits were identified. The final sampling population for chart review consisted of 1,714 children (70.9%) who could be linked to a provider having complete contact information. In an attempt to obtain an adequate number of cases to test this measure, we set a target sample of 200 abstracted charts. Patient medical records were then requested from provider offices and healthcare facilities for data abstraction. Patient medical records were sent to a centralized location for data abstraction. The first 204 charts received were abstracted for measure testing; 86 children (42.2%) were female, and the average age was 12.0 (SD = 3.9).

Of the 204 abstracted charts, one (0.5%) was excluded based on clinical documentation of suspected child abuse or neglect and five (2.5%) were excluded due to documentation of a medical condition that could otherwise warrant neuroimaging. There were 65 charts (31.9%) with clinical documentation of trauma occurring within 24 hours of the ED visit; among those, eight were excluded, as they had concussion as a diagnosis without evidence of a headache as a symptom, leaving 57 charts (27.9%) in the eligible study population.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
The administrative dataset and chart review sample described above were used for all aspects of testing.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
Patient-level demographic and socioeconomic characteristics were generally unavailable from the medical records reviewed for measure testing. Therefore, we used ZIP-code level race and ethnicity, median household income, and urbanicity, collected for the 2010 United States Census and the 2011 American Community Survey (ACS), as proxy variables to characterize the population. The small numbers of eligible numerator and denominator cases (n=8 and n=57, respectively) do not allow for meaningful comparisons of overuse of

neuroimaging among children evaluated in EDs with post-traumatic headache across different socio-demographic groups.

Race and Ethnicity Census Characteristics

On average, children with post-traumatic headache who underwent neuroimaging resided in ZIP codes reporting primarily white race (80.2%) and modest levels of Hispanic ethnicity (9.8%). The children included in the denominator group resided in ZIP codes reporting a higher proportion of white residents (81.8%) and a similar proportion of Hispanic ethnicity (10.0%). The children included in the numerator group resided in ZIP codes reporting a still higher proportion of white residents (84.9%) and a slightly lower proportion of residents of Hispanic ethnicity (6.6%). These demographic characteristics differ from the population of the United States as a whole, as the 2010 US Census data indicates that approximately 72.4% of the population was white, 13.2% was black, and 16.3% was of Hispanic ethnicity. The summary statistics for race and ethnicity within ZIP code across the sampled subgroups of children with valid ZIP codes are reported in Tables 1 and 2 (see pages 6 and 7).

Socioeconomic Status – Census Characteristics

On average, the ZIP code-level median household income for children with post-traumatic headache who underwent neuroimaging was $69,540. The children in the denominator group resided in ZIP codes with higher median household incomes (mean $81,430), and those included in the numerator group resided in ZIP codes with lower median household incomes (mean $64,401). The median household income for the ZIP codes in which these children resided was substantially higher than the median household income of the population of the entire United States, as reported in the American Community Survey in 2011, which was $50,502. The summary statistics for distribution of the ZIP-code level median household income for sampled groups of children with valid ZIP codes and complete census data are reported in Table 3 (see page 8).

Urbanicity – Census Characteristics

Children with post-traumatic headache who underwent neuroimaging primarily reside in urban ZIP codes (75.4%). The subset of children meeting denominator criteria resided in ZIP codes that were slightly more urban (77.9%), and those children meeting numerator criteria resided in substantially less urban ZIP codes (52.8%). The proportion of children in this sample who resided in urban ZIP codes is similar to the rest of the United States, where approximately 79% of the population resides in an urban area. The summary statistics for urbanicity within ZIP code for sampled groups of children with valid ZIP codes are reported in Table 4 (see page 9).

Table 1. Mean (SD) Proportion of Racial Groups within Sampled ZIP Codes of Residence‡

Sampled Group Description	American Indian or Alaska Native Mean (SD) ‡	Asian Mean (SD) ‡	Black or African American Mean (SD) ‡	Native Hawaiian or Other Pacific Islander Mean (SD) ‡	White Mean (SD) ‡	Two or More Races Mean (SD) ‡	Other Mean (SD) ‡
Eligible children with post-traumatic headache (n=5,807)*	0.5 (1.1)	5.1 (8.0)	8.5 (13.5)	0.1 (0.2)	79.6 (17.7)	2.6 (1.4)	3.7 (6.0)
Subset who had a CT or MRI (n=2,918)**	0.5 (1.0)	4.9 (7.9)	8.0 (12.6)	0.1 (0.2)	80.2 (16.8)	2.6 (1.4)	3.7 (6.1)

Subset following claims denominator exclusions (n=2,386)***	0.5 (1.0)	4.8 (7.8)	8.2 (12.7)	0.1 (0.2)	80.3 (16.8)	2.6 (1.4)	3.6 (6.0)
Subset following claims numerator exclusions (n=1,985)****	0.5 (0.9)	4.9 (8.1)	8.1 (12.6)	0.1 (0.2)	80.3 (16.8)	2.6 (1.5)	3.6 (5.9)
Subset with reviewed and abstracted medical records (n=200)+	0.4 (0.3)	5.4 (8.1)	7.0 (10.1)	0.1 (0.3)	80.5 (15.6)	2.6 (1.4)	3.9 (6.9)
Children meeting denominator criteria (n=57)++	0.4 (0.4)	6.1 (9.1)	5.1 (8.2)	0.1 (0.1)	81.8 (15.1)	2.6 (1.3)	3.9 (7.6)
Children meeting numerator criteria (n=8)+++	0.3 (0.1)	3.3 (4.3)	7.1 (12.4)	0.03 (0.05)	84.9 (15.5)	1.8 (1.2)	2.7 (2.3)

SD = standard deviation
‡Data summarize characteristics of the broader population residing in ZIP codes of sampled cases.
*Among eligible children who had a post-traumatic headache (n=5,912), no information available for 105 members (1.8%) due to missing or unmatched ZIP code, yielding n=5,807 (98.2%).
** Among the subset of children who had a CT or MRI (n=2,967), no information available for 49 members (1.7%) due to missing or unmatched ZIP code, yielding n=2,918 (98.3%).
*** Among the subset of children following denominator exclusions (n=2,419), no information available for 33 members (1.4%) due to missing or unmatched ZIP code, yielding n=2,386 (98.6%).
**** Among the subset of children following numerator exclusions (n=2,009), no information available for 24 (1.2%) members due to missing or unmatched ZIP code, yielding n=1,985 (98.8%).
+ Among the subset of children with abstracted medical records (n=204), no information available for 4 members (2.0%) due to missing or unmatched ZIP code, yielding n=200 (98.0%).
++ Among children meeting denominator criteria (n=57), information was available for all members, yielding n=57 (100%).
+++ Among children meeting numerator criteria (n=8), information was available for all members, yielding n=8 (100%).

Table 2. Mean (SD) Proportion Reporting Hispanic Ethnicity within Sampled ZIP Codes of Residence‡

Sampled Group Description	Hispanic Ethnicity Mean (SD) ‡
Eligible children with post-traumatic headache (n=5,807)*	9.7 (13.5)
Subset who had a CT or MRI (n=2,918)**	9.8 (13.6)
Subset following claims denominator exclusions (n=2,386)***	9.5 (13.3)
Subset following claims numerator exclusions (n=1,985)****	9.4 (13.2)
Subset with reviewed and abstracted medical records (n=200)+	10.3 (14.7)
Children meeting denominator criteria (n=57)++	10.0 (13.9)
Children meeting numerator criteria (n=8)+++	6.6 (5.3)

SD = standard deviation

‡Data summarize characteristics of the broader population residing in ZIP codes of sampled cases.
*Among eligible children who had a post-traumatic headache (n=5,912), no information available for 105 members (1.8%) due to missing or unmatched ZIP code, yielding n=5,807 (98.2%).
** Among the subset of children who had a CT or MRI (n=2,967), no information available for 49 members (1.7%) due to missing or unmatched ZIP code, yielding n=2,918 (98.3%).
*** Among the subset of children following denominator exclusions (n=2,419), no information available for 33 members (1.4%) due to missing or unmatched ZIP code, yielding n=2,386 (98.6%).
**** Among the subset of children following numerator exclusions (n=2,009), no information available for 24 (1.2%) members due to missing or unmatched ZIP code, yielding n=1,985 (98.8%).
+ Among the subset of children with abstracted medical records (n=204), no information available for 4 members (2.0%) due to missing or unmatched ZIP code, yielding n=200 (98.0%).
++ Among children meeting denominator criteria (n=57), information was available for all members, yielding n=57 (100%).
+++ Among children meeting numerator criteria (n=8), information was available for all members, yielding n=8 (100%).

Table 3. Median Household Income within Sampled ZIP Codes of Residence‡

Sampled Group Description	Median Household Income (Mean)‡	SD	Min	25th Percentile	Median	75th Percentile	Max
Eligible children with post-traumatic headache (n=5,805)*	$ 69,886	$29,495	$15,473	$47,570	$63,878	$85,462	$219,688
Subset who had a CT or MRI (n=2,917)**	$69,540	$29,624	$16,036	$47,028	$63,542	$85,462	$219,688
Subset following claims denominator exclusions (n=2,386)***	$69,188	$29,630	$16,036	$46,964	$63,158	$85,380	$219,688
Subset following claims numerator exclusions (n=1,985)****	$68,934	$29,430	$16,036	$46,733	$63,269	$85,011	$219,688
Subset with reviewed and abstracted medical records (n=200)+	$74,498	$30,834	$20,673	$51,920	$69,214	$93,236	$167,037
Children meeting denominator criteria (n=57)++	$81,430	$32,839	$30,085	$57,712	$76,014	$99,041	$167,037
Children meeting numerator criteria (n=8)+++	$64,401	$33,345	$32,267	$39,979	$51,366	$84,965	$130,319

‡Data summarize characteristics of the broader population residing in ZIP codes of sampled cases.
*Among eligible children who had a post-traumatic headache (n=5,912), no information available for 107 members (1.8%) due to missing or unmatched ZIP code or missing census data, yielding n=5,805 (98.2%).
** Among the subset of children who had a CT or MRI (n=2,967), no information available for 50 members (1.7%) due to missing or unmatched ZIP code or missing census data, yielding n=2,917 (98.3%).
*** Among the subset of children following denominator exclusions (n=2,419), no information available for 33 members (1.4%) due to missing or unmatched ZIP code or missing census data, yielding n=2,386 (98.6%).
**** Among the subset of children following numerator exclusions (n=2,009), no information available for 24 members (1.2%) due to missing or unmatched ZIP code or missing census data, yielding n=1,985 (98.8%).
+ Among the subset of children with abstracted medical records (n=204), no information available for 4 members (2.0%) due to missing or unmatched ZIP code or missing census data, yielding n=200 (98.0%).
++ Among children meeting denominator criteria (n=57), information was available for all members, yielding n=57 (100%).
+++ Among children meeting numerator criteria (n=8), information was available for all members, yielding n=8 (100%).

Table 4. Proportion of Sampled ZIP Codes Categorized as Urban‡

Sampled Group Description	Urban (Mean)‡	SD	Min	25th Percentile	Median	75th Percentile	Max
Eligible children with post-traumatic headache (n=5,807)*	77.3	32.8	0	66.4	95.0	100	100
Subset who had a CT or MRI (n=2,918)**	75.4	33.7	0	63.0	93.7	100	100
Subset following claims denominator exclusions (n=2,386)***	74.5	34.2	0	61.6	93.0	100	100
Subset following claims numerator exclusions (n=1,985)****	74.0	34.5	0	61.1	92.9	100	100
Subset with reviewed and abstracted medical records (n=200)+	77.5	32.3	0	63.6	95.2	100	100
Children meeting denominator criteria (n=57)++	77.9	30.9	0	68.3	94.2	100	100
Children meeting numerator criteria (n=8)+++	52.8	42.3	0	6.3	67.7	87.1	100

‡Data summarize characteristics of the broader population residing in ZIP codes of sampled cases.
*Among eligible children who had a post-traumatic headache (n=5,912), no information available for 105 members (1.8%) due to missing or unmatched ZIP code, yielding n=5,807 (98.2%).
** Among the subset of children who had a CT or MRI (n=2,967), no information available for 49 members (1.7%) due to missing or unmatched ZIP code, yielding n=2,918 (98.3%).
*** Among the subset of children following denominator exclusions (n=2,419), no information available for 33 members (1.4%) due to missing or unmatched ZIP code, yielding n=2,386 (98.6%).
**** Among the subset of children following numerator exclusions (n=2,009), no information available for 24 (1.2%) members due to missing or unmatched ZIP code, yielding n=1,985 (98.8%).
+ Among the subset of children with abstracted medical records (n=204), no information available for 4 members (2.0%) due to missing or unmatched ZIP code, yielding n=200 (98.0%).
++ Among children meeting denominator criteria (n=57), information was available for all members, yielding n=57 (100%).
+++ Among children meeting numerator criteria (n=8), information was available for all members, yielding n=8 (100%).

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
See section **2b2.** for validity testing of data elements.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Validity of Exclusion Criteria

Denominator: We tested the validity of administrative claims to exclude cases from the denominator based on two ICD-9-CM code-based criteria: 1) suspected child abuse or neglect and 2) history of a medical condition that could otherwise warrant neuroimaging, against the gold standard, the medical record. Children with ICD-9-CM codes associated with these claims-based exclusions were removed from the chart review sample. In other words, none of the charts sampled for medical record review contained ICD-9-CM codes associated with these claims-based exclusions. We tested the accuracy of the assumption that the absence of these ICD-9-CM codes in administrative claims would mean the absence of clinical documentation indicative of these exclusionary conditions in the medical record.

Numerator: We tested administrative claims against chart review data to determine the potential to exclude cases from the numerator using administrative claims for two numerator criteria: 1) seizure or convulsion and 2) indicators of increased intracranial pressure. Data for these two numerator criteria were abstracted from charts and ICD-9-CM codes were identified in administrative claims. The medical chart was considered the gold standard. Sensitivity, specificity, and negative and positive predictive values were calculated.

Conversion of ICD-9-CM to ICD-10-CM Codes

The goal of ICD-9-CM to ICD-10-CM conversion was to translate this measure to a new code set, fully consistent with the intent of the original measure. Codes are attached in S.2b of the Measure Submission Form. All ICD-9-CM diagnosis codes were converted to ICD-10-CM codes using the Centers for Medicare and Medicaid Services (CMS) 2015 diagnosis code General Equivalence Mappings (GEM) and diagnosis code description files, accessed on August 26, 2015. The ICD-9-CM codes were converted to ICD-10-CM using the GEM file and manually reviewed for consistency using the diagnosis code descriptions for the source ICD-9-CM and converted ICD-10-CM codes. In addition, the resultant ICD-10-CM codes

were back-translated to ICD-9-CM to verify the accuracy of the coding. Source files from CMS were acquired from these files:

1. ICD-9 to 10 diagnosis GEM -*2015_I9gem.txt* *https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html*
2. ICD-10 to 9 diagnosis GEM - 2015_10gem.txt *https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html*
3. ICD-9 description file - *CMS32_DESC_SHORT_DX.txt* *https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html*
4. ICD-10 description file - *icd10cm_order_2015.txt* *https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html*

The resultant ICD-10-CM codes were clinically reviewed. We removed the ICD-10-CM code G44.32x (chronic post-traumatic headache), as this measure is focused on imaging that occurs within 24 hours of an injury. The chronicity of a post-traumatic headache was not characterized in ICD-9-CM codes. We also excluded ICD-10-CM codes specific to psychological abuse and observation following alleged adult physical abuse. The original list of ICD-9-CM codes included one E-code (E934.2 Anticoagulants causing adverse effects in therapeutic use) that did not convert to an ICD-10-CM code.

ICD-9-CM procedure codes for head CT and brain MRI were converted using an online tool: http://www.icd10data.com/Convert

Validity of Data Abstraction from the Medical Record

Validity of medical record data was determined through re-abstraction of patient record data by a senior abstractor, considered the gold standard for medical record review. We calculated the inter-rater reliability (IRR) comparing abstractors with the senior abstractor. IRR was determined by calculating percent agreement and Cohen's kappa statistic. Sensitivity, specificity, and negative and positive predictive values were calculated.

Face Validity of Performance Measure Score

The face validity of this measure was established by a national panel of experts and parent representatives for families of children with headache and seizures convened by Q-METRIC. The Q-METRIC panel included nationally recognized experts in the area of imaging children, representing general pediatrics, pediatric radiology, pediatric neurology, pediatric neurosurgery, pediatric emergency medicine, general emergency medicine, and family medicine. In addition, measure validity was considered by experts in state Medicaid program operations, health plan quality measurement, health informatics, and health care quality measurement. In total, the Q-METRIC imaging panel included 15 experts, providing a comprehensive perspective on imaging children and the measurement of quality metrics for states, health plans, and EDs. The expert panel assessed whether the performance of this measure would result in improved quality of care for children with headache and seizures in relation to neuroimaging. Specifically, the panel weighed the evidence to determine if this measure of overuse could reduce unnecessary imaging among children with post-traumatic headache. The voting process to prioritize the measure was based on the ability of the measure to distinguish good from poor quality.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Validity of Exclusion Criteria

Denominator: Of the 204 charts that were reviewed, one (0.5%) had clinical documentation of suspected child abuse or neglect and five (2.5%) contained clinical documentation of a medical condition that could otherwise warrant neuroimaging in the absence of ICD-9-CM codes associated with these two claims-based denominator exclusions. Therefore, 97% (198 of 204) of the charts reviewed were in agreement with the administrative claims regarding the absence of these denominator exclusions.

Numerator: Among children eligible for the denominator after chart review (n=57), the sensitivity of claims for identification of seizure was 0% (95% CI; 0.0, 97.5) and the specificity was 100% (95% CI; 93.6, 100); positive predictive value could not be calculated because there were no true or false positives and negative predictive value was 98.3%

(95% CI; 90.6, 99.9). The sensitivity of claims for identification of indicators of increased intracranial pressure was 8.1% (95% CI; 1.7, 21.9) and the specificity was 90.0% (95% CI; 68.3, 98.8); positive predictive value was 60.0% (95% CI; 14.7, 94.7) and negative predictive value was 34.6% (95% CI; 22.0, 49.1). Contingency tables for both variables are shown below (Tables 5 and 6).

Table 5: Contingency Table for Presence of Seizure in Administrative Claims and Charts

		Seizure in Claims Based on ICD-9-CM Codes (345.2x, 345.3x, 780.33, 780.39)		
		Present	Absent	Total
Evidence of Seizure or Convulsions Documented in Charts	Present	0	0	0
	Absent	1	56	57
	Total	1	56	57

Table 6: Contingency Table for Presence of Indicators of Increased Intracranial Pressure in Administrative Claims and Charts

		Indicators of Increased Intracranial Pressure in Claims Based on ICD-9-CM Codes (368.2x, 374.3x, 377.0x, 387.5x, 379.50, 386.2x, 780.02, 780.03, 780.09, 780.4x, 780.97, 781.2x - 781.4x, 781.93, 536.2x, 781.94, 780.0x, 379.41, 348.4x, 348.5x)		
		Present	Absent	Total
Evidence of Indicators of Increased Intracranial Pressure in Charts	Present	3	2	5
	Absent	34	18	52
	Total	37	20	57

Conversion of ICD-9-CM to ICD-10-CM Codes

We found the majority of ICD-9-CM codes utilized to narrow the number of eligible charts to sample for chart review for the calculation of this measure mapped to ICD-10-CM codes that remain relevant to our intended specifications. This measure could not be tested in administrative data using ICD-10-CM codes since this testing occurred prior to the clinical adoption of ICD-10-CM coding.

Validity of Data Abstraction from the Medical Record

Of the 204 abstracted medical records, 30 (15%) were reviewed for IRR; percent agreement and kappa were calculated. IRR was assessed by comparing individual abstractor agreement with a senior abstractor as the gold standard on the 16 data elements abstracted from charts for this measure (corresponding to 441 eligible items after accounting for skip patterns). Disagreement was identified for two of the 16 data elements: 1) "Was there documentation of increased intracranial pressure? (indications include: swelling of the optic disc (papilledema), double vision (diplopia), abnormal face or eye movements, dizziness (vertigo), abnormal gait (ataxia), abnormal

coordination (dysmetria), confusion)"; percent agreement was 96.7% (kappa 0.84); and 2) "Was there documentation of altered mental status including comments such as "not acting like himself" per parent report?"; percent agreement was 96.7% (kappa 0.90).

Overall, abstractor agreement was 99.3% (kappa 0.98). The sensitivity of the abstractors to identify chart-based exclusions compared with the senior abstractor was 100% (95% CI; 94.6, 100); specificity was 99.5% (95% CI; 98.1, 99.9); positive predictive value was 97.1% (95% CI; 89.9, 99.7) and negative predictive value was 100% (95% CI; 99.0, 100.0). The related contingency table is below (Table 7).

Table 7: Contingency Table for Presence of Chart Review Exclusions

		Senior Abstractor Identified Chart-Based Exclusion		
		Present	Absent	Total
Abstractor Identified Chart-Based Exclusion	Present	67	2	69
	Absent	0	372	372
	Total	67	374	441

Face Validity of Performance Measure Score

The Q-METRIC expert panel concluded that this measure has a high degree of face validity through a detailed review of concepts and metrics considered to be essential to the appropriate imaging of children. Concepts and draft measures were rated by this group for their relative importance. This measure was highly rated, receiving an average score of 7.0 (with 9 as the highest possible score). In addition, the expert panel concluded that this measure of overuse of neuroimaging for the evaluation of children with post-traumatic headache could reduce unnecessary imaging for this population of children, and the measure would be able to distinguish good from poor quality.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Validity of Exclusion Criteria

Denominator: Our results demonstrate that we were able to exclude nearly all children with clinical evidence of child abuse or neglect or a medical condition that could otherwise warrant neuroimaging through exclusions based on associated ICD-9-CM codes present in administrative claims. Therefore, the use of administrative claims is an appropriate and valid method to narrow the population of charts sampled within this measure specification. However, the presence of these exclusionary conditions in the medical record indicates that medical record abstraction is necessary to accurately identify these two denominator exclusions. The abstraction of this information should be conducted in conjunction with the chart review necessary to identify children with post-traumatic headache, an ED visit within 24 hours of trauma, and the numerator exclusions required for calculation of this measure.

Numerator: The low sensitivity of administrative claims compared with the gold standard of the medical record for the two variables tested indicates that chart review is required for the accurate and complete collection of numerator exclusion criteria.

This measure relies on chart review for the identification of inclusion criteria for which there are no ICD-9-CM codes (e.g., documentation of trauma within 24 hours of the ED visit). Chart review also provides a secondary opportunity to identify exclusion criteria that may not fully be captured in ICD-9-CM codes contained in administrative data. Therefore, we conclude that administrative claims alone are insufficient for calculating neuroimaging overuse percentages at this time.

Conversion of ICD-9-CM to ICD-10-CM Codes

The ICD-9-CM to ICD-10-CM code mapping procedure outlined above can be applied to obtain relevant ICD-10-CM codes for the identification of charts eligible for the chart review sample for the calculation of this measure.

Validity of Data Abstraction from the Medical Record

A kappa of greater than 0.81 is considered almost perfect agreement (Landis and Koch, 1977). A percent agreement of 99.3% and kappa statistic of 0.98 indicate that a very high level of agreement was achieved. Given this evidence, the data elements needed for calculation of the measure can be abstracted with a high degree of accuracy.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159-174.

Face Validity of Measure

Given the high rating assigned by the Q-METRIC expert panel, we feel this measure has a very high degree of face validity.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

There are several exclusion criteria that can be applied to administrative claims to narrow the population eligible for chart review. The degree to which these exclusion criteria affect overuse percentage calculations is unknown. Therefore, we performed a sensitivity analysis of exclusion criteria described below.

Among 2,967 children that visited the ED with a post-traumatic headache, concussion, or general headache and underwent CT or MRI, 450 children (15.2%) had the presence of at least one ICD-9-CM code indicative of child abuse and neglect, loss of consciousness, skull fracture or intracranial hemorrhage. This group of ICD-9-CM codes was flagged as present or absent in the administrative data available to the Q-METRIC team. Claims-based exclusions for medical conditions that could otherwise warrant neuroimaging (n=98) were applied to the denominator with a unique flag. Claims-based exclusions for indicators of increased intracranial pressure (n=343) and seizure/convulsions (n=67) were applied to the numerator with unique flags.

To perform the sensitivity analysis of exclusion criteria, we varied the number of children among the 450 (originally classified as having abuse/neglect, or loss of consciousness, skull fracture, or intracranial hemorrhage) with the denominator exclusion of child abuse and neglect by 25%, 50%, 90% and 100%. In each scenario, children who were not excluded for abuse/neglect were counted as having numerator exclusions for loss of consciousness, skull fracture, or intracranial hemorrhage. We held constant the claims-based exclusions for medical conditions that could otherwise warrant neuroimaging (increased intracranial pressure, seizure/convulsions). In each scenario, we calculated the overuse percentage based solely on administrative claims data.

In the sample of abstracted charts (n=204), we determined the overuse percentage using exclusions that were identified in administrative claims before chart review. We subsequently calculated the overuse percentage using criteria abstracted during medical record review.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Table 8: Sensitivity Analysis with Variation in Claims-Based Exclusion Criteria among Children with Neuroimaging during an ED Visit for Post-Traumatic Headache, Concussion, or General Headache, N = 2967

	% Variation in Number of Children with Denominator Exclusion for Child Abuse and Neglect				
	0%	**25%**	**50%**	**90%**	**100%**
DENOMINATOR EXCLUSIONS					
Abuse and neglect	450	333	225	45	0
Medical conditions that could otherwise warrant neuroimaging	98	98	98	98	98
DENOMINATOR	2,419	2,536	2,644	2,824	2,869
NUMERATOR EXCLUSIONS					
Loss of consciousness, skull fracture, intracranial hemorrhage	0	117	225	405	450
Indicators of increased intracranial pressure	343	343	343	343	343
Seizure or convulsions	67	67	67	67	67
NUMERATOR	2,009	2,009	2,009	2,009	2,009
Overuse percentage using administrative claims	**83.0%**	**79.2%**	**76.0%**	**71.1%**	**70.0%**

Table 9: Overuse Percentage within the Chart Review Sample using Claims or Chart Review Criteria

	Chart Review Sample N=204	
	Criteria identified in Claims	**Criteria identified in Charts**
Denominator	198	57
Numerator	187	8
Overuse percentage	**94.4%**	**14.0%**

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The results of the exclusion analysis demonstrate that without the use of chart review, the overuse percentage would be substantially over-estimated 56 to 80 percentage points. Although the initial application of ICD-9-CM code-based exclusions decreases the burden of reviewing charts unlikely to meet final inclusion criteria for calculation of this measure as specified, the application of exclusions obtained exclusively from chart review substantially changes the neuroimaging overuse percentage as compared to claims alone. Therefore, identification of exclusions in both administrative claims and chart review are necessary for calculation of this measure. Lastly, it is important to note that there are key elements for this measure that cannot be captured in any form using administrative claims. For example there are currently no ICD-9-CM or ICD-10-CM codes for the major inclusion criteria of having an injury within 24 hours of the ED visit. The vast differences between

overuse percentages calculated using data available in administrative claims alone and through chart review justify the burden of chart review for the calculation of this measure.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

We calculated a single neuroimaging overuse percentage for the evaluation of children with post-traumatic headache within a sample of charts obtained by HealthCore after cases were narrowed using administrative claims from the HIRD. Due to the small sample size of charts eligible for inclusion in the numerator and denominator after chart review, we were unable to perform comparisons between health plans or hospital EDs.

A two-sided two-proportion z-test was conducted to determine if the observed overuse percentage in our sample was statistically different than the observed rate of overuse within the 2009 PECARN study of children with traumatic brain injury to develop a clincial decision rule for CT imaging. In addition, we conducted a one-sided one-proportion z-test to determine if the observed overuse percentage in our sample was greater than a theoretical target overuse percentage of 5%.

In order to inform future applications of this measure, we calcuated the sample size needed to achieve 95% confidence intervals with half widths of 2.5%, 5%, and 8% for anticpated overuse percentages ranging from 5% to 25%. We also calculated the per group sample size needed to detect 5% to 20% differences in two prorportions with power of 80, 90 and 95. We used a control overuse percentage of 15% for this calculation.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The overuse percentage in our chart review sample was 14.0%. The sample sizes needed to achieve 95% confidence intervals with half widths of 2.5%, 5%, and 8% for anticipated overuse percentages ranging from 5% to 25% are presented in Table 10.

Table 10: Sample Size Calculation for 95% Confidence Intervals around Expected Overuse Percentages

Expected Overuse Proportion	Required Sample Size for CI half-width=2.5%	Required Sample Size for CI half-width=5%	Required Sample Size for CI width=8%
5.0%	292	73	N/A
8.0%	292	114	45
10.0%	554	139	55
15.0%	784	196	77
20.0%	984	246	97
25.0%	1153	289	113

When comparing our overuse percentage (14.0%) with the rate of imaging in children lacking all of the six predictors of clinically important traumatic brain injury in the PECARN derivation sample (25.4%), a z-score of -2.0704 was obtained, corresponding to a two-sided p-value of 0.038. When comparing our overuse percentage (14.0%) with the rate of imaging in children lacking all of the six predictors of clinically important traumatic brain injury in the PECARN validation sample (24.1%), a z-score of -1.8339 was obtained, corresponding to a two-sided p-value of 0.067. When comparing our observed overuse percentage with the theoretical target overuse percentage of 5%, a z-score of 3.017 was obtained, corresponding to a one-sided p-value of 0.0013. The results of our sample size calculation to guide future testing of this measure are presented in Table 11.

Table 11: Per Group Sample Size for Two-Sample Proportion Test (control proportion 15%)

Power	5% Difference	10% Difference	15% Difference	20% Difference
80	906	250	121	73
90	1,212	335	161	97
95	1,498	413	199	119

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*) Despite the small number of charts eligible for our overuse percentage calculations, this measure was successfully able to distinguish statistically significant differences ($p<0.05$) in overuse of imaging from the historical PECARN rates around 25% and the theoretical target percentage of 5%. Sample size calculations for both 95% confidence interval half-widths and two-sample proportion tests are provided to guide appropriate sample size targets for use of this measure in quality improvement and quality performance reporting. A minimum of 196 charts included in the denominator after chart review would be recommended to obtain a 95% confidence interval with a 5% half-width around an expected overuse percentage of 15%. A per group minimum of 335 charts included in the denominator after chart review would be recommended for a two-sample proportion test to detect a 10% difference from a control proportion of 15% with power of 0.90 and alpha of 0.05.

The neuroimaging overuse percentage in this sample is significantly lower, both statistically and clinically, than unnecessary imaging rates reported for the derivation sample in the PECARN study. A reduction in the overuse of neuroimaging by more than 10 percentage points is clinically significant when considering the large number of children who undergo neuroimaging for evaluation of post-traumatic headache. However, overuse percentages in this sample remain significantly higher than a target percentage of 5%, which indicates less than optimal performance.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
N/A

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

Our results indicate that the data elements required for the calculation of this measure are typically recorded in electronic health record (EHR) systems. However, important information required for numerator or denominator exclusion criteria may be recorded in an unstructured format in problem lists, as well as in nursing and physician notes. Order entry systems can provide structured information about orders placed for neuroimaging studies; this furnishes key information necessary for future applications of the measure. Importantly, for this measure to be accurate, it may be necessary to combine data from multiple EHR systems. The use of Health Information Exchange (HIE), especially using the DIRECT protocol for exchange across individual electronic medical records (EMRs), would be an important tactical step to enable this measure.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

This measure was tested using medical record data after administrative claims were used to identify the population to sample for chart review. Administrative data needed for this measure include date of birth, diagnosis codes, and procedure codes and dates. These data are generally available, although obtaining them may require a restricted-use data agreement and Institutional Review

Board (IRB) approval.

Testing this measure using medical record data required the development of an abstraction tool and the use of qualified nurse abstractors. We provide an example data abstraction tool for chart review (see S.1. above). Review of clinical documentation was required to ensure that exclusions were appropriately captured for the determination of overuse of neuroimaging (i.e., imaging obtained in the absence of indicators of intracranial hemorrhage or basilar skull fracture).

Our review of medical charts indicated that 71.1% (145/204) of the children who were included in the chart review sampling population after the application of administrative claims exclusions were subsequently excluded from the denominator based on information in the medical chart. Importantly, the majority of numerator exclusions (i.e., symptoms of intracranial injury that represent a clinical indication for neuroimaging) were not adequately captured in administrative claims. As a consequence, using administrative data alone would result in a substantial overestimation of the degree to which neuroimaging is overused in the evaluation of children with post-traumatic headache. This finding is not unexpected, as there are several exclusions that can only be accurately captured through review of clinical documentation contained within the medical record. As an example, one denominator exclusion criterion, time of injury greater than 24 hours, cannot be identified through the use of administrative claims.

Chart review also may be beneficial to confirm that individuals with claims-based denominator exclusions have been appropriately identified and removed from the final eligible population, although we found high validity between data elements available within administrative claims compared with data elements documented within the medical chart (see testing form). Additionally, chart review is necessary to determine cases meet measure inclusion criteria for post-traumatic headache. Some of the ICD-9-CM codes used to identify cases for chart review were intentionally non-specific, such as 'general symptoms of headache' (ICD-9-CM code 784.0), as they reflect codes that are used in clinical practice to bill for care delivered to children with post-traumatic headache but require determination of a trauma history within 24 hours of the ED visit based on chart review.

This measure was tested using a target sample of 200 abstracted charts for eligible children during the measurement year. The yield of charts eligible after the application of denominator exclusions was lower than expected; 27.9% of the 204 charts abstracted for testing were eligible for the denominator. In addition, 67 of the 75 children had a 'general symptoms of headache' diagnosis code had no clinical documentation of trauma occurring within 24 hours of the ED visit. The inclusion of this non-specific code contributed substantially to the attrition of eligible cases. Larger samples of charts would be required for abstraction in order to ensure adequate sample size remains in the denominator after application of exclusion criteria. To detect differences between two health plans, hospital emergency departments, or providers with overuse percentages of 20% and 10% would require a sample size of at least 199 denominator eligible cases per group with a p-value of 0.05 and 80% power.

Continuing advances in the development and implementation of EHRs may prompt providers to document key elements needed for application of inclusion and exclusion criteria necessary for this measure. Advances would further allow for electronic capture of structured clinical information needed to determine if and when neuroimaging has been overused in the evaluation of children experiencing a post-traumatic headache.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
N/A

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Payment Program Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Not in use	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This measure provides families, purchasers, providers, and policy makers with a straightforward measure to assess any potential overuse of imaging in the ED involving the care of children with post-traumatic headache. The primary information needed for this measure comes from administrative claims and medical record data and includes basic demographics, diagnostic codes, and procedure codes and times of services, all of which are widely available.

The measure is in the process of being submitted to the National Quality Measures Clearinghouse of the Agency for Healthcare Research and Quality as part of the Pediatric Quality Measures Program (PQMP). Measure dissemination efforts are currently in progress. We anticipate that NQF endorsement will lead to widespread use of the measure by purchasers, providers, and policy makers.

The only issue impeding broad use is dissemination of the measure and NQF endorsement.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Implementation plan:

Year 1-2: Dissemination of measure specifications to targeted certifying/regulatory organizations (e.g., Joint Commission), health plans, hospital groups, professional organizations (i.e., Children's Hospital Association), payers, CMS and other government agencies, and health care institutions (hospitals, urgent care centers).

Year 2-3: Public reporting of initial data collected from measure use across hospitals. This may occur in a number of ways: (1) via a collaborative approach with a group of hospitals through an organization such as the Children's Hospital Association, (2) assessment of data from Medicaid claims, (3) data aggregation from payers regarding their enrolled patients, or (4) data aggregation across hospital groups that implement the measure (e.g., Tenet).

Year 3: Refinement and review of updated literature available following initial publication of the measure. Re-calibration of the measure based on updated specifications and initial data from use.

Year 4-5: Re-introduction and ongoing testing of measure.

Year 6: Refinement and review of updated literature available following initial publication of the measure. Public reporting of measure performance at the level of the health care facility.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No unintended negative consequences were identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0668 : Appropriate Head CT Imaging in Adults with Mild Traumatic Brain Injury

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
N/A

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

This pediatric measure is aligned with NQF Measure number 0668: Appropriate Head CT Imaging in Adults with Mild Traumatic Brain Injury. The measures are harmonized in terms of the basic clinical criteria (imaging obtained in the ED within 24 hours of a head injury among patients with a Glasgow Coma Scale score of greater than or equal to 14) used to identify the population eligible for inclusion in the denominator. The pediatric measure differs in several ways, including consideration of current trends in neuroimaging, the ability to use administrative claims to narrow the population considered eligible for the more labor intensive chart review process, and the available evidence on the need for neuroimaging of children with post-traumatic headache. The endorsed adult measure is focused on CT imaging alone; this pediatric measure was tested to assess the overuse of neuroimaging more broadly, including both CT and MRI, for children who are evaluated for post-traumatic headache. The inclusion of MRI is important with recent shifts toward imaging modalities that avoid radiation exposure but still subject patients to risks from sedation/anesthesia, incidental findings, and costs associated with overuse of imaging studies. The pediatric measure was tested in a two-stage approach that first used administrative claims to identify the potentially eligible population and then used chart review in order to account for exclusions that could be documented in the provider notes but not captured with a relevant ICD-9-CM code. We included an extensive list of ICD-9-CM codes indicative of conditions in which neuroimaging for post-traumatic headache could be warranted (for example, coagulopathy or cerebral cyst) in order focus this measure on clear cases of overuse of neuroimaging. Finally, we applied the specific factors that were identified by Kuppermann and colleagues as relevant to the risk of clinically important brain injury in children with minor trauma based on results of the largest prospective cohort study of pediatric traumatic brain injury.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

There is overlap between the NQF Measure number 0668: Appropriate Head CT Imaging in Adults with Mild Traumatic Brain Injury and the pediatric measure in that the adult measure includes children 16 to 18 years old. The pediatric measure is more narrowly focused on children between 2 and 18 years of age. Because of the unique nature of child illness and injury, a pediatric-focused measure is needed.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** Q-METRIC_IMG_Post-TraumaHD_NQF_Appendix.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Q-METRIC – The University of Michigan
Co.2 **Point of Contact:** Michelle, Macy, MD, MS, mlmacy@med.umich.edu, 734-936-8338-
Co.3 **Measure Developer if different from Measure Steward:** Q-METRIC – The University of Michigan
Co.4 **Point of Contact:** Gary, Freed, MD, MPH, gfreed@med.umich.edu, 734-232-0657-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

The face validity of this measure was established by a national panel of experts and parent representatives for families of children with headaches and seizures convened by Q-METRIC. The Q-METRIC Representative Panel included nationally recognized experts in the area of imaging children, representing general pediatrics, pediatric radiology, pediatric neurology, pediatric neurosurgery, pediatric emergency medicine, general emergency medicine, and family medicine. The Q-METRIC Feasibility Panel included experts in state Medicaid program operations, health plan quality measurement, health informatics, and health care quality measurement. In total, the Q-METRIC imaging panel included 15 experts, providing a comprehensive perspective on imaging children and the measurement of quality metrics for states and health plans.

The Q-METRIC expert panels concluded that this measure has a high degree of face validity through a detailed review of concepts and metrics considered to be essential to appropriately imaging children. Concepts and draft measures were rated by this group for their relative importance. This measure was highly rated, receiving an average score of 7.0 (with 9 as the highest possible score).

Representative Panel:
Dana Cook, Parent Representative, Paw Paw, MI
Peter Dayan, MD, MSc, Division of Pediatric Emergency Medicine, Morgan Stanley Children´s Hospital, New York, NY
Lisa Dover, Parent Representative, Ann Arbor, MI
Danny Greig, MD, FAAFP, Emergency Room Physician, MidMichigan Medical Center, Midland, MI
Blaise Jones, MD, Director of Clinical Services, Chief of Neuroradiology, Cincinnati Children´s Hospital Medical Center, Cincinnati, OH
Steven Leber, MD, PhD, Professor of Pediatrics and Neurology, University of Michigan, Ann Arbor, MI
Cormac Maher, MD, Associate Professor of Neurosurgery, University of Michigan, Ann Arbor, MI
L. Kendall Webb, MD, Assistant Professor of Emergency Medicine, Director of IT for the Emergency Department, University of Florida, Jacksonville, Jacksonville, FL
Neal Weinberg, MD, FAAP, General Pediatrician, IHA Pediatric Healthcare – Arbor Park, Ann Arbor, MI

Feasibility Panel:
Cathy Call, RN, BSN, MSEd, MSN, CPHQ, PMP, LSS, Practice Area Leader for Health Quality Research, Health Care Analytics Group, Altarum Institute, Alexandria, VA
Andrea DeVries, PhD, Director of Research Operations, HealthCore Inc., Wilmington, DE
J. Mitchell Harris, PhD, Director of Research and Statistics, Children's Hospital Association, (formerly NACHRI), Alexandria, VA
Don Lighter, MD, MBA, Director, The Institute for Health Quality Research and Education, Knoxville, TN
Sue Moran, BSN, MPH, Director of the Bureau of Medicaid Program Operations and Quality Assurance, Michigan Department of Community Health, Lansing, MI
Stuart Weinberg, MD, Assistant Professor of Biomedical Informatics, Assistant Professor of Pediatrics, Vanderbilt University, Nashville, TN

Q-METRIC Investigators:
Michelle L. Macy, MD, MS, Assistant Professor, Departments of Emergency Medicine and Pediatrics, School of Medicine, University of Michigan, Ann Arbor, MI
Gary L. Freed, MD, MPH, Professor of Pediatrics, School of Medicine and Professor of Health Management and Policy, School of Public Health, University of Michigan, Ann Arbor, MI (principal investigator)
Kevin J. Dombkowski, DrPH, MS, Research Associate Professor of Pediatrics, School of Medicine, University of Michigan, Ann Arbor, MI

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?** N/A
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** N/A
Ad.7 **Disclaimers:** This work was funded by the Agency for Healthcare Research and Quality (AHRQ) and the Centers for Medicare & Medicaid Services (CMS) under the CHIPRA Pediatric Quality Measures Program Centers of Excellence grant number U18 HS020516. AHRQ, in accordance to CHIPRA 42 U.S.C. Section 1139A(b), and consistent with AHRQ´s mandate to disseminate research results, 42 U.S.C. Section 299c-3, has a worldwide irrevocable license to use and permit others to use products and materials from the grant for government purposes, which may include making the materials available for verification or replication by other researchers and making them available to the health care community and the public, if such distribution would significantly increase access to a product and thereby produce substantial or valuable public health benefits. The Measures can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial

use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the measures require a license agreement between the user and the Quality Measurement, Evaluation, Testing, Review and Implementation Consortium (Q-METRIC) at the University of Michigan (U-M). Neither Q-METRIC/U-M nor their members shall be responsible for any use of the Measures. Q-METRIC/U-M makes no representations, warranties or endorsement about the quality of any organization or physician that uses or reports performance measures, and Q-METRIC/U-M has no liability to anyone who relies on such measures. The Q-METRIC performance measures and specifications are not clinical guidelines and do not establish a standard of medical care.

This statement is signed by Gary L. Freed, MD, MPH, who, as the principal investigator of Q-METRIC, is authorized to act for any holder of copyright on the submitted measure.

Gary L. Freed, MD, MPH
Percy and Mary Murphy Professor of Pediatrics, School of Medicine
Professor of Health Management and Policy, School of Public Health
Principal Investigator, Q-METRIC
Child Health and Evaluation Research (CHEAR) Unit
Division of General Pediatrics
University of Michigan Hospital and Health Systems
Ann Arbor, MI 48109-5456

Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2803
Measure Title: Tobacco Use and Help with Quitting Among Adolescents
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: Percentage of adolescents 12 to 20 years of age during the measurement year for whom tobacco use status was documented and received help with quitting if identified as a tobacco user.
Developer Rationale: The Tobacco Use and Help with Quitting Among Adolescents measure addresses an issue of significant importance. Tobacco use can have both immediate and long-term serious health consequences, yet data show that, despite some successes, many adolescents continue to begin or use tobacco products. Research has shown that health care providers can play an important role in promoting tobacco-use abstinence and cessation. Thus, this measure encourages standardized documentation of tobacco use status among adolescents and appropriate follow-up for those who are users.

Numerator Statement: Adolescents who are not smokers OR Adolescents who are smokers but are receiving cessation counseling.
Denominator Statement: Adolescents who turn 12 through 20 years of age during the measurement year.
Denominator Exclusions: N/A

Measure Type: Process
Data Source: Electronic Clinical Data
Level of Analysis: Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this process measure:

- The Level of Analysis is Clinician/Group Practice.
- This measure addresses standardized documentation of tobacco use status among individuals ages 12-20 years adolescents as well as appropriate follow-up for those who are users.
- The evidence for this process measure is based on guidelines/recommendations of two bodies: the U.S. Preventive Services Health Task Force (USPSTF) and the American Academy of Pediatrics (AAP). The developer focuses on the USPSTF recommendations because it is a systematic review; the AAP recommendations are provided within an AAP policy statement.
  - USPSTF Overall Recommendation: The USPSTF recommends that primary care clinicians provide

interventions, including education or brief counseling, to prevent initiation of tobacco use among school-aged children and adolescents. Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.
    - USPSTF Recommendation 1: Clinicians should ask pediatric and adolescent patients about tobacco use and provide a strong message regarding the importance of totally abstaining from tobacco use. Grade C Strength of Evidence—"reserved for important clinical situations in which the Panel achieved consensus on the recommendation in the absence of relevant randomized controlled trials."
    - USPSTF Recommendation 2: Counseling has been shown to be effective in treatment of adolescent smokers. Therefore, adolescent smokers should be provided with counseling interventions to aid them in quitting smoking. Grade B Strength of Evidence—"some evidence from randomized clinical trials supported the recommendation, but the scientific support was not optimal. For instance, few randomized trials existed, the trials that did exist were somewhat inconsistent, or the trials were not directly relevant to the recommendation."
    - The systematic review assessed the Quantity, Quality, and Consistency of the literature: 19 trials, of which 4 were rated "good" and 15 were rated "fair".
    - The USPSTF found no evidence on the harms of behavioral interventions to prevent tobacco use and concluded the magnitude of potential harms is probably small to none.
    - Overall, the USPSTF concluded with moderate certainty that primary-care interventions to prevent tobacco use in school-aged children and adolescents have a moderate net benefit.
- The developer notes that while the USPSTF recommendation focuses on primary-care based intervention, the measure includes assessment. The developer notes that USPSTF assumes assessment as a logical necessary precursor to intervention.
- Per the NQF Evidence Algorithm, the evidence is based on a systematic review that includes grading and an articulation of the quantity, quality, and consistency of the evidence and so is eligible for a HIGH, MODERATE, or LOW rating (box 5a).

Question for the Committee

- *Has the developer established a clear relationship of this measure to patient outcomes that is supported by evidence and is it of [HIGH, MODERATE, LOW] strength?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following information:

- Data from the 2011 National Youth Tobacco Survey show that only 32% of adolescent respondents reported being asked about tobacco use or 31% were advised about abstaining/quitting by a healthcare provider.
- The developer reports that results from testing the measure (2010-2011 data) found the overall rate of adolescents with documentation of tobacco use and help with cessation was 61.6%, with a range of 44.5% to 85.3% across three testing sites.
- Non-Hispanic white and African American adolescents had similar rates of tobacco use/help documented (65.9% and 66.8%, respectively). Hispanic/Latino and Asian/Native American/Pacific Islander adolescents had lower rates (38.1% and 25.0%, respectively). Those of other/multiple races had a rate of 54.6%.
- Respondents with commercial insurance who received help with quitting reported a rate of 82%, while Medicaid patients had a rate of 60% and those with other insurance statuses a rate of only 39%.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Is the Committee aware of performance gap information since 2011?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when*

performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- USPST B recommendation supports asking tobacco use status, however, not for assistance to quit. What evidence is there to support a specific intervention from a physician will result in tobacco cessation.
- adequate evidence
- The proposed measure is focused on a process of care and has two distinct subparts. One is the recommendation that all adolescents be screened for tobacco use and the second is that counseling and assistance with quitting be offered to those who screen positive. The logical connection to health outcomes is clear, given the enormous body of evidence of the harms of tobacco use.
- The authors provide documentation that counseling interventions in the primary care setting may be effective, based on a number of RCTs. These have weaknesses and have shown mixed results, but the US Public Health Service gave this a B rating. The USPTF has give this a grade of B, and includes in its recommendation both counseling for prevention and intervention (for current tobacco users). The evidence for screening/prevention counseling comes from evidence of lack of harm so USPTF concludes moderate certainly of moderate benefit. An evidence review by the EPC found that only 4 of 19 RCTs were of good quality (15 were fair), on which the above conclusions seem to be based.
- In summary, there is sufficient evidence for the measure (at the level of Moderate), and the recommendations by the bodies above add to its credibility.
- Process measure - the developer cites the USPSTF recommendations of Grade C support for the portion of the measure focus related to identifying tobacco use and Grade B support for the portion of the measure related to referral for interventions in those identified as tobacco users. The USPSTF review included almost 20 studies graded as "good" or "fair." The relationship of the measure to outcomes is supported by the USPSTF recommendations and would fall under NQF MODERATE criteria, I believe.
- Limited evidence base to support the measure.
- Evidence in the form of professional consensus guidelines
- In this case, it appears that the evidence applies tangentially to the goal of decreasing tobacco use in children and adolescents. If, during well child visits, clinicians address the need to abstain or quit smoking and provide support, the incidence of tobacco use will decrease. SR of evidence indicates a Moderate level by box 5A. Also Grade B by USPSTF is noted. The relationship between the measured outcome and the clinician addressing this issue as proposed is identified and supported by the rationale.

1b. Performance Gap.

- There appears to be gaps in performance. Highest disparities appear to be with payer source of care.
- Measure shows performance gap exists with differences amongst minority groups.
- Information is provided that documents a considerable performance gap, though the magnitude of that gap is variable depending on the source of the information (2011 survey data vs. the chart review at 3 clinical sites undertaken by the developers).
- There does seem to be disparity by race/ethnicity, with teens of Hispanic/Latino and Asian/Pacific Islander race/ethnicity having lower rates.
- The data cited demonstrates a performance gap. I would be curious to see more recent data with implementation of meaningful use to see if these numbers are increased more recently. There appears to be significant disparity between the non-Hispanic white and AA adolescents as compared to Hispanic/Latino and Asian/Native American/Pacific Islander adolescents.
- Performance gaps exist for Hispanic youth as well as those with Medicaid, suggesting the need for a performance measure.
- This measure should be considered sensitive to disparities due to performance gaps by SES and minority status.
- Yes, several surveys were noted with a variety of scores for intervention/addressing the smoking issue. Scores ranged: 2011 study - 32% reported being asked about tobacco use to 82%. This high group was from commercial

insurers. Medicaid covered patients saw ranges of 30s to 60s. So, yes, there is a gap. It appears to me that the overall quality of studies show a fair to good quality by algorithm.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The developer defines the numerator as: Adolescents who are not smokers OR Adolescents who are smokers but are receiving cessation counseling. The denominator is defined as: Adolescents who turn 12 through 20 years of age during the measurement year and had documentation of a face-to-face visit with a primary care practice during the 12 months prior to the measurement year.
- Type of score is by rate and better quality is associated with higher score. Detailed steps for the calculation algorithm are provided, which appears straightforward.
- The measure is not risk-adjusted.

Questions for the Committee

- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer reports the following:

- The measure was tested using data abstracted from electronic health records for group/practice level of analysis.
- Empirical testing was performed at three pediatric centers during the periods: October 1, 2010 to December 31, 2011, and October 1, 2010 and March 31, 2012, depending on the test.
  - Pediatric clinics were 1) affiliated with a children's hospital (this sample was selected from adolescents enrolled in Medicaid); 2) a network of clinics serving homeless and vulnerable adolescents; and 3) an adolescent medicine clinic affiliated with a children's hospital (which primarily provides behavioral health and gynecology care to young women).
  - The participating sites were in different states and used different EHR vendors.
- The testing dataset was consistent with the measure specifications for the target populations and reporting entities.
- Reliability testing was done at the level of data elements using a sub-sample of 75 adolescents from the initial sample of 597 (25 from each of the three field testing clinical sites).
  - Inter-rater reliability was performed, yielding a Kappa coefficient for smoking status (based on the *Meaningful Use* definition) of 0.94, 95% Confidence Interval 0.84, 1.0. Kappas for five other data elements were not provided because neither abstractor found these data elements in the charts sampled. In this regard, the Kappa could be considered 1.0. The five data elements were: current tobacco use; documentation of advice to quit smoking/using tobacco; counseling on the benefits of quitting smoking/using tobacco; referral to smoking/tobacco cessation support program; enrolled in a

smoking/tobacco use program.
    - Comparison between manual chart review to an automated EHR extraction using the full sample of 597 records was performed. Specific to the proposed measure, the percentage of adolescents whose tobacco use is documented and who received help with quitting if they are users was 61.6% in the manual review versus 47.4% in the automated extract, with a Kappa of 0.52 (moderate agreement). Additionally, there was documentation of tobacco status for 70.9% of adolescents in the manual reviews compared to 53.9% in the automated extracts, with a Kappa of 0.52 (moderate agreement), and 13.6% of adolescents were identified as smokers in the manual reviews compared to 7.7% in the automated extract, with a Kappa of 0.66 (substantial agreement). The developer reports "substantial variations by site both in the results and the agreement between manual review and automated EHR extract."
- Per the NQF Evidence Algorithm, empirical reliability testing with patient-level data (boxes 8-10) may be rated MODERATE, LOW, or INSUFFICIENT.

Questions for the Committee

- *Are the test samples adequate to generalize for widespread implementation?*
- *Is the testing complete for the manual inter-rater reliability given the lack of reliability statistics for five of the data elements because they were not present in the sampling of charts? That is, testing included the data elements and detecting the absence of the data elements was highly reliable (Kappa = 1.0) , but does the Committee wish to comment on reliability to detect the data elements?*
- *Does the Committee find the EHR to manual abstraction reliability testing results demonstrate sufficient reliability so that differences in performance can be identified?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The goal of the measure is to assess how many adolescents are receiving counseling/education for tobacco use cessation. The numerator is the number of adolescents who are not smokers OR are smokers but are receiving cessation counseling. The denominator is adolescents who had documentation of a face-to-face visit with a primary care practice.
- The specifications are consistent with the USPSTF recommendations; assessment is included in the measure, as well as the intervention (counseling), because, as noted by the developer, it is a necessary precursor to the intervention.

Question for the Committee

- *Does the Committee concur that the specifications are consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer reports the following information:

- Empirical testing was performed at the performance measure score.
    - The developer hypothesized that the performance rates for adolescents who had one or more well-care child visits during the measurement period would exceed those for adolescents who had no well-care visits during the period October 1, 2010 to December 31, 2011.
    - Documentation of tobacco use and help with quitting was significantly higher among adolescents who had at least one well-care visit in the measurement period compared to adolescents without designated well

care visits. Specifically, the results were (N=400):
    - Had Well-Child Visit(s): 58.9% of adolescents whose tobacco use was documented and who received help with quitting if they were a user vs. No Well-Child Visit = 39.2%.
    - The developer reports the results were significant (p-value <0.0001).
- Face validity also was assessed by a multi-stakeholder expert panel along with three targeted multi-stakeholder panels. All concluded the measure is a valid way to assess tobacco status, use, and follow-up among adolescents, which we interpret as assessment at the performance score level. (Per the NQF guidance and Validity Testing Algorithm, face validity must be assessed at the computed measure score, not measure construct, data element accuracy, feasibility, etc. [box 5]).
- Per the NQF Algorithm for Validity Testing, empirical testing at the computed measure score level score may be rated HIGH, MODERATE, LOW, or INSUFFICENT depending on the certainty or confidence that the performance scores are a valid indicator of quality (boxes 6-8). Reliance *only* on face validity of the computed measure score means the rating may be MODERATE, LOW, or INSUFFICIENT (boxes 4-5).

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- There are no exclusions.

Questions for the Committee:

- *Should there be any exclusions for this measure?*
- *Does the Committee believe there are other threats to validity?*

2b4. Risk adjustment:

- The measure is not risk adjusted.

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer states that it did not have sufficient data to conduct statistical tests to identify meaningful differences. The developer presents only performance rates from each of the testing sites.
- The overall rate of tobacco use and help with quitting among adolescents was 61.6%, with rates at the three sites of 44.5%, 55.5%, and 85.3%.
- The developer acknowledges that site-to-site variation can be explained, in part, by differences in the availability of data elements, content of free-text notes, and site characteristics. The developers believe, however that "variations in these results would imply variations among providers."

Question for the Committee:

- *Although no statistical analyses are provided, does the Committee feel this measure identifies meaningful differences in quality?*

2b6. Comparability of data sources/methods:

- This is not needed.

2b7. Missing Data

- According to the developer, there are no missing data, so this is not applicable.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- The numerator statement seems to include two separate and distinct groups: adolescents who don't smoke and adolescents who do smoke and have been counseled to quit. The numerator seems to have two intentions: tobacco use status; and, if tobacco user, and advised to quit. Why are these two groups in the numerator combined, rather than having 2 separate quality measures?
- Measure adequately shows why tobacco screening and help with quitting is high priority.
- The measure is generally well specified, but like all measures based on documentation in the medical record there will be borderline cases. These should be further specified. For example, if a chart says "non-smoker" will that be taken as evidence that the teen is not a tobacco user in other ways (e.g. chewing tobacco, snuff).
- In 5.5 the authors state a look back period of 18 mos. I'm not sure if that's correct given that the measure could include any one turning 12-20 in the measurement year, and the measurement is about visits in the year PRIOR to the measurement year. It seems that at least 24 months of data is required.
- From S.18 it's not clear that the visit to be reviewed is from the calendar year prior to the measurement year (which is how I interpret "the 12 months prior to the measurement year" in S.9. It's possible that the developers did not intend this, and would measure any visit within or before the measurement year.
- The test samples are small - 3 pediatric clinics in different states using different EHRs. EHRs faired relatively poorly compared to manual review for each of the data points. I don't think the inter-rater reliability can be commented upon when the data elements aren't found in the charts sampled. I'll look forward to discussion of the evidence algorithm - I'm not sure how to classify the empirical evidence submitted.
- Data elements are clearly defined and the calculation algorithm is clear. It appears the measure can be consistently implemented.
- The specifications are consistent with the evidence.
- The data elements are clearly defined. The steps in the logic are clear.
- For this measure to succeed, we must accept that if the clinician addresses the smoking issue and intervenes if there is smoking, the rate of smoking will decrease.

2a2. Reliability testing

- The specifications need additional clarity. For example, it is not clear what is meant by "receiving cessation counseling" which is not the same in other areas of the document. In other areas, other terminology is used: "received help" "intervention" "assistance with quitting" "help with quitting"
- Concern re: most effective strategies to assist with quitting smoking. no mention of motivational interviewing strategies.
- Reliability testing (for manual abstraction) was performed on 25 records from each of three clinical settings. Reliability (kappa) is good for chart review by different abstractors. However, there was less good detection and agreement between chart review and automated extraction (with the former showing evidence of higher performance).
- By the algorithm there is Moderate reliability for the manual abstraction only, and Low reliability for the automated extract.
- Measure may not be consistently applied in various practice settings. Inquiries about smoking and cessation activities have been found to be sporadic in practice settings.
- I have some concern that testing of a chart review measure in only 3 clinical settings is insufficient to understand the wide variation in how things might be documented.
- The high agreement is present only because most of the time there was no mention of the data elements. This does not provide an estimate of agreement when such elements are present.
- Reliability testing was carried out in 3 pediatric centers with a somewhat varied populations Inter rater reliability testing was used with a Kappa Coef. for smoking status. Results: 0.94, 95% confidences level. EHR and manual review of charts was used - Kappa - 0,52 =moderate agreement.
- Using the Algorithm - Moderate reliability

2b1. Validity Specifications

- The numerator statement is ambiguous with the combination of 2 groups and the lack of clarity for "receiving cessation counseling"

- If automated data extraction is the only methodology used when the measure is operationalized, the measure is not reliable.
- No obvious concerns.
- The major issue for this measure is whether these two things-- documentation of no tobacco use and counseling for cessation among users belong in the same measure. Clearly, cessation depends on knowing a teen is a tobacco user. However, clinicians could have a high score by screening while providing cessation counseling to a relatively small proportion (depending on the frequency of tobacco use in their practice). The developers should discuss the pros and cons of having these reported as two separate measures.
- The specifications seem consistent with the evidence.
- Kappa coefficients could not be calculated for most data elements because they were not documented.
- Sample size was low for the reliability testing.
- The specifications are consistent with the evidence.

2b2. Validity Testing

- How many tobacco users were in the numerators for each test site?
- Face validity was assessed by the developers by convening a number of stakeholder groups including patients, parents, physicians, health plans and Medicaid Directors. Partnerships with AAP to National Partnership for Women and Families were also useful for convening. There is limited information on what these groups discussed or whether there was any significant variability in the views of the various stakeholders.
- The "known groups" validity test provides little additional empirical evidence for validity. The result is expected based on typical patterns for documentation.
- The face validity process suggests Moderate validity. But, there remains the issue of whether the two parts of this measure would be better reported separately.
- Established face validity.
- It appears that the specifications are consistent with the evidence presented.
- Well care visits were highly correlated with documentation of tobacco use and assistance with quitting.
- Validity testing involved a study of 400 patients who had well care visits where a 58% documented tobacco use and received help to stop smoking.
- Face validity was also noted.
- Using the algorithm (box 5) there appears to be sufficient validity for the use of this measure.
- I am not able to respond to some of these questions for this measure.

2b3-2b7. Threats to Validity

- Would like to have seen the test site for behavioral health included since tobacco use is higher among populations with behavioral health conditions.
- No
- Missing data is described as not being an issue. This is because, by definition, the absence of documentation is considered a failure. It is possible that some clinicians did counsel patients, but it was not recorded in the EHR. The measure will favor EHRs that drive documentation of this particular service.
- The developer states there are no missing data.
- No missing data reported.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer states:

- These elements can be manually abstracted from a healthcare provider's record.
- Some components of this measure are aligned with the Meaningful Use definition of tobacco use status, and the remaining are available in structured fields, narrative notes, or other non-structured fields.
- The measure has been specified as an eMeasure, but is not being submitted at this time (though it may be in the future).
- Results from testing demonstrate that the measure is feasible for clinicians to report.
- The measure has been added to the Physician Quality Reporting system (PQRS)for 2015, so there is no experience on operational use yet.

Questions for the Committee

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Why is PQRS being used within the context of this measure. PQRS is primarily a Medicare initiative. WOuld pediatricians be engaged in PQRS?
- No concerns
- The measure appears feasible with the caveat that it requires manual abstraction. The relationship to meaningful use is complex, and the developers do not provide information on how this will make electronic abstraction more effective in the future, compared to the results of the current test.
- The measure is currently included in the PQRS by CMS.
- There is harmonization with another NQF measure targeting adults.
- These elements can be manually abstracted from a chart and are acquired in the course of routine patient care, but cannot reliably be extracted from electronic sources as demonstrated by the developer.
- Information could be gathered from structured fields in the EHR. Currently, many of the elements would be available in un-structured formats, making it a challenging measure to implement across sites.
- It appears that in some cases the required information must be abstracted from health care provider records and therefor is less available that easily found dat in the EHR.
- It is apparent that the data is included in CMS Meaningful Use and as of 2015 in the Physician Quality Reporting System (PQRS), so the data elements can be generated.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

The developer reports:

- The measure is currently in use in PRQS for 2015 and the EHR Incentive Program (Meaningful Use).
- This is a new measure and improvement results are not yet available.
- No unintended consequences have been reported thus far.

Question for the Committee

o *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- If this measure is used for any provider P4P program, those providers who see a higher proportion of smoking adolescents will have a greater challenge in meeting performance expectations than those who see a lower proportion of smoking adolescents.
- I can't identify any unintended adverse consequences nor do I think these would outweigh the benefits of this measure.
- Measure is being used by the Physician Quality Reporting System. Tobacco Use is included in Meaningful Use, and can be included in the EHR incentive program.

Criterion 5: Related and Competing Measures

- This measure, #2803, is related to one NQF-endorsed measure, NQF 0028: Preventive Care & Screening: Tobacco Use: Screening & Cessation Intervention.
- NQF 0028 has a different target population (18 years and older), while this measure covers the population from 12 years to 20 years of age during the measurement period.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Click here to enter measure title

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: Click here to enter a date

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Adolescent Tobacco Cessation

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Adolescent is seen by physician >> Physician assesses the adolescent as a tobacco user or non-tobacco user >> If adolescent is a tobacco user, physician provides assistance with quitting >> Adolescent ceases using tobacco >> Adolescent's risk of developing tobacco-related morbidity/mortality decreases

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☒ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Fiore MC, Jaén CR, Baker TB, et al. Treating Tobacco Use and Dependence: 2008 Update. Clinical Practice Guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service. May 2008.

http://www.ahrq.gov/sites/default/files/wysiwyg/professionals/clinicians-providers/guidelines-recommendations/tobacco/clinicians/update/treating_tobacco_use08.pdf

American Academy of Pediatrics. Committee on Environmental Health, Committee on Substance Abuse, Committee on Adolescence, and Committee on Native American Child Health. 2009. Tobacco Use: A Pediatric Disease. Pediatrics 124(5): 1474. [Reaffirmed May 2013]

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

U.S. Public Health Service

Recommendation 1: Clinicians should ask pediatric and adolescent patients about tobacco use and provide a strong message regarding the importance of totally abstaining from tobacco use.

Recommendation 2: Counseling has been shown to be effective in treatment of adolescent smokers. Therefore, adolescent smokers should be provided with counseling interventions to aid them in quitting smoking.

American Academy of Pediatrics

For patients and their family members

• Counsel children and parents about the harms of tobacco use.

• Include tobacco in all discussions of substances of abuse and risky behaviors. Discussion and anticipatory guidance about tobacco use should ideally begin by 5 years of age and emphasize resisting the influence of advertising and rehearsal of peer refusal skills. Be aware of confidentiality issues related to tobacco use and other substance abuse, including testing for nicotine and its metabolites.

• Encourage parents to start discussions of tobacco use with their children early in their life and continue to do so throughout childhood and adolescence; these discussions should include delivery of clear messages disapproving of tobacco use. Both parents and children should be counseled that it is not safe to “experiment” with tobacco, because nicotine is so highly addictive and there is no safe way to use tobacco. Tobacco dependence can begin almost as soon as use begins, with some users exhibiting signs of dependence with only occasional or monthly use.83,84 As a result, prevention of tobacco use is one of the most important messages you can deliver.

For patients or family members who use tobacco

• Advise all families to make their homes and cars smoke free, and urge all tobacco users to quit. Provide appropriate advice and counseling to foster tobacco users to quit. Routinely offer help and referral to those who use tobacco— even if the person is not your patient. Be familiar with evidence-based guidelines for treatment of tobacco use and dependence and apply them to patients and their families.14 There is a growing body of literature on the effectiveness of pediatric clinician-provided treatment for parental nicotine addiction that demonstrates a role for pediatricians in this effort.

- Pharmacotherapy is an effective component of tobacco use-cessation treatment in adults. Encourage tobacco users to include these medications in their quit plan, whenever appropriate. Be familiar with and

offer information and instruction on correct use. Many nicotine replacement products are available without a prescription, although prescriptions are required for any nicotine-containing product if the patient is younger than 18 years.

- Pediatricians who choose not to prescribe pharmacotherapies should make referrals to cessation services and recommend that parents discuss pharmacotherapies with their health care providers or purchase over-the-counter products.

• Be familiar with tobacco use– cessation services in your community and provide referrals to these programs for your patients and their families. Memorize the national quit line telephone number (1-800-QUIT NOW), prominently post it, and provide it to all tobacco users. Whenever possible, proactively enroll tobacco users in cessation programs, using "fax-back" or similar programs. Such referrals are more effective in connecting the tobacco user to the resource than referrals that require the tobacco user to initiate the contact.

• Counsel all parents, including those who smoke, on how to deliver anti-tobacco messages and ways to discuss the addictive nature of nicotine.

- When parents or caregivers use tobacco, their children are more likely to experiment with tobacco and to begin to use tobacco regularly. Maintain a high index of suspicion for early onset of tobacco use by these children. It can be a particularly powerful message when the parent or caregiver who uses tobacco advises the child never to start using tobacco.
- Help patients and families understand that even casual use of tobacco by children and adolescents, regardless of amount or frequency, is illegal and associated with adverse health consequences.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

U.S. Public Health Service

Recommendation 1: Strength of Evidence C - Reserved for important clinical situations in which the Panel achieved consensus on the recommendation in the absence of relevant randomized controlled trials.

Recommendation 2: Strength of Evidence B - Some evidence from randomized clinical trials supported the recommendation, but the scientific support was not optimal. For instance, few randomized trials existed, the trials that did exist were somewhat inconsistent, or the trials were not directly relevant to the recommendation.

American Academy of Pediatrics (AAP)
The guidelines above are from an AAP Policy Statement.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

U.S. Public Health Service

Strength of Evidence A: Multiple well-designed randomized clinical trials, directly relevant to the recommendation, yielded a consistent pattern of findings.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

NA

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section <u>1a.7</u>***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Final Update Summary: Tobacco Use in Children and Adolescents: Primary Care Interventions. U.S. Preventive Services Task Force. July 2015.

http://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/tobacco-use-in-children-and-adolescents-primary-care-interventions

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

The USPSTF recommends that primary care clinicians provide interventions, including education or brief counseling, to prevent initiation of tobacco use among school-aged children and adolescents.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Grade A: The USPSTF recommends this service. There is high certainty that the net benefit is substantial.

Grade C: The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small

Grade D: The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. Discourage the use of this service. I statement.

I Statement: The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

http://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

We have provided two guidelines to support this measure: the U.S. Preventive Services Task Force (USPSTF) Guideline, which focuses on primary care interventions to prevent tobacco use initiation in children and adolescents, and the U.S. Public Health Service, which recommends clinicians discuss the risk of tobacco use and offer counseling on tobacco cessation to active tobacco users.

In this section, we focus on the findings of the systematic review of the evidence upon which the U.S. Preventive Services Task Force based its recommendation.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The Evidence-based Practice Center (EPC) that conducted the review assessed the evidence for the efficacy and harms of primary care-relevant interventions that aim to reduce tobacco use among children and adolescents.

Though the USPSTF focuses on primary-care based interventions and our measure includes assessment, in the USPSTF guideline, the assessment is assumed and a logical necessary step towards providing interventions.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Evidence included 19 trials of which four were good-quality and 15 were fair-quality trials.

Randomized Controlled Trials and Cohort Studies Criteria:

Initial assembly of comparable groups: ◦ For RCTs: adequate randomization, including first concealment and whether potential confounders were distributed equally among groups.

◦ For cohort studies: consideration of potential confounders with either restriction or measurement for adjustment in the analysis; consideration of inception cohorts.

Maintenance of comparable groups (includes attrition, cross-overs, adherence, contamination).

Important differential loss to follow-up or overall high loss to follow-up.

Measurements: equal, reliable, and valid (includes masking of outcome assessment).

Clear definition of interventions.

All important outcomes considered.

Analysis: adjustment for potential confounders for cohort studies, or intention to treat analysis for RCTs.

Definition of ratings based on above criteria:

Good: Meets all criteria: Comparable groups are assembled initially and maintained throughout the study (follow-up at least 80 percent); reliable and valid measurement instruments are used and applied equally to the

groups; interventions are spelled out clearly; all important outcomes are considered; and appropriate attention to confounders in analysis. In addition, for RCTs, intention to treat analysis is used.

Fair: Studies will be graded "fair" if any or all of the following problems occur, without the fatal flaws noted in the "poor" category below: Generally comparable groups are assembled initially but some question remains whether some (although not major) differences occurred with follow-up; measurement instruments are acceptable (although not the best) and generally applied equally; some but not all important outcomes are considered; and some but not all potential confounders are accounted for. Intention to treat analysis is done for RCTs.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

In addition to *Good* and *Fair*, the evidence-based practice center includes a category of *Poor* for RCTs and Cohort Studies

Poor: Studies will be graded "poor" if any of the following fatal flaws exists: Groups assembled initially are not close to being comparable or maintained throughout the study; unreliable or invalid measurement instruments are used or not applied at all equally among groups (including not masking outcome assessment); and key confounders are given little or no attention. For RCTs, intention to treat analysis is lacking.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1980-2013

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence?

The EPC identified 19 trials that examined the efficacy of primary care-relevant interventions in preventing tobacco use initiation, promoting cessation among youth, and/or harms of interventions. While the study designs varied, all were rated as Good or Fair quality by the EPC. Below is a general description, per topic studied, of the trials that were assessed.

- Combined prevention and cessation interventions
  - 7 trials: 5 randomized controlled trials (RCTs), 2 cluster-randomized trials
- Prevention interventions
  - 10 trials: 4 of the trials from the assessment for combined treatment were included in addition to 6 trials on behavior-based interventions to prevent tobacco initiation
  - Two of the studies were rated good quality based on their methods (e.g., valid randomization techniques, good intervention fidelity)
  - The remaining were rated as fair quality, as randomization procedures were not reported or uncertain
  - The EPC noted there was also a lack of blinding for outcome assessors but concluded this was unlikely to produce bias in the studies using standardized data collection tools (e.g., computer-assisted telephone interviewing)
- Cessation interventions (Behavior and Bupropion)
  - 10 trials: 9 trials examined cessation among baseline smokers. Of these, two were rated good quality based on their methods (e.g., valid randomization techniques and good intervention fidelity)
  - The remaining were rated fair quality due to issues including attrition and concerns with participant compliance

- Adverse effects associated with interventions
  - None of the trials of behavior-based interventions explicitly reported on treatment harms, but three medication-specific studies reported on harms
  - All three medication-specific trials included randomization techniques.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Fair to good quality: most of the trials included in the evidence review included randomization and had good intervention fidelity.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

The USPSTF found adequate evidence that behavioral counseling interventions can reduce the risk of smoking initiation in school-aged children and adolescents.

The U.S. Public Health Service reached a similar conclusion and also recommends clinicians provide adolescent tobacco users assistance with quitting, citing research that has shown that a provider's advice to quit can be effective.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The USPSTF found no evidence on the harms of behavioral interventions to prevent tobacco use and concluded the magnitude of potential harms is probably small to none.

Overall, the USPSTF concluded with moderate certainty that primary-care interventions to prevent tobacco use in school-aged children and adolescents have a moderate net benefit.

Taken together with the U.S. Public Health Service recommendations, clinician-provided interventions to prevent tobacco use in adolescents is associated with net benefits.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

We are not aware of any major new evidence reviews conducted since this systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Tobacco__Evidence.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The Tobacco Use and Help with Quitting Among Adolescents measure addresses an issue of significant importance. Tobacco use can have both immediate and long-term serious health consequences, yet data show that, despite some successes, many adolescents continue to begin or use tobacco products. Research has shown that health care providers can play an important role in promoting tobacco-use abstinence and cessation. Thus, this measure encourages standardized documentation of tobacco use status among adolescents and appropriate follow-up for those who are users.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
N/A, New Measure

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Guidelines recommend that clinicians provide tobacco use screening, education and counseling to adolescents during annual visits (Fiore 2008, USPSTF 2013). However, data from the 2011 National Youth Tobacco Survey showed that only a third of adolescent respondents reported being asked about tobacco use or advised about abstaining/quitting by a health care provider (32.2% and 31.4%, respectively) (Schauer 2014).

As part of the measure field testing (described more fully in the Testing attachment), we found that the overall rate of adolescents with documentation of tobacco use and help with quitting was 61.6%, with a range of 44.5 to 85.3% across three sites.

Fiore MC, Jaén CR, Baker TB, et al. Treating Tobacco Use and Dependence: 2008 Update. Clinical Practice Guideline. Rockville, MD: US Department of Health and Human Services. Public Health Service; 2008.
Institute for Clinical Systems Improvement (ICSI). Healthcare Guideline: Preventive Services for Children and Adolescents. Bloomington, MN: Institute for Clinical Systems Improvement; 2009.
Moyer VA, US Preventive Services Task Force. Primary care interventions to prevent tobacco use in children and adolescents: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2013;159(8):552–557.
Schauer GL, Agaku IT, King BA, Malarcher AM. Health care provider advice for adolescent tobacco use: results from the 2011 National Youth Tobacco Survey. Pediatrics. 2014 Sep;134(3):446-55. doi: 10.1542/peds.2014-0458.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We have some limited data on disparities from testing. Across our three clinician sites, the percentage of adolescents whose tobacco use was documented and, if smokers, who received help with quitting varied by race/ethnicity and insurance status. In our sample, non-Hispanic white and African American adolescents had similar rates of tobacco use/help documented (65.9 and 66.8%,

respectively). Hispanic/Latino and Asian/Native American/Pacific Islander adolescents had lower rates (38.1 and 25.0%, respectively). Those of other/multiple races had a rate of 54.6%.

By insurance status, those with Medicaid had a rate of 60.6%, commercial had a rate of 82.0%, and those who self-paid/had other insurance had a rate of 39.2%.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
From the 2005 National Health Interview Survey, white smokers (85%) were slightly more likely to be asked about tobacco use than black (77%) or Hispanic (72%) smokers (Cokkinides et al., 2008). In addition, minority groups were also less likely to be advised to quit (63% in whites, 55% in black and 48% in Hispanics) (Cokkinides et al., 2008).

Persons (across ages) whose household incomes were below or near the federal poverty level had substantially higher prevalence of smoking, compared with persons whose household incomes were above the federal poverty level. Yet people who have a low socio-economic status are less likely to have adequate access to primary care providers and information about the harms of tobacco use (Fiore et al., 2008).

Cokkinides, V. E., M. T. Halpern, et al. 2008. "Racial and Ethnic Disparities in Smoking-Cessation Interventions." American Journal of Preventive Medicine 34(5): 404-412.
Fiore MC, Jaén CR, Baker TB, et al. Treating Tobacco Use and Dependence: 2008 Update. Clinical Practice Guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service. May 2008.

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
A leading cause of morbidity/mortality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Over 2.6 million adolescents 18 years of age and younger are current tobacco users, with nearly one-fifth of all adolescents becoming current smokers before finishing high school (NSDUH 2010 and University of Michigan 2011). This issue is important, as early onset of tobacco use is correlated to tobacco use in adulthood. Of adults that smoke on a daily basis, 82 percent reported trying their first cigarette before the age of 18, and 53 percent reported becoming daily smokers before the age of 18 (Campaign for Tobacco-Free Kids [TFK] 2011b).

Tobacco use is associated with some of the most serious and costly diseases, including lung cancer, heart disease and emphysema. Tobacco use can affect an individual's reproductive health and damage almost every organ in the body. In addition to these long-term complications, there are also a number of health concerns that can appear immediately in otherwise young and healthy adolescents, such as increased heart rate, increased blood pressure and shortness of breath (TFK 2011a, 2012). Additionally, tobacco use can lead to engagement in other risky behaviors. Adolescents who smoke or use tobacco products are three times more likely than their non-smoking counterparts to use alcohol; eight times more likely to use marijuana; and 22 times more likely to use cocaine (TFK 2011a and Fox et al. 2010).

The financial burden incurred from tobacco use is significant. From 2000 to 2004, annual expenditures (public and private) related to smoking were $96 billion, and another $97 billion can be attributed to lost productivity each year (TFK 2012 and Clark et al., 2010). When taking into account additional costs related to engagement in other risky behaviors, the costs total over $200 billion (distributed among direct costs such as medical expenses and indirect costs such as costs related to lost productivity and drug-related crimes) (Clark et al., 2010).

Studies suggest tobacco use prevention efforts can lead to cost savings. The Campaign for Tobacco-Free Kids found that for

prevention or early intervention efforts, for every percentage-point decline in youth smoking, there is a corresponding $13.2 million reduction in health care costs (accrued over the lifetime of adolescents who do not become adult smokers) (TFK 2010).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Campaign for Tobacco-Free Kids (TFK). 2012. Tobacco Overview. http://www.tobaccofreekids.org/facts_issues/tobacco_101/ (April 2012).
Campaign for Tobacco-Free Kids. 2011a. Health Harms from Smoking and Other Tobacco Use. http://www.tobaccofreekids.org/research/factsheets/pdf/0194.pdf.
Campaign for Tobacco-Free Kids. 2011b. Tobacco Use Among Kids. http://www.tobaccofreekids.org/research/factsheets/pdf/0002.pdf.
Campaign for Tobacco-Free Kids. 2010. Benefits & Savings From Each One Percentage Point Decline in the USA Smoking Rates. http://www.tobaccofreekids.org/research/factsheets/pdf/0235.pdf.
Clark RE et al. 2010. Substance Abuse and Healthcare Costs Knowledge Asset, Web site created by the Robert Wood Johnson Foundation's Substance Abuse Policy Research Program. http://saprp.org/knowledgeassets/knowledge_detail.cfm?KAID=21. http://www.surgeongeneral.gov/tobacco/treating_tobacco_use08.pdf.
Fox HB, McManus MA and Arnold KN. 2010. Significant Multiple Risk Behaviors Among U.S. High School Students. http://www.thenationalalliance.org/pdfs/FS8.%20Significant%20Multiple%20Risk%20Behaviors.pdf.
National Survey on Drug Use and Health (NSDUH). 2010 National Survey on Drug Use and Health data. http://oas.samhsa.gov/NSDUH/2k10NSDUH/tabs/Sect2peTabs17to21.pdf.
University of Michigan, Monitoring the Future Study, 2011. http://www.monitoringthefuture.org/data/11data/pr11cig1.pdf.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health : Alcohol, Substance Use/Abuse, Behavioral Health : Tobacco Use, Prevention : Tobacco Use

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Prevention, Prevention : Screening

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
None

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
New Measure

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Adolescents who are not smokers OR Adolescents who are smokers but are receiving cessation counseling.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Look back period – 18 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Documentation that the adolescent is not a tobacco user
OR
Documentation that the adolescent is a tobacco user AND any of the following:
-Advice given to quit smoking or tobacco use
-Counseling on the benefits of quitting smoking or tobacco use (e.g., "5-A" Framework)
-Assistance with or referral to external smoking or tobacco cessation support programs (e.g., telephone counseling 'quit line')
-Current enrollment in smoking or tobacco use cessation program

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Adolescents who turn 12 through 20 years of age during the measurement year.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Adolescents who turn 12 through 20 years of age during the measurement year and had documentation of a face-to-face visit with a primary care practice during the 12 months prior to the measurement year.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
N/A

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Identify the eligible population (denominator).
Step 1a: Identify adolescents who turn 12 through 20 years of age during the measurement period.
Step 1b: Identify adolescents in Step 1a who had a face-to-face visit.
Step 2: Identify tobacco users (numerator).
Step 2a: From the denominator, identify adolescents documented as non-tobacco users.
Step 2b: From the remaining adolescents in the denominator, identify adolescents documented as tobacco users who received help with quitting.
Step 3: Sum adolescents identified in Steps 2a and 2b.
Step 4: Divide the total in Step 3 by the denominator to get the rate.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure has been newly added to the Physician Quality Reporting System, which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Tobacco_Use_-_Testing-635803585109241250.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Tobacco Use and Help with Quitting Among Adolescents
Date of Submission: 9/30/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset: N/A

1.3. What are the dates of the data used in testing? Testing of data element reliability and validity was performed using data from two study groups. In study group 1, data were obtained for care occurring from October 1, 2010 to December 31, 2011 (a 15-month observation period). For study group 2, data were obtained for care occurring between October 1, 2010 and March 2012 (an 18-month observation period).

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)?

We field tested this measure in three pediatric centers. The participating sites included pediatric clinics affiliated with a children's hospital (this sample was selected from adolescents enrolled in Medicaid); a network of clinics serving homeless and vulnerable adolescents, and an adolescent medicine clinic affiliated with a children's hospital (which primarily provides behavioral health and gynecology care to young women). The participating sites were in different states and used different EHR vendors.

In addition, we tested for face validity using advisory panels, which included experts in measures development, adolescent medicine, and quality improvement (i.e. individuals well positioned to speak to a measure's face validity). See Submission form for the names and affiliations of panel members.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
Potentially eligible adolescents were 12 to 19 years old as of December 31, 2010 (thus adolescents in the study ranged from age 12 to age 20) and had at least one visit to the same primary care office or adolescent medicine clinic in both 2010 and 2011.

A total of 597 adolescents comprised the final study group. Site personnel assigned site-specific identification numbers to protect the confidentiality of the adolescents' records and maintained a crosswalk with the patient identifiers. The mean age of the sample was 15.5 years (range: 12 to 19 years). Slightly more than two-thirds of the sample was female (68.2%). African-American adolescents represented the largest proportion of the overall sample (44.4%) followed by non-Hispanic whites (30%). Approximately 93% of adolescents lived in households where English was the preferred language spoken at home.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Reliability Testing

The sample for reliability testing was a sub-sample of 75 adolescents from the initial sample of 597 (25 from each of the three sites).

Validity Testing

For validity testing, we compared performance against well-care visits. Thus, we used the sample from two of our sites; site 3 was excluded because it was an adolescent medicine clinic that served primarily female adolescents for behavioral health and gynecology care. The resulting sample was 400 adolescents from the initial sample.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
Adolescent's health insurance coverage (commercial, Medicaid, self-pay/other) was used as a proxy measure of family socioeconomic status.

2a2. RELIABILITY TESTING
***<u>Note</u>**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests
Reliability was tested using two methods: inter-rater reliability and manual review/EHR extract comparison.

To assess inter-rater reliability, two reviewers independently collected data on 75 patients. Inter-rater reliability assessed the level of agreement for data elements related to tobacco cessation assistance between two independent abstractors reviewing the same data from the same data source. Agreement between abstractors was measured using the kappa statistic, which is a measure of agreement adjusted for agreement that could occur by chance. Kappa coefficients greater than 0.75 are indicative of excellent agreement.

In the manual review/EHR comparison, we assessed the agreement between rates calculated using manual EHR review compared to rates calculated automatically through an EHR extract. We report the kappa statistic here as well.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*).

Inter-Rater Reliability

Table 1 presents the levels of agreement between the two manual reviewers together for the data elements of tobacco cessation assistance.

Table 1. Inter-rater Reliability of Manual Reviews for *Tobacco Use and Help With Quitting Among Adolescents* Data Elements[1]

	TOTAL	
Data elements	**Kappa Coefficient**	**95% Confidence Interval[2]**
Smoking Status as defined in CMS EHR Meaningful Use objectives	0.94	0.87, 1.00
Current tobacco use	n/a	n/a
Documentation of advice to quit smoking/using tobacco	n/a	n/a
Counseling on the benefits of quitting smoking/using tobacco	n/a	n/a
Referral to smoking/tobacco cessation support program	n/a	n/a
Enrolled in a smoking/tobacco use program	n/a	n/a

[1] Based on n=75 repeated ratings by two manual reviewers
[2] 95% confidence intervals listed as n/a are because neither rater could find any data available in these charts for those data elements, though in these cases percent agreement can be considered 100%

Comparison between manual review and automated EHR extract
Table 2 compares information on tobacco use documentation and help with quitting calculated from manual EHR review versus automated EHR data extracts for the same sample of adolescents.

Table 2. Agreement between Manual EHR Review and Automated EHR Extract: Information on Tobacco (N=597)

	Manual EHR Review %	**Automated Data Extract %**	**Kappa Coefficient**	**95% Confidence Interval**

Percentage of Adolescents with Tobacco Status Documented	70.9%	53.9%	0.52	0.45, 0.58
Percentage of Adolescents Who Are Current Tobacco Users	13.6%	7.7%	0.66	0.56, 0.76
Percentage of Adolescents Whose Tobacco Use Is Documented and Who Received Help With Quitting If They Are Users	61.6%	47.4%	0.52	0.45, 0.59

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Inter-rater Reliability

The agreement between the two reviewers was high for a large proportion of data elements (approximately 200). While we are reporting on our testing for the tobacco use measure, the variables we assessed in our overall field test also included aspects of care related to demographics, sexual activity, chlamydia screening, depression screening, vaccinations, and other common well-care visit items.

There was high agreement for smoking status based on the *Meaningful Use* definition data element that was included in our measure (Kappa coefficient =0.94). The kappa coefficients for the remaining data elements could not be calculated because there was no variance in the ratings of either reviewer, primarily because the data elements were not documented.

Comparison between manual review and automated EHR extract

Overall, there was documentation of tobacco status for 70.9% of adolescents in the manual reviews compared to 53.9% in the automated extracts; the Kappa score (0.52) shows moderate agreement. In the manual reviews, 13.6% of adolescents were identified as smokers compared to 7.7% in the automated extract; the agreement is substantial (Kappa=0.66). For the proposed measure, the percentage of adolescents whose tobacco use is documented and who received help with quitting if they are users was 61.6% in the manual review versus 47.4% in the automated extract; the agreement was moderate (Kappa=0.52). There were substantial variations by site both in the results and the agreement between manual review and automated EHR extract.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Method of assessing face validity

Validity refers to whether the measure represents the concept being evaluated. To assess different perspectives on the measure's face validity, NCINQ reviewed the specifications and field test results with our advisory panels, which included individuals well positioned to speak to a measure's face validity. We convened a multi-stakeholder advisory panel with representation from a wide range of stakeholders, including consumers, pediatricians, family physicians, adolescent medicine physicians, health plans, state Medicaid agencies and researchers. In addition, we convened three targeted panels of stakeholders with particular relevance to the measures: we partnered with the National Partnership for Women and Families to convene a panel of consumer and family advocates; we partnered with the American Academy of Pediatrics to convene a panel of pediatricians, including adolescent medicine physicians; and we convened a panel of state Medicaid and CHIP representatives.

Method of assessing known groups validity

While any clinical encounter with adolescents, including sports physicals or acute care visits, represents an opportunity to discuss risky behaviors, designated well-care visits provide an important opportunity for these conversations. For this reason, NCINQ chose to evaluate the known-groups validity, defined as the ability of the measure to meaningfully differentiate distinct groups, by comparing the performance rates of adolescents who did not have any well-care visits in the measurement period to those who had one or more well-care visits. The manual reviewers abstracted the total number of well-care visits that were completed from October 1, 2010 to December 31, 2011. We defined well-care visits based on diagnosis or procedures codes or a visit that included documentation of health and developmental history, a physical exam, and health education/anticipatory guidance. The total number of well-care visits was transformed into a dichotomous variable to indicate whether the adolescent had any well-care visits (yes/no). NCINQ excluded Site 2 from the known groups validity analysis; this site is an adolescent medicine clinic that served primarily female adolescents for behavioral health and gynecology care.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face Validity

Our advisory panels concluded the measure is a valid way to assess tobacco status, use and follow-up among adolescents.

Known Groups Validity

As shown in Table 3, documentation of tobacco use and help with quitting was significantly higher among adolescents who had at least one well-care visit in the measurement period compared to adolescents without designated well care visits. The results were significant (p-value <0.0001).

Table 3. Known Groups Validation: Tobacco Use and Help with Quitting Among Adolescents with and Without Designated Well Care Visits[1]

	Had 1 or More Well-Care Visits in Measurement Period		
Percentage of adolescents whose tobacco use is documented and who received help with quitting if they are users	Yes	No	*p*-value
Sites 1 and 3 (combined)	58.9%	39.2%	<0.0001

[1] Data from EHR manual review (N=400)

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Face Validity

Our advisory panels concluded the measure as specified is a valid way to assess tobacco status, use and follow-up in adolescents. Our interpretation of these results is that the measure has sufficient face validity.

Known Groups Validity

Documentation of tobacco use and help with quitting was significantly higher among adolescents who had at least one well-care visit in the measurement period compared to adolescents without designated well care visits. Based on these results, we conclude the measure shows what we would expect from measure performance among these two groups. We expect patients with more well care visits to be compliant for the tobacco-use measure, and our results aligned with this expectation.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We did not conduct a test on exclusions, as this measure applies to a general population of adolescents and does not does not have relevant exclusions. Clinical guidelines recommend assessing and providing any needed treatment for all adolescents for tobacco use.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

NA

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

NA

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

This measure applies to a general population of adolescents and does not require risk adjustment.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical*

significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:
2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
Our sample of three sites did not provide sufficient data to conduct statistical tests. We provide information on the mean rates across the three sites below.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
Performance rates for the Tobacco Use and Help with Quitting Among Adolescents measure based on manual EHR review are presented by site and total sample in the table below. The overall rate was 61.6%. Rates vary from 44.5 percent documentation to 85.3 percent.

Performance Rates for Tobacco Use and Help with Quitting among Adolescents in Manual EHR Review, Overall and by Site

	Overall	Site 1	Site 2	Site 3
Percentage of Adolescents whose tobacco use is documented and who received help with quitting if they are users	61.6%	55.5%	85.3%	44.5%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e.*, *what do the results mean in terms of statistical and meaningful differences?*)
Rates varied from a low of 44.5 percent documentation to a high of 85.3 percent documentation. While site-to-site variation can be explained, in part, by differences in the availability of data elements, content of free-text notes, and site characteristics, we believe that variations in these results would imply variations among providers.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
NA

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e.*, *what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
This measure is collected with a complete sample, there is no missing data on the overall measure.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
NA

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
NA

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

Some components of this measure are aligned with the Meaningful Use definition of tobacco use status. These elements are available in structured fields. The remaining components are available in structured fields, narrative notes or other non-structured fields. We anticipate that, as measures evolve to capture data in real time, more elements will become available in structured formats. This measure has been specified as an eMeasure. Per NQF instructions, an eMeasure may be submitted separately in future.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

This measure was newly added to PQRS in 2015, so we do not have experience with operational use yet. However, testing revealed that this measure was feasible for clinicians to report.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, “commercial use” refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting CMS Physician Quality Reporting System (PQRS) http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ CMS EHR INCENTIVE PROGRAM (MEANINGFUL USE) https://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives Payment Program CMS Physician Quality Reporting System (PQRS) http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ https://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives CMS EHR INCENTIVE PROGRAM (MEANINGFUL USE)

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

CMS PHYSICIAN QUALITY REPORTING SYSTEM: This measure is used in the Physician Quality Reporting System (PQRS) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). PQRS is a voluntary individual reporting program that provides an incentive payment to identified EPs who satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Medicare Part C–Medicare Advantage beneficiaries are not included in claims-based reporting of individual measures or measures groups.

CMS EHR INCENTIVE PROGRAM (MEANINGFUL USE): The Medicare and Medicaid Electronic Health Care Record (EHR) Incentive Programs provide incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology. Tobacco status as a structured field is included in Meaningful Use.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for**

implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No unintended consequences were identified for this measure.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0028 : Preventive Care & Screening: Tobacco Use: Screening & Cessation Intervention

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
NQF 0028 measures tobacco use in adults aged 18 and older. The proposed measure will assess tobacco use in adolescents who are between the ages of 12 and 20.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3500-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
National Collaborative for Innovation in Quality Measurement (NCINQ) Consumer Panel
Joan Alker, MPhil, Georgetown Center for Children and Families
Roni Christopher, MEd, OTR/L, PCMH-CCE, The Greater Cincinnati Health Collaborative
Daniel Coury, MD, Nationwide Children's Hospital
Eileen Forlenza, Colorado Medical Home Initiative, Children and Youth with Special Health Care Needs Unit
Michaelle Gady, JD, Families USA
Janis Guerney, JD, Family Voices
Jocelyn Guyer, MPA, Georgetown Center for Children and Families
Catherine Hess, MSW, National Academy for State Health Policy
Carolyn Muller, RN, Montgomery County Health Department
Cindy Pellegrini, March of Dimes
Judith Shaw, EdD, MPH, RN, VCHIP
Stuart Spielman, JD, LLM, Autism Speaks
Michelle Sternthal, PhD, March of Dimes

NCINQ Measurement Advisory Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Katie Brookler, Colorado Department of Health Care Policy and Financing
Cathy Caldwell, MPH, Alabama Department of Public Health
Ted Ganiats, MD, University of California, San Diego
Darcy Gruttadaro, JD, National Allegiance on Mental Illness
Jennifer Havens, MD, NYU School of Medicine
Virginia Moyer, MD, MPH, FAAP, Baylor College of Medicine, USPSTF
Edward Schor, MD, Lucile Packard Foundation for Children's Health
Xavier Sevilla, MD, FAAP, Whole Child Pediatrics
Gwen Smith, Illinois Department of Healthcare and Family Services/Health Management Associates
Janet (Jessie) Sullivan, MD, Hudson Health Plan
Kalahn Taylor-Clark, PhD, MPH, George Mason University
Craig Thiele, MD, CareSource
Jeb Weisman, PhD, Children's Health Fund
Charles Wibbelsman, MD, Kaiser Permanente Medical Group, Inc.

NCINQ Clinician Advisory Panel
Elizabeth Alderman, MD, FAAP, Albert Einstein College of Medicine
Sarah Brewington, MD, Sandhills Pediatrics Inc
Gale Burstein, MD, MPH, FAAP, FSAHM, Women and Children's Hospital of Buffalo, NY
Barry Bzostek, MD, FAAP, Women and Children's Hospital of Buffalo, NY
Danielle Casher, MD, FAAP, St. Christoper's Hospital for Children
Edward Curry, MD, FAAP, Emergency Department, St. Christopher's Hospital for Children, PA
Eve Kimball, MD, FAAP, Southern California Permanente Medical Group
Paul Melinkovich, MD, FAAP, Kaiser Permanente
Jackie Nelson, MD, FAAP, Lander Regional H
Ellen Squire, MD, FAAP, HaysMed Pediatric Center

NCINQ State Panel
Mary Applegate, MD, Ohio Department of Job and Family Services
Sharon Carte, MHS, State of West Virginia Children's Health Insurance Program
Susan Castellano, Minnesota Department of Human Services
Catherine Hess, MSW, National Academy for State Health Policy
Michael Hogan, PhD, New York State office of Mental Health
Barbara Lantz, MN, RN, State of Washington Department of Social and Health Services, Medicaid Purchasing Administration
Judy Mohr Peterson, PhD, Oregon Health Authority
Tracy Plouck, MPA, Ohio Department of Mental Health
Gina Robinson, Colorado Department of Health Care Policy and Financing
Janet Stover, Illinois Association of Rehabilitation Facilities
Eric Trupin, PhD, University of Washington

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2015

Ad.3 **Month and Year of most recent revision:**

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2016

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2805
Measure Title: Pediatric Psychosis: Timely Inpatient Psychiatric Consultation
Measure Steward: Seattle Children´s Research Institute
Brief Description of Measure: Percentage of children/adolescents age >=5 to <=19 years-old admitted to the hospital with psychotic symptoms who had a psychiatric consult (in person or by telepsychiatry) within 24 hours of admission.
Developer Rationale: In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.

The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of indicators related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. Pediatrics 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is an indicator designed to fill this key measurement gap.

The COE4CCN Mental Health Working Group (see item Ad.1 for more details on this group) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group (Bardach et al. Pediatrics 2014). We found that psychosis was the third most common reason for pediatric mental health hospitalizations. Literature reviews were then conducted separately for each of the most common conditions, and one of these reviews focused on children evaluated and treated for psychosis in the ED and inpatient settings. See Evidence form for conceptual model underlying the rationale for the measures.

Based on this review, we developed a suggested list of indicators to assess the quality of pediatric mental health care in the hospital setting, including specific indicators measuring care for children with psychotic symptoms. The validity and feasibility of these indicators were then evaluated by an expert panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in hospitals in Washington state, Ohio, and Minnesota. This proposal presents the results of this development and validation work.

Numerator Statement: Eligible patients with documentation of an in-person or telemedicine psychiatric consult within 24 hours of inpatient admission.
Denominator Statement: Patients aged 5 to 19 years-old admitted to the hospital with psychotic symptoms.
Denominator Exclusions: No patients were excluded from the target population.

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The

preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. Evidence for this process measure should demonstrate that a psychiatric consult within 24 hours of admission in children/adolescents admitted to the hospital with psychotic symptoms should improve outcomes.

The developer provides the following information for this facility-level process measure:

- The developer states that the evidence supporting this measure derives primarily from American Academy of Child and Adolescent Psychiatry (AACAP) 2013 guidelines, a Cochrane review, and a review of the literature by the developer.
- The AACAP guideline recommendations cited by the developer are:
  - *Recommendation 4.* Antipsychotic medication is a primary treatment for schizophrenia spectrum disorders in children and adolescents. [Clinical Standard, highest recommendation]
  - *Recommendation 9.* Psychotherapeutic interventions should be provided in combination with medication therapies. [Clinical Guideline]
  - The developer posits the applicability of these guidelines as being appropriate to this measure because they are the recommended interventions and can only be implemented by a psychiatrist or qualified mental healthcare professional.
  - The systematic review encompassed 12 randomized controlled trials that aimed to improve outcome in first-episode psychosis, using a heterogeneous group of interventions, including early access (within 24 hours) to psychiatric evaluation, a family orientation to treatment, psychoeducational interventions, a variety of medications (including omega-3 fatty acids) and specialized treatment teams.
    - All treatment arms in all trials (both control and intervention), included standard psychiatric care.
    - None of the studies specifically assessed early access to care as a stand alone intervention.
    - Mean age among the studies is low 20s, with several in the teens.
  - The systematic review assesses the quality (moderate), quantity, and consistency of the evidence, but does not deploy a grading system for each study; it does categorize the studies by type.
  - The developer states the evidence supports psychiatric specialty consultation, given the complexity of medication choices and the need for careful monitoring of side effects and given the need for a trained psychiatrist or psychotherapist to deliver specific psychotherapeutic intervention.
  - The developer acknowledges: "Overall, though there is not extensive literature supporting this process measure, the benefits of measurement likely far outweigh the risks."
- Per the **NQF Algorithm for Evidence**, the evidence submitted includes a systematic review, but does not appear to include a grading of the evidence; the recommendations are graded based on the review (box 7-->). The eligible ratings are MODERATE or LOW, depending on the Steering Committee's assessment of whether the evidence submitted 1) is applicable to the process of care being measured, and 2) indicates a high degree of certainty that the benefits clearly outweigh undesirable effects.

Questions for the Committee

- *Is the evidence directly applicable to the process of care being measured?*
- *Has the developer provided sufficient evidence between the relationship of this measure to patient outcomes? For the timeframe (24 hours) that must be met to achieve the measure's specifications? For the age range?*
- *How strong is the evidence for this relationship?*
- *If the Committee concludes the empirical evidence is not sufficiently specific, does the Committee wish to*

consider the INSUFFICIENT WITH EXCEPTION path (boxes 10-->12)?

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following information:

- Mental health hospitalizations for pediatrics represented 9.1% of all hospitalizations for children ages >2 years in 2009, with psychosis the third most common mental health diagnosis (12.1%), after depression (44.1%) and bipolar disorder (18.1%).
- Children and adolescents with a diagnosis of a psychotic disorder face a number of challenges medically, socially, and developmentally. Several studies found a high risk of educational and/or occupational impairment for patients with early-onset schizophrenia.
- The measure captures two elements of performance: access to specialty psychiatric care for these patients, and timeliness of that access.
- The developer tested the measure using data aggregated over two years from three children's hospitals: Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital. Patients (N=253) included in the test were discharged from one of the three hospitals over the two-year period (January 1, 2012-December 31, 2013).
  - Mean performance was 88.4% and the range was 76.6% to 95.1%
  - With respect to disparities, the developer did not find statistically significant difference in performance across the demographic groups it examined.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, is the Committee aware of evidence that disparities exist in this area of healthcare?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities])*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- Evidence is from a different age cohort (young adults and teens) for whom psychosis related to schizophrenia is more common. Psychosis is not just schizophrenia and can be related to drugs (e.g. diphenhydramine toxicity), ingestions, metabolic disorders, etc. for which early psychiatric intervention with antipsychotics may not be appropriate. Psychosis in young children is rarely schizophrenia. Early psychiatric evaluation may be unnecessary in the case of medical causes of psychosis and may have the potential to be harmful if a psychiatric diagnosis is given when it is really a self-limiting medical condition or something different not related to psychiatry at all. The stigma attached to psychiatric disease is great and has the potential to be harmful especially in a case where it is not relevant. I don't think that the premise of treatment of psychosis as schizophrenia with antipsychotics early is necessarily appropriate until a medical cause is ruled out which might take longer than 24 hours. Also it is not appropriate in young children for whom any psychosis is unlikely to be related to schizophrenia. There is not really evidence that early psychiatric involvement (within the first 24 hours) in all cases of psychosis in this age group would change the outcome.
- The evidence is largely face validity; the argument that a consult within 24 hours constitutes "early access to care" doesn't really hold up since the papers about early access refer to early in the course of disease development, not the early hours of an acute episode. The available evidence is not directly applicable to this measure.
- The measure examines a key component of the process of care in a child with an acute psychosis. Early access to

psychiatric care is a key component of caring for children with psychosis.

- The recommendation with the stronger evidence (Recommendation 4) is based on pharmacological interventions, which presumably should be guided by a psychiatrist who supposedly has better understanding of the complexities of antipsychotic medications. However, since these typically require a few weeks before benefit is seen, does the clinical evidence support the need for this 24-hour specific time-frame?
- It doesn't appear that the evidence is directly related to access to a psychiatric consult within 24 hours of admission to the hospital. However, it seems that the benefits from this early evaluation would outweigh the risks.
- The evidence does not directly relate to this measure. There are practices that are dependent on a Psych consultation, but outcomes of the consult itself not been studied.

1b. Performance Gap.

- The performance in these children's hospitals is pretty similar, however, no data is available about other areas or about differences in outcome related to the intervention. Unlike some of the other treatment guidelines this is not clearly linked with improvements in care. Disparities are always an issue in mental health care, but this is an institution measure specific to a protocol and therefore shouldn't be applied differently to different populations.
- Performance is moderately good (all above 75%, mean almost 90%) so there may be a ceiling effect; however significant differences between sites were found. Thus the gap is relatively modest.
- Performance was measured at only three hospitals, and showed little variance in care.
- Would be interested to know how the performance gap that was quoted is related to the type of hospital: tertiary/quaternary care facility vs other. May represent issue of access. Subgroup analysis difficult since only two of the hospitals had access to more comprehensive information (PMCA data)
- Performance data was aggregated over 2 years from 3 children's hospitals. Developers provided information to prove that children with mental health disorders exhibit numerous challenges.
- There is no disparities information given therefore might not be considered disparities sensitive.
- The range of performance was relatively high, but may not be as high on non-children's hospitals. However, no data on non-Children's hospital provided. The gap may be greater than the data provided indicates.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

The developer provides the following information:

- This is a facility level measure.
- The data sources are administrative claims and electronic clinical data, including Electronic Health Record and paper medical records. The developer provides an attachment for the applicable codes.
- The developer defines the numerator as: Eligible patients with documentation of an in-person or telemedicine psychiatric consult within 24 hours of inpatient admission. The denominator is defined as: Patients aged 5 to 19 years-old admitted to the hospital with psychotic symptoms. There were no denominator exclusions (no patients were excluded from the target population).

Questions for the Committee :

- *Are the codes provided complete and appropriate?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer provides the following information:

- The measure was tested at the critical data element level and the performance measure score level. Two existing data sources were used, and the testing period was January 1, 2012 through December 31, 2013.
  - Critical data elements were tested for inter-rater reliability of medical record abstraction.
    - The total population sample size was N=252
    - For this specific measure the N=14—too few to calculate a Kappa. The developer reports, however, 100% agreement.
  - Performance measure score reliability was assessed using the intra-class correlation coefficient (ICC). The ICC assesses the ratio of between site variation and within site variation on performance. Higher ICC implies that the between site variation (signal) is higher than the within site variation (noise)
    - ICCs were computed using STATA SE 13.
    - The developer reports the hospital-level ICC=0.154 (95%CI 0.023-0.587); N=3
    - The developer reports that ICCs ≥0.10 indicate that there are meaningful between-site performance differences.
- Per the **NQF Algorithm for Reliability**, empirical testing was performed at the level of the computed performance measure score and so the eligible ratings are HIGH, MODERATE, or LOW (box3-->6)

Questions for the Committee

- *Does the Committee concur with the developer's conclusion that the results demonstrate sufficient reliability so that differences in performance can be identified?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The goal of the measure is to improve outcomes for pediatric patients admitted with psychotic symptoms by ensuring a consult occurs within 24 hours of inpatient admission.
- The numerator is the number of eligible patients with the documentation of a psychiatric consult within 24 hours of inpatient admission. The denominator is patients aged 5 to 19 years old admitted to the hospital with psychotic symptoms.
- The evidence for the specifications provided by the developer center on AACAP recommendations that 1) antipsychotic medication is a primary treatment for schizophrenia spectrum disorders in children and adolescents, and 2) psychoteherapeutic interventions should be provided in combination with medication therapies.
  - The developer states the evidence supports psychiatric specialty consultation, given the complexity of medication choices and the need for careful monitoring of side effects and give the need for a trained psychiatrist or psychotherapist to deliver specific psychotherapeutic intervention.
  - The developer notes that none of the studies in the systematic review specifically assessed early access to care as a stand alone intervention.
  - No evidence appears to address the specified 24-hour timeframe.

Question for the Committee

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer provides the following information:

- Empirical testing was done at the level of the computed measure score. The developer also tested face validity of the performance measure score.
- For the empirical testing at the measure score level, the developer assessed the relationship between performance on the measure and three utilization outcomes: 30-day readmission to the same hospital; 30-day return Emergency Department (ED) visit to the same hospital; and Length of Stay (LOS).
  - Multivariable regression was used to assess the relationship between performance on this measure and the validation metric.
  - The measure was adjusted for gender and insurance type (identified by face validity) and admitting hospital and patient race/ethnicity (associated with the outcome measure).
  - The developer found no statistical difference (OR=1) between passing/failing the measure and two of the validation metrics—readmissions and ED visits. The developer hypothesizes that the low event rate may have led to limited power to demonstrate an association.
  - The developer found LOS was statistically significantly shorter for those patients passing the measure (1.5 hours difference), therefore supporting the validity of the measure that patients with more timely psychiatric consultation have an improved outcomes (i.e., have their psychiatric needs more rapidly addressed and so are able to return home more rapidly).
- The developer performed systematic face validity assessment (RAND-UCLA Modified Delphi) of "whether panelists would consider providers who adhere more consistently to the quality measure to be providing higher quality care," which we interpret as face validity assessment at the level of the **computed measure score** (as required by NQF). The panelists did conclude there was face validity, although other factors were bundled with the assessment (i.e., the question was not scored in isolation).
- Per the **NQF Algorithm for Validity**, empirical testing was performed at the level of the computed performance measure score and so the eligible ratings are HIGH, MODERATE, or LOW (box 6-->8). Relying only on face validity, the eligible ratings are MODERATE OR LOW (box 4-->5).

Questions for the Committee

- *Was the empirical validity testing methodology appropriate and of the measure as specified?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- There are no exclusions

2b4. Risk adjustment:

- No risk adjustment or risk stratification

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provides the following information:

- The developer tested the difference in performance across the three hospitals using an omnibus test for difference, and then performed individual comparisons between each hospitals performance and the performance of the group as a whole.
- The developer used Fisher's exact test to assess statistical significance for all comparisons.
- Results were: statistically and clinically meaningful difference in hospital performance. The P-value for the omnibus test was 0.0002 and the P-value for the difference from overall mean of others was 0.19, 00001, and 0.0013.

Question for the Committee

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

The developer notes the following:

- While it is unlikely that missing data contributes to substantial or meaningful biases of performance estimates, two potential areas for missing data are at the level of the administrative claims and medical abstraction stage.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- Repeat of above questions. Is this meant to be reliability--testing? The numbers of this are so small (probably because schizophrenia and psychosis are relatively rare compared to other mental health conditions). Small numbers make the work difficult to interpret and of questionable utility for widespread use.
- The 24 hour deadline seems arbitrary, but is consistent with other timelines for establishing care.
- Allows for 24-48 hour consultation with justification for delay; no specific reason need be documented. Is this appropriate?
- The developers tested the measure for a 2 year period. The total population was 252, however the specific measure was used on 14 people therefore a kappa could not be calculated. The developers report 100% agreement. Current results show that the measure is reliable.
- Sufficient

2a2. Reliability testing

- Concerns about the potential for non-schizophrenia overlaps in diagnosis for which a child psych evaluation might not be helpful. Post hospital diagnoses might be different from acute admission ones and so pulling data after might miss kids that presented with psychiatric symptoms, but did not have a psychiatric disorder.
- Some concern that the denominator appears to be based on hospital discharge diagnosis of psychosis, but the issue is around actions at the time of admission for suspected psychosis. Some concern that the date/time of the consult note might not reflect when the action was actually taken, and that this could be "gamed" but overall, likely that this can be consistently implemented.
- Numbers of subjects seems too small to assess reliability.
- Does the "clinician extender" group include residents? Students?
- Numerator and denominator fields are clearly defined with ICD9 and ICD 10 codes defined.
- Agree with developer's reliability estimate.

2b1. Validity Specifications

- The evidence does not match the specifications and the specifications do not take into account the heterogeneity of etiology for presentation of acute psychosis in children.
- The evidence that is quoted is generally not really directly applicable to the measure; the face validity of timely access to appropriate care providers is good. This could be eligible for insufficient exception.
- The measure has been validated for both face validity and relationship to positive outcome.
- Early intervention is not a standalone intervention in any of the data quoted. Evidence in these studies does not include patients as young as 5 years old.
- The evidence does not directly address early access to a psych consult. Instead it provides evidence that children with psychiatric disorders are in need of complex medication therefore careful monitoring needs to happen...and this can happen if an early consult occurs.

2b2. Validity Testing

- Quality in treatment of psychiatric disorders assumes appropriate ruling out of medical causes which this measure doesn't appear to address. Looking at readmission rates to the ED or hospital may not be a measure relevant to all acute presentations of psychosis. An evaluation within 24 hours may not be the right timing and may not be necessary in the case of non-psychiatric conditions.
- Validity testing done with outcomes (30-day readmit and return to ER) that don't have an obvious theoretical connection to the measure. LOS does have an obvious connection and it was confirmed. The measure seems to relate most closely to appropriateness, access, and efficiency as parameters of care.

- The numbers are small, making the association with good outcome difficult to validate.
- Not certain the score is an indicator of quality--had minimal impact on any of the outcome measures.
- Empirical testing occurred at the level of computed measure score looking at the relationship between performance on the measure and three utilization outcomes: 30-day readmission to the same hospital; 30-day return Emergency Department (ED) visit to the same hospital; and Length of Stay (LOS).
- No statistical difference between passing/failing the measure and two of the validation metrics—readmissions and ED visits (small sample size?).
- LOS was statistically significantly shorter for patients passing the measure (1.5 hour difference - does that mean they stayed 1.5 hours shorter? if so, is that number clinically meaningful)

2b3-2b7. Threats to Validity

- Unclear
- Not likely
- Since the measure relies to some extent on chart abstraction, that is a threat.
- There are 2 potential areas for missing data: The level of the administrative claims and medical abstraction stage.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer provides the following information:

- The data sources are administrative claims and clinical data, including Electronic Health Record and paper medical records.
- In testing, the developer abstracted data from both paper charts and electronic health records. It found that EHR abstractions were easier due to the structured notes that automatically identified provider names, titles, and departments, facilitating efficient identification of the consultant note.

Questions for the Committee:

- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Data collection using clinical abstraction may not be as easy for institutions not set up for this study although a claim for a psych consultation and a matching ICD code might work. However, discharge codes might be different than the psychosis ones depending on the reason for the psychosis.
- Feasible
- Feasible in the context of structured EHR notes.
- The consultation is currently by manual chart abstraction but there are ways to make this electronic.
- The data sources are administrative claims and clinical data, including Electronic Health Record and paper medical records.
    - Data abstracted from both paper charts and electronic health records.
    - EHR abstractions were easier
    - Not sure how many hospitals are still using paper documentation. Paper documentation seems to be more difficult.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

The developer provides the following information:

- The planned uses of this measure include quality improvement with benchmarking (external benchmarking to multiple organizations) and quality improvement (internal to the specific organization).
- This measure has not been implemented as the development, validation, and testing were just recently completed. The tools needed to abstract the measures are publicly available and non-proprietary
- There were no unintended consequences identified during testing.

***Questions for the Committee**:*

- *The developer indicates use for benchmarking and quality improvement. NQF endorsement focuses on primarily accountability, and then appropriateness for quality improvement. Is this measure appropriate for accountability purposes?*
- *Can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- This measure has neither shown an improvement in quality nor an increase in efficiency based on the intervention. There might be decreased efficiency in requiring an early child psychiatry consultation in every case of psychosis especially if a medical reason is identified for which there would be no role of either psychiatric medication or psychotherapeutic intervention.
- The measure has not yet been used across systems.
- Benefits of the measure outweigh the unintended consequences. Appears to be a good quality improvement/benchmarking measure.
- The unintended consequence of increasing consultations that are not necessary is outweighed by the benefit of timely consultations when necessary.

Criterion 5: Related and Competing Measures

- There are no related and/or competing measures.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Pediatric Psychosis: Timely Inpatient Psychiatric Consultation

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

- ☐ Health outcome: Click here to name the health outcome
- ☐Patient-reported outcome (PRO): Click here to name the PRO

  PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

- ☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Psychiatric consultation within 24 hours of admission for psychosis

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

NA

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

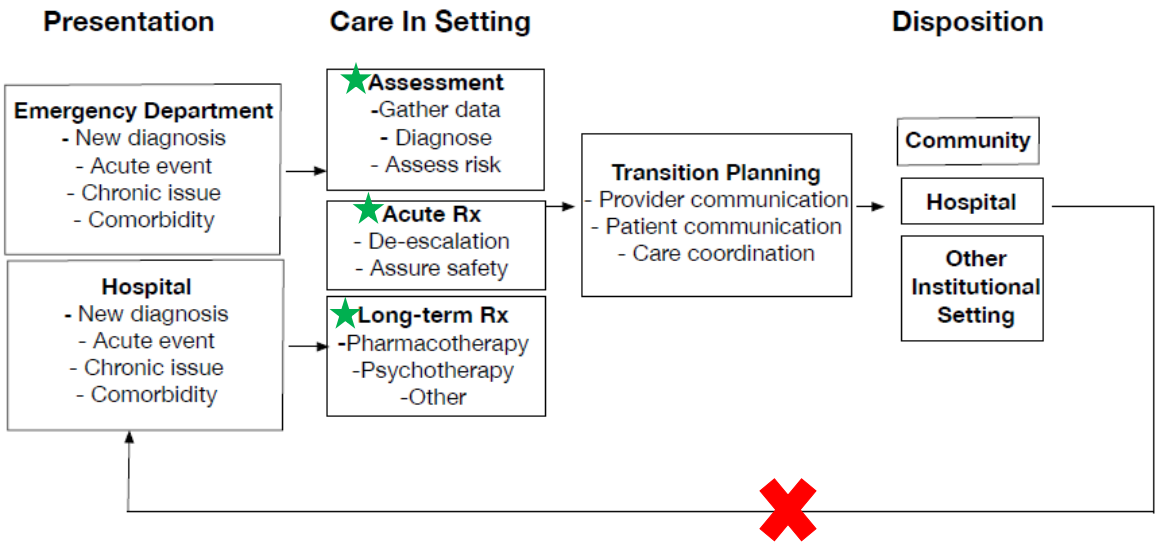
NA

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The evidence supporting this measure derives from three main sources: the American Academy of Child and Adolescent Psychiatry (AACAP) guidelines on treatment of early onset psychosis, a Cochrane systematic review, and the results of a Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CNN) multi-stakeholder Delphi panel convened specifically for the assessment of pediatric mental health measures.



The figure above depicts the underlying conceptual model describing how measured processes of care might reduce re-presentation with acute psychosis. The green stars mark the processes of care of interest, which this measure proposes be done with specialty psychiatric consultation, and that they be done in a timely manner. The red X marks the pathway back to the undesirable health outcome that would be blocked if measure performance is optimal.

The proposed measure captures two elements of performance: access to specialty psychiatric care for these patients, and timeliness of that access. The evidence for access to psychiatric management is summarized in the AACAP guidelines (items ***1a.4*, *and* *1a.7)***, and the evidence for early access is summarized in the Cochrane review (item #***1a.6* *and* *1a.7)***.

Summary: Overall, though there is not extensive literature supporting this process measure, the benefits of measurement likely far outweigh the risks. This measure also had high face validity according to the Delphi panel convened as part of the COE measure development work for pediatric mental health measures (see section 1a.8).

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

McClellan J, Werry J, Bernet W, Arnold V, Beitchman J, Benson RS, Bukstein O, Kinlan J, Rue D, Shaw J, Kroeger K: Practice parameter for the assessment and treatment of children and adolescents with schizophrenia, Journal of the American Academy of Child and Adolescent Psychiatry 2013, Volume 52, Issue 9, Pages 976–990

http://www.jaacap.com/article/S0890-8567(13)00112-3/fulltext

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Two recommendations from the guideline support the need for timely psychiatric consultation for children presenting to the inpatient general medical setting with psychotic symptoms. The interventions recommended below would only be appropriately instituted by a psychiatrist or other qualified mental health care provider:

Recommendation 4. Antipsychotic medication is a primary treatment for schizophrenia spectrum disorders in children and adolescents. [CS]

Recommendation 9. Psychotherapeutic interventions should be provided in combination with medication therapies. [CG]

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

The AACAP guidelines granted the two recommendations the highest [CS] (*Recommendation 4)* and second-highest [CG] (*Recommendation 9)* gradings:

•Clinical Standard [CS] is applied to recommendations that are based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials) and/or overwhelming clinical consensus

•Clinical Guideline [CG] is applied to recommendations that are based on strong empirical evidence (e.g., nonrandomized controlled trials, cohort studies, case-control studies) and/or strong clinical consensus

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

The subsequent, lower levels of ordered grades is below:

•Clinical Option [OP] is applied to recommendations that are based on emerging empirical evidence (e.g., uncontrolled trials or case series/reports) or clinical opinion, but lack strong empirical evidence and/or strong clinical consensus

•Not Endorsed [NE] is applied to practices that are known to be ineffective or contraindicated

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

SEE SECTION 1A. 7 PART 1 FOR SUMMARY OF AACAP GUIDELINE EVIDENCE

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Early intervention for psychosis.[1]

Marshall M, Rathbone J.

Cochrane Database Syst Rev. 2011 Jun 15;(6):CD004718. doi: 10.1002/14651858.CD004718.pub3. Review.

http://www.ncbi.nlm.nih.gov/pubmed/21678345

http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004718.pub3/epdf

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7 **SEE SECTION 1A. 7 PART 2 FOR SUMMARY OF COCHRANE FINDINGS**

SECTION 1A.7 PART 1: RESPONSES FOR USE OF PSYCHIATRIC SPECIALTY CONSULTATION

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

SUMMARY: Overall, though the evidence for use of antipsychotics and psychotherapeutic interventions that we present below is limited, the evidence supports the need for psychiatric specialty consultation, given the complexity of therapeutic choices, and the need for specialty training in delivering psychotherapeutic interventions, as recommended by AACAP guidelines.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Recommendation 4. Antipsychotic medication is a primary treatment for schizophrenia spectrum disorders in children and adolescents. [CS]

Recommendation 9. Psychotherapeutic interventions should be provided in combination with medication therapies. [CG]

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Summary of strength of the quoted evidence is as follows:

For **Recommendation 4**, the quality of the evidence ranged from case series (weakest evidence) to randomized controlled trial (strongest evidence).

For **Recommendation 9**, the quality of the evidence ranged from uncontrolled trial (3rd level of 4, with 1 being strongest evidence) to randomized controlled trial (strongest evidence).

The categories of empirical evidence are defined in the guideline as follows, presented in descending order:

- Randomized, Controlled Trial [rct] is applied to studies in which subjects are randomly assigned to two or more treatment conditions
- Controlled Trial [ct] is applied to studies in which subjects are nonrandomly assigned to two or more treatment conditions
- Uncontrolled Trial [ut] is applied to studies in which subjects are assigned to one treatment condition
- Case series/report [cs] is applied to a case series or a case report

Overall grades for the quality of the evidence were not given, beyond the grading for the Recommendations noted above in Section 1a.4.3.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

All grades are presented above.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: January 2004-August 2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

For Recommendation 4, there were 13 randomized controlled trials and 2 case series.

For Recommendation 9, there were 2 randomized controlled trials and 1 controlled trial.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Recommendation 4. The overall quality of evidence across the studies is strong, with a large number of well designed randomized controlled trials. Study outcomes generally focused on treatment of symptoms, with no assessment of prevention of relapse, representation to the Emergency department or readmission to the hospital.

Recommendation 9. The overall quality of the evidence across the studies is weak-moderate, with two small RCTs (n=25 and n=40) with heterogeneous interventions and one small non-randomized controlled trial (n=24). The non-randomized trial found a decrease in hospitalization rates with the intervention. The other studies assessed effects on cognitive skills (executive function and cognitive flexibility).

A notable strength of the evidence for both recommendations is that much of it derives from studies with patient populations in the age range of the proposed measure, focusing on youth with early onset schizophrenia (EOS), who range in age from early teens to mid-late 20s. 11 studies for **Recommendation 4** and 4 studies from **Recommendation 9** focus on this population.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Recommendation 4. The trials cited in the guideline mostly find that second-generation (atypical) antipsychotics are weakly effective or not effective, and that tolerability is low. The estimates of benefit are not high.

Briefly, several controlled trials of antipsychotic agents for EOS have been conducted, although all have limitations and more studies are needed. Older studies support the use of loxapine94[rct] and haloperidol. 95[rct] For adolescents with EOS, industry-sponsored randomized controlled acute trials support the efficacy of risperidone (n = 257)96[rct] and aripiprazole (n = 302).97[rct] An industry-sponsored trial found olanzapine to

be superior to placebo on symptom ratings of psychosis (n = 107).However, the overall response rate for olanzapine was low (38%) and did not differ from placebo.98[rct]

There are few studies comparing the efficacy and safety of different agents for EOS. In youth with more broadly defined psychotic disorders (n = 50), olanzapine was maintained significantly longer than risperidone and haloperidol.99[rct] The proportion of responders at 8 weeks for olanzapine (88%), risperidone (74%), and haloperidol (53%) was not significantly different. Sedation, extrapyramidal side effects (EPSs), and weight gain were common in all three groups. A small randomized controlled 6-week trial of adolescents with first-onset psychosis (n = 22) found no significant differences in efficacy or tolerability between risperidone and quetiapine.100[rct] Similarly, an 8-week study of youth with different psychotic illnesses (n = 30) found no differences in efficacy among olanzapine, risperidone, and quetiapine.101[rct]

Recommendation 9. The estimates of benefit are weak, though the studies all found a benefit with the intervention arm.

Briefly, there are few studies of psychosocial treatments for youth with schizophrenia. Psychoeducation, including parent seminars, problem-solving sessions, milieu therapy (while the subjects were hospitalized), and networks (reintegrating the subjects back into their schools and communities), was associated with lower rates of rehospitalization in a small sample of adolescents with EOS.125[ut] In a separate study, youth who received cognitive remediation plus psychoeducational treatment showed greater improvements in early visual information processing at 1-year follow-up, although no significant short-term improvements were found.126[rct],127[rct] A 3-month trial of cognitive remediation therapy, in comparison with standard therapy, was associated with improvements in planning ability and cognitive flexibility in adolescents with schizophrenia.128[rct]

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Recommendation 4. Tolerability (e.g., sleepiness), changes to metabolic profiles (weight gain, glucose tolerance), and neurologic side effects (extrapyramidal side effects) were studied. One atypical antipsychotic, olanzapine, while better tolerated than other atypicals, is associated with greater weight gain, leading to a suggestion in the guideline to defer it as a first line agent.

Recommendation 9. No harms were assessed, though costs of programs were discussed.

SUMMARY: Overall, the evidence presented supports psychiatric specialty consultation, given the complexity of medication choices and need for careful monitoring of side effects, and given the need for a trained psychiatrist or psychotherapist to deliver specific psychotherapeutic interventions.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

Intervention for adolescents with early-onset psychosis and their families: a randomized controlled trial. 9-month, randomized, rater-blinded clinical trial involving 55 adolescent patients with early-onset psychosis and either or both of their parents. A psychoeducational problem-solving group intervention (n = 27) was compared with a nonstructured group intervention (n = 28). At the end of the group intervention, 15% of patients in the

psychoeducational group and 39% patients in the nonstructured group had visited the emergency department (p = .039).[3]

http://www.ncbi.nlm.nih.gov/pubmed/24839887

2. Calvo A, Moreno M, Ruiz-Sancho A, et al. Intervention for adolescents with early-onset psychosis and their families: a randomized controlled trial. *J Am Acad Child Adolesc Psychiatry*. 2014;53(6):688-696.

Summary: This study does not change the conclusions of the systematic review, though it strengthens the evidence that psychotherapeutic interventions (in this case, a group intervention) can decrease utilization and re-presentation for care.

SECTION 1A.7—PART 2: EARLY INTERVENTION FOR EARLY ONSET PSYCHOSIS (COCHRANE REVIEW)

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Overview: This section summarizes data in support of early intervention for psychosis, from a Cochrane review in 2011. As noted in section **1a.6.1.**, the citation and URL for the review is:

Early intervention for psychosis.[1]

Marshall M, Rathbone J.

Cochrane Database Syst Rev. 2011 Jun 15;(6):CD004718. doi: 10.1002/14651858.CD004718.pub3. Review.

http://www.ncbi.nlm.nih.gov/pubmed/21678345

http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004718.pub3/epdf

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The specific overall intervention that is addressed in the evidence review is the effect of early intervention programs in the prevention and treatment of psychosis. Early intervention in psychosis has two elements that are distinct from standard care: early detection and phase-specific treatment.

We focus on the findings from the review regarding phase specific treatment. Phase-specific treatments are defined as treatments (psychological, social or physical) that are especially targeted at people in the prodrome or early stages of schizophrenia.[4] Phase-specific treatments may be directed at preventing progression to psychosis (in people with prodromal symptoms), or at promoting recovery (in people who have recently experienced their first episode of psychosis).

We are focusing on the evidence base for phase specific treatment.

Until relatively recently, the orthodox approach to treating schizophrenia was to concentrate therapeutic resources on those people who developed severe and chronic disabilities.[5] This approach has been challenged by proponents of early

intervention, who have argued that greater investment of resources in the early stages of the disorder, such as during young adulthood and adolescence, might substantially reduce the numbers of people developing chronic disabilities.[6] This argument has been strengthened by the observation that there may be an association between various outcome parameters and the duration of untreated psychosis (the time from the development of the first psychotic symptom to the receipt of adequate drug treatment).[7] This has led to the proposition that untreated psychosis may be 'toxic' and that early intervention might prevent irreversible harm.[8] *The proposed indicator reflects the emphasis in the literature on the benefits of early intervention with appropriate psychiatric expertise.*

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

GRADE Working Group grades of evidence

High quality: Further research is very unlikely to change our confidence in the estimate of effect.

Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.

Very low quality: We are very uncertain about the estimate.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1994-2009

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

12 randomized controlled trials.

The 12 trials aimed to improve outcome in first-episode psychosis, using a heterogeneous group of interventions, including early access (within 24 hours) to psychiatric evaluation, a family orientation to treatment, psychoeducational interventions, a variety of medications (including omega-3 fatty acids) and specialized treatment teams. *None of the studies specifically assessed early access to care as a stand alone intervention.*

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The quality of the evidence is moderate. All studies were randomized, although in terms of allocation concealment, the quality of included studies was acceptable but not good, since precise details of the method of randomization were lacking for most studies. Because the studies were small, except for one, it is likely that larger trials could have an important impact on confidence in the estimate of effect and may change the estimate.

A relevant strength of the evidence is that the measure's target population is similar in age to populations in the included studies. Mean ages in the reviewed studies were in the low 20s, with multiple studies including age ranges down to the low teens.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

There is some support for phase-specific treatment focused on employment and family therapy, but this should be replicated with larger and longer trials. The effect of treatment teams specialized in early intervention for psychosis is equivocal as was treatment with omega-3 fatty acids. All treatment arms in all trials (both control and intervention), included standard psychiatric care. There were no trials that did not include psychiatric assessment and care.

A meta-analysis was not performed due to heterogeneity across studies in interventions and outcomes.

Summary: The proposed indicator addresses only one element of early intervention – timely access to psychiatric evaluation. This element was only tested in combination with other interventions in the reviewed trials. However, the strength of the specialty care evidence in early intervention supports a likely strong benefit to timely psychiatric care over generalist care (hospitalist or primary care provider) in the inpatient setting.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Side effects of specific medications were studied, but harms of early intervention were not asssessed outside of the effects of specific medications.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

A recent review of early intervention in psychosis, published in August 2015 (citation below),[9] did not cite any new evidence, and a Pubmed search for "early intervention psychosis" did not return any new clinical trials assessing the effects of early intervention programs for treatment of psychosis.

Early intervention services in psychosis: from evidence to wide implementation.

Csillag C, Nordentoft M, Mizuno M, Jones PB, Killackey E, Taylor M, Chen E, Kane J, McDaid D.

Early Interv Psychiatry. 2015 Sep 11. doi: 10.1111/eip.12279. [Epub ahead of print] PMID: 26362703

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

References

1. Marshall M, Rathbone J. Early intervention for psychosis. *Cochrane Database Syst Rev.* 2011(6):CD004718.
2. Haas M, Eerdekens M, Kushner S, et al. Efficacy, safety and tolerability of two dosing regimens in adolescent schizophrenia: double-blind study. *Br J Psychiatry.* 2009;194(2):158-164.
3. Calvo A, Moreno M, Ruiz-Sancho A, et al. Intervention for adolescents with early-onset psychosis and their families: a randomized controlled trial. *J Am Acad Child Adolesc Psychiatry.* 2014;53(6):688-696.
4. Miller R, Mason SE. Phase-specific psychosocial interventions for first-episode schizophrenia. *Bull Menninger Clin.* 1999;63(4):499-519.
5. McGorry P, Jackson H. *Pathways to care in early psychosis: clinical and consumer perspectives.* Cambridge: Cambridge University Press; 1999.
6. Wyatt RJ. Early intervention with neuroleptics may decrease the long-term morbidity of schizophrenia. *Schizophr Res.* 1991;5(3):201-202.
7. Norman RM, Malla AK. Duration of untreated psychosis: a critical examination of the concept and its importance. *Psychol Med.* 2001;31(3):381-400.
8. Sheitman BB, Lieberman JA. The natural history and pathophysiology of treatment resistant schizophrenia. *J Psychiatr Res.* 1998;32(3-4):143-150.
9. Csillag C, Nordentoft M, Mizuno M, et al. Early intervention services in psychosis: from evidence to wide implementation. *Early Interv Psychiatry.* 2015.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
P3_Inpt_Consult_in_24_hours_evidence_attachment_2015_09_29_SUBMITTED.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.

The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of indicators related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. Pediatrics 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is an indicator designed to fill this key measurement gap.

The COE4CCN Mental Health Working Group (see item Ad.1 for more details on this group) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group (Bardach et al. Pediatrics 2014). We found that psychosis was the third most common reason for pediatric mental health hospitalizations. Literature reviews were then conducted separately for each of the most common conditions, and one of these reviews focused on children evaluated and treated for psychosis in the ED and inpatient settings. See Evidence form for conceptual model underlying the rationale for the measures.

Based on this review, we developed a suggested list of indicators to assess the quality of pediatric mental health care in the hospital setting, including specific indicators measuring care for children with psychotic symptoms. The validity and feasibility of these indicators were then evaluated by an expert panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in hospitals in Washington state, Ohio, and Minnesota. This proposal presents the results of this development and validation work.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In a field test of the measure, performed as part of the funded development work, we measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital. Included patients were discharged from one of the three hospitals over the two year period (January 1, 2012-December 31, 2013).

# of hospitals: 3
# of patients: 253
Mean (SD): 88.4% (10.2)
Min-Max: 76.6%-95.1%

IQR: N/A

See Testing form, item 2b.5.2a, for individual hospital performance.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We performed a field test of the measure as part of the funded development work. We measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital. Patients included in the field test were discharged from one of the three hospitals over the two year period (January 1, 2012-December 31, 2013).

We did not find any statistically significant differences in performance across groups. Please see Testing form, item 2b.5.2b for data.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
Psychosis in pediatric patients is a high priority aspect of healthcare, with substantial inpatient utilization and high severity of illness, in addition to a number of associated costs to the healthcare system and to patients and families. Mental health hospitalizations for pediatrics represented 9.1% of all hospitalizations for children ages >2 in 2009, with psychosis the third most common mental health diagnosis (12.1%), after depression (44.1%) and bipolar disorder (18.1%).1 A significant increase in the diagnosis of psychotic disorders from 8.3 to 12.0 percent of hospital discharges was found in a national survey of inpatient mental health services for children and adolescents from 1999 to 2000.2 Specific predictors of poor long term outcomes include more than two inpatient-treated episodes of schizophrenia3 and a longer duration of first inpatient treatment.3 Lay et al.3 found that 12 years after their initial diagnoses of schizophrenia only 17% of adolescents had not been readmitted for further inpatient treatment, and there was a median of 4 subsequent inpatient-treated episodes. Similarly, Fleischhaker et al.4 found an average of 3 readmissions for 40% of patients in a 10-year follow-up for adolescent-onset schizophrenia.

Children and adolescents with a diagnosis of a psychotic disorder face a number of challenges medically, socially, and developmentally. Several studies found a high risk of educational and/or occupational impairment for patients with early-onset schizophrenia.3,4 The long-term prognosis for psychosis with onset before the age of 18 years is poor in the majority of cases. For childhood-onset schizophrenia up to 50% of cases become chronic, 25% achieve partial remission, and only 25% achieve full remission.3-6 In addition, long-term studies of patients with childhood-onset schizophrenia found high rates of depression4,5 and significantly higher rates of suicide4,5 compared with other psychiatric inpatients. Compared to an estimated 15% of young adults in their community in Germany3, 42% of young adults with adolescent-onset schizophrenia were living with their parents, and another 32% were institutionalized. Lay et al.3 assessed delays or impairment in educational and/or occupational functioning and found

significant impairment for 58% of participants, mild impairment for 24%, and only 18% with no impairment. Social disability was assessed using items related to performance of specific social roles on the Psychiatric Disability Assessment Schedule; serious dysfunction was observed in 79% of patients and only 12% exhibited no dysfunction.

A number of costs have been associated with early-onset psychosis for the medical system as well as the patient and family. Length of stay for inpatients with psychosis has been found to typically be longer than for other mental health diagnoses.7 In addition, in a comparison of mental health versus non-mental health ED visits from 2001-2008, patients with a mental health diagnosis had fewer referrals to outpatient care7 and a higher number of inpatient admissions.7 Long-term studies of patients with early-onset psychosis have found that as adults, most were financially dependent on family or receiving public assistance.3,4

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Bardach NS, Coker TR, Zima BT, et al. Common and costly hospitalizations for pediatric mental health disorders. Pediatrics. 2014;133(4):602-609.
2. Case BG, Olfson M, Marcus SC, Siegel C. Trends in the inpatient mental health treatment of children and adolescents in US community hospitals between 1990 and 2000. Arch Gen Psychiatry. 2007;64(1):89-96.
3. Lay B, Blanz B, Hartmann M, Schmidt MH. The psychosocial outcome of adolescent-onset schizophrenia: a 12-year followup. Schizophr Bull. 2000;26(4):801-816.
4. Fleischhaker C, Schulz E, Tepper K, Martin M, Hennighausen K, Remschmidt H. Long-Term Course of Adolescent Schizophrenia. Schizophrenia Bulletin. 2005;31(3):769-780.
5. Remschmidt H, Martin M, Fleischhaker C, et al. Forty-two-years later: the outcome of childhood-onset schizophrenia. J Neural Transm. 2007;114(4):505-512.
6. Hassan GAM, Taha GRA. Long term functioning in early onset psychosis: Two years prospective follow-up study. Behavioral and Brain Functions. 2011;7.
7. Case SD, Case BG, Olfson M, Linakis JG, Laska EM. Length of stay of pediatric mental health emergency department visits in the United States. J Am Acad Child Adolesc Psychiatry. 2011;50(11):1110-1119.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

Behavioral Health, Behavioral Health : Serious Mental Illness, Mental Health, Mental Health : Serious Mental Illness

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Access

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

Measure specifications can be found at the following URL under the heading: "Mental Health Measures":
http://www.seattlechildrens.org/research/child-health-behavior-and-development/mangione-smith-lab/measurement-tools/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of

the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PSYCHOSIS_ICD9_and_ICD10_Codes_for_Denominator_Identification_SUBMITTED.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Eligible patients with documentation of an in-person or telemedicine psychiatric consult within 24 hours of inpatient admission.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
24 month period of data, retrospectively collected. We propose using 24 months due to the low prevalence of the condition. This is the period used in the field testing of the measure.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients passing the quality measure are identified during medical record abstraction using the guidelines below. The item numbers match the "Medical Records Abstraction Tool Guidelines" under "Mental Health Measures" provided on the website in S.1. This language is also in the "Medical Records Electronic Abstraction and Scoring Tool" on the website.

12) Psychiatric Consult –The patient had a psychiatric consult within 24 hours of admission (or prior to discharge if the admission was less than 24 hours in duration) [choose response 1]. The end of the 24-hour time frame is computed (based on admission time) and displayed in the online tool. Include in this interval any psychiatric consult that may have been done in the marker emergency department (ED) prior to admission if the patient was admitted via the marker ED. The consult may be in person or by telemedicine. The consult must have been done by a psychiatrist or PhD psychologist. If the consult was done by a clinician-extender (nurse practitioner, advanced practice nurse, physician assistant, licensed social worker, or licensed counselor), this is acceptable as long as the assessment is co-signed by a psychiatrist. If an appropriate person did not assess the patient during the first 24 hours, choose response 2 (No/No data), and continue to Q12a).

12a) Response 1 -The patient had a psychiatric consult within 48 hours of admission AND a justification for the delay. (The end of this time frame is computed based on the date and time of admission and is displayed in the question text.) If a qualifying MH provider assessed the patient by the indicated time, select response 1 only if a justification was noted for the delay that prevented the consult from occurring within the first 24 hours. The abstractor is not asked to evaluate the content or acceptability of the justification. Any justification that specifically refers to the time delay for the assessment is acceptable. If there was a consult within >24 to 48 hours and there is no justification noted for the delay, select response 2 (No/No data). If the consult did not occur or only occurred more than 48 hours after admission, select response 3 (Neither of the above/No data).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients aged 5 to 19 years-old admitted to the hospital with psychotic symptoms.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

Cases are identified from hospital administrative data.

Patients aged >=5 to <=19 years-old

Patients have at least one of the following ICD9 codes for psychosis, as a primary or secondary diagnosis: 291.3, 291.5, 292.11, 292.12, 293.81, 293.82, 295.30, 295.31, 295.32, 295.33, 295.34, 295.40, 295.41, 295.42, 294.43, 295.44, 295.70, 295.71, 295.72, 295.73, 295.74, 295.90, 295.91, 295.92, 295.93, 295.94, 296.24, 296.44, 297.1, 297.2, 297.3, 298.0, 298.1, 298.2, 298.3, 298.4, 298.8, 298.9

These codes were chosen by Members of the COE4CCN Mental Health Working Group (see Ad.1) co-chaired by Psychiatric Health Services Researchers Drs. Michael Murphy and Bonnie Zima.
Patients were included regardless of source of admission (from ED, direct admission, or transferred from outside hospital)

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
No patients were excluded from the target population.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

- S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)

No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Identify the hospital's eligible target denominator population (N)
2. Identify the cases meeting the target process, the numerator population (n)

3. Calculate the hospital score (n/N)

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A. Given the low prevalence of the condition, the measured group is the entire population of eligible patients.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data collection tool is publicly available on the website in S.1. and also attached in the Appendix materials.

Title: "Medical Record Measure Electronic Abstraction and Scoring Tool" under "Mental Health Measures"

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health/Psychiatric : Inpatient, Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
P3_Inpt_Consult_in_24_hours_Testing_Attachment_2015_10_13_SUBMITTED.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Pediatric Psychosis: Timely Inpatient Psychiatric Consultation
Date of Submission: 9/30/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient

preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

As described in the submission form, the validity and feasibility of the COE4CCN pediatric mental health quality measures were evaluated by an expert panel using the RAND-University of California, Los Angeles (UCLA) modified Delphi method.[1]

Detailed measure specifications were developed for the endorsed pediatric mental health measure. These specifications were then used to develop an electronic excel macro data collection tool for use with medical records data. The tool has automated scoring capability and is available on the website listed in item S.1. Abstraction and scoring guidelines are provided as an appendix to this submission.

Field Testing of the Delphi Panel Endorsed Pediatric Mental Health Quality Measures

Three tertiary care children's hospitals participated in the field test of the ***Pediatric Psychosis*** Mental Health quality measures. For each hospital, two research nurses were trained to use the medical record abstraction tool and the companion abstraction tool guidelines. For training purposes, the nurses abstracted excerpts from several sample charts targeting the abstraction content for the mental health conditions and including both ED and inpatient care. Their abstractions were compared to gold-standard abstractions previously completed by the developer of the measure specifications. Abstractors were considered fully trained when the trainer observed that they could reliably abstract the applicable gold-standard medical record excerpts.

Case Selection

Cases for the field test were selected using International Classification of Diseases 9th Revision Clinical Modification (ICD-9) codes for psychosis from administrative databases from each hospital for discharges occurring between January 1st,2012 and December 31st, 2013 (see Appendix for a list of ICD-9 codes used to select cases for abstraction).

The final sample goal for psychosis was a total of 100 cases selected from the two larger hospitals and 35 from the smaller hospital, with 25% replacement cases in order to have adequate sample after patients were excluded during the medical record abstraction phase. Because of limited sample sizes at each hospital for psychosis, all eligible patients were included in the final sample. See **Table 2b5.1** for sample sizes in each hospital.

Medical Record Abstractions

At each hospital, the two trained nurse abstractors were each assigned half of the case sample for psychosis. Data for each case were entered by the nurses into the electronic Pediatric Mental Health abstraction tool and both the raw data and auto-generated measure scores were uploaded to a central research database for further analysis.

At the two larger tertiary care hospitals, each nurse abstracted Pediatric Psychosis measures from 14 additional charts that were randomly selected from the other nurse's sample to facilitate assessment of inter-rater reliability (see inter-rater reliability testing results in **2a2.3** below). The 14 charts were among a total of 60 (10% sample) pulled for inter-rater reliability testing of quality measures we developed and tested across three different mental health diagnoses (psychosis, danger to self/suicidality, and substance abuse).

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**

☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Two existing administrative datasets were used to sample patients using the ICD9 codes.

The Pediatric Health Information System (PHIS) database was used to sample the medical records from two of the children's hospitals. This is a comparative pediatric database, and includes clinical and resource utilization data for inpatient, ambulatory surgery, emergency department and observation unit patient encounters for 45 children's hospitals. (More information about PHIS is available at: https://www.childrenshospitals.org/Programs-and-Services/Data-Analytics-and-Research/Pediatric-Health-Information-System)

The hospital administrative discharge databases were used to sample the medical records from the other hospitals.

1.3. What are the dates of the data used in testing? January 1, 2012-December 31st, 2013

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Three hospitals that admit children were included in the field test. All three are stand-alone children's hospitals. They are located in Washington state (Seattle Children's Hospital), Minnesota (University of Minnesota Children's Hospital), and Ohio (Cincinnati Children's Hospital). All have dedicated inpatient psychiatric units.

These hospitals were selected as they are all member organizations of the COE4CCN multi-stakeholder consortium of organizations that took part in the Center's measure development activities.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Table 1.6 Testing: Sociodemographic Characteristics of Patients Eligible for Measurement with Pediatric Psychosis: Timely Inpatient Psychiatric Consultation (N=252)

	N	%
Child gender		
Male	157	61
Female	95	37
Missing	1	0
Child race/ethnicity		
Hispanic	6	2
White	120	47
Black	69	27
Other	46	18
Missing	12	5
Insurance type		
Public	144	56
Private	103	40
Uninsured	5	2
Missing	1	0
PMCA category*		
Non-chronic condition	41	18
Non-complex chronic condition	88	40
Complex chronic condition	93	42
Missing	0	0

* PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015).[2] Available only at 2 of the 3 participating hospitals.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

N/A

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

To measure patient-level sociodemongraphic variables, we used patient gender, race, ethnicity, insurance type, and chronic disease status. These variables were derived from the administrative claims data from each participating hospital. Chronic disease status was captured using the Pediatric Medical Complexity Algorithm (PMCA), which categorizes pediatric inpatients using diagnostic ICD9 codes as having an acute medical condition only (non-chronic condition), a non-complex chronic condition, or a complex chronic condition.[2] Retrospective claims data needed to run PMCA were only available from 2 of the field test hospitals.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Critical data elements used in the measure were tested for inter-rater reliability of medical record abstraction. Reliability was measured using the prevalence adjusted bias adjusted kappa (PABAK) statistic for patient eligibility for measurement, and for the patient score for the quality measure. Kappa is a statistic that captures the proportion of agreement beyond that expected by chance, that is, the *achieved* beyond-chance agreement as a portion of the *possible* beyond-chance agreement.[3] PABAK is a measure of inter-rater reliability that adjusts the magnitude of the kappa statistic to take account of the influences of high or low prevalence and of inter-rater differences in assessment of prevalence. The PABAK statistic adjusts for high or low prevalence and is what we used in our calculations of inter-rater reliability.

Performance measure score was assessed for reliability across performance sites using the intra-class correlation coefficient (ICC). The ICC assesses the ratio of between site variation and within site variation on performance. Higher ICC implies that the between site variation (signal) is higher than the within site variation (noise). ICCs were computed using STATA SE 13.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Critical data elements:

There are two stages of medical record abstraction for which we tested inter-rater reliability for all Pediatric Mental Health Measures: patient eligibility for the measure; and patient score for the quality measure. For this measure, because there were no medical record exclusions, we did not measure patient eligibility kappas, since there were no abstractions for that stage.

The specific measure addressed in this submission was one of 6 psychosis measures included in the field test as part of the broader COE4CCN Pediatric Mental Health Measures in the Hospital Setting Project.

Across all 6 psychosis measures tested in the field, 120 records were sampled and abstracted by both nurse abstractors.
Kappa for patient measure score for all 6 psychosis measures (n=98 eligible patient charts):
0.62.
PABAK for patient measure score for all 6 psychosis measures (n=98 eligible patient charts):
0.72.

For the specific submitted measure, there was only a small subset (n=14) of the randomly sampled charts that were eligible. There were too few patients eligible for this measure to calculate kappa. Instead, we present the percent agreement.
Percent agreement for patient scores on the quality measure under consideration:
100%

Performance measure score:
We performed ICC testing for performance variation at the level of the hospital, since that is the intended level of measurement. However, despite adequate sample size at the patient level within each site (see Table 2b5.1 Testing below), the number of higher level clusters in our field test is limited to the 3 participating hospitals. Future measurement across a larger number of participating hospitals will give more generalizable estimations of ICC for this measure.
Hospital-level ICC=0.154 (95%CI 0.023-0.587). N=3 hospitals

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Critical data elements: Interpretation of Kappas is generally cited as follows[3,4]: ≤0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect.
Hence, inter-rater reliability for psychosis measures was substantial. For the specific submitted measure, percent agreement was perfect.

Performance measure score: Hospital level ICC based on the three hospitals is relatively high. ICCs ≥0.10 are considered relatively high.[5] Hence, the ICCs indicate that there are meaningful between-site performance differences.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☒ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

CRITICAL DATA ELEMENTS

ICD10 CONVERSION (no testing performed)

1. Statement of intent for the selection of ICD-10 codes:
    a. **The goal is to convert this measure to a new code set, fully consistent with the intent of the original measure.**
2. Excel spreadsheet with original ICD-9 codes from the Field test and the ICD9-ICD10 conversion table is attached at S2.b
3. Description of the process used to identify ICD-10 codes, including:
    a. Experts who assisted in the process:
        i. Bonnie Zima (co-chair Mental Health Working Group, see Ad.1)
        ii. Michael Murphy (co-chair Mental Health Working Group, see Ad.1)
    b. Name of the tool used to identify/map to ICD-10 codes:
        i. Transformation was based on the Centers for Medicaid and Medicare Services Gems tool.
    c. Stakeholder input was obtained from the COE4CCN Mental Health Multi-stakeholder Working Group. See below.

Psychosis ICD9 to ICD10 Conversion: Stakeholder Comments

A) Researcher and practitioner stakeholder #1:

"Psychosis - F44.89 - I usually think of dissociative disorders and conversion as not being delusional or psychotic. They are more loss of function than hallucinations, etc. So, I am not sure that this code belongs."

Response: consultation with stakeholder #3 and then deleted this code.

B) Researcher and practitioner stakeholder #2:

"I read all the new ICD 10 dx for both psychosis and substance abuse and they all seemed appropriate. They also all seemed to correspond pretty well to their ICD 9 antecedents. I am signing off on these lists. I think that the codes make sense."

Response: none needed

C) Researcher and practitioner stakeholder #3:

"re: Psychosis - F44.89, agree with [stakeholder #1] re: conversion is a somatoform disorder. Would delete."

"re: Psychosis - F44.89, I've honestly never heard of the dx "reactive confusion" and it's not in either the DSM 5 or DSM IVR. Thus I agree with [stakeholder #1]. I also wonder whether during this exercise we are getting caught up with a more historical shift within the DSM to align with the ICD…."

Response: Deleted F44.89

D) State Medicaid office stakeholder #4:

"The mental health folks in my agency are ahead of the rest of us as they have created crosswalks that make sense for our programs. Basically the codes are being based off of the DSM-5. The DSM-5 diagnoses lists both ICD-9 and ICD-10 codes with the diagnoses."

Response: Because we went through the DSM for psychosis and chose specific ICD9s for the field testing, and there is a consistent 1:1 match with ICD9 and ICD10, we decided to keep the crosswalk for ICD9-ICD10 for psychosis.

PERFORMANCE MEASURE SCORE

EMPIRICAL VALIDITY TESTING

We assessed the patient-level relationship between meeting the quality measure and three utilization outcomes that, per our conceptual model, were outcomes of interests and which we hypothesized a priori might have a relationship with the measure.

Multivariable regression was used to assess the independent relationship between meeting the measure and the validation metric of interest, independent of other confounders. Covariates were chosen based on face validity (gender and insurance type) and based on empirical evidence that they were associated with both the measure and the outcome measure (admitting hospital, and child race/ethnicity).

30 day readmission to the hospital (measured as readmission within 30 days of discharge, to the same hospital, since we did not have data on readmissions to other hospitals). (logistic model)

30-day return ED visit (measured as return visit within 30 days of discharge, to the same hospital, since we did not have data on readmissions to other hospitals). (logistic model)

Length of stay, measured in two ways:

Average inpatient days, truncated at 99th percentile, due to a few large outliers (linear model)

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—The RAND-UCLA Modified Delphi Method

The content validity of the group of quality measures developed in the COE4CCN Pediatric Mental Health measures effort, which included the psychosis measure proposed, was established using the RAND-UCLA Modified Delphi Method. The process began with the nomination of 10 individuals by 8 stakeholder organizations including the American Academy of Child and Adolescent Psychiatry, the AAP Committee on Pediatric Emergency Medicine, the AAP Task Force on Mental Health, the Medicaid Medical Directors Learning Network, the AAP Section on Hospitalist Medicine, Family Voices, the Society for Adolescent Medicine, and the Substance Abuse and Mental Health Services Administration. Nine of the nominees agreed to be members of our multi-stakeholder Delphi panel. All panelists were people deemed by the nominating organizations to have substantial expertise and/or experience related to child mental health (see Ad.1 for a list of panel members). The panel read the psychosis literature review written by project staff and reviewed and scored each proposed quality measure on validity. This method is a well-established, structured approach to measure evaluation that involves two rounds of independent panel member scoring, with group discussion in between.[1] After reviewing literature review and draft psychosis quality measures, panel members were asked to rate each measure's validity on a scale from 1 (low) to 9 (high). Validity was assessed by considering whether there was adequate scientific evidence or expert consensus to support its link to better outcomes; whether there would be health benefits associated with receiving measure-specified care; whether they would consider providers who adhere more consistently to the quality measure to be providing higher quality care; and whether adherence to the measure is under the control of health care providers and/or systems. The Delphi method has been found to be reliable and to have content, construct and predictive validity.[6-10] For a quality measure or measure component to move to the next stage of measure development, it had to have a median validity score $\geq$ 7 (1-9 scale) and be scored without disagreement based on the mean absolute deviation from the median after the second round of scoring. This process ensures that only measures widely judged to be valid moved forward into measure specification. See **Table 2b.2.3** for Delphi panel scores on the measure for this submission.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

CRITICAL DATA ELEMENTS

I CD10 CONVERSION (no testing performed)

PERFORMANCE MEASURE SCORE
EMPIRICAL VALIDITY TESTING

Table 2b2.3. Validation Metrics Pediatric Psychosis: Timely Inpatient Psychiatric Consultation (N=251)

	Met measure (n=213)	Did not meet measure (n=38)	Adjusted OR (95% CI)*	p-value
30-day readmissions, n (%)	28/213 (13.1%)	5/38 (13.2%)	1.00 (0.99-1.01)	0.64
30-day ED revisits, n (%)	21/213 (9.9%)	3/38 (7.9%)	1.00 (0.99-1.01)	0.77
			Adjusted coefficient (95% CI)*	
Length of stay (mean, days)**	13.2 (16.1)	15.9 (21.5)	-0.06 (-0.12-0.00)	**0.04**

*Adjusted for hospital, race/ethnicity, gender, and insurance type. OR assessed using logistic regression.
**Length of stay was available for n=249. This measure was truncated at the 99th percentile, to handle large outliers.

PERFORMANCE MEASURE SCORE
SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—The RAND-UCLA Modified Delphi Method The scores for this measure from the 9 members of the panel after round 2 of Delphi scoring (scoring done after discussions at the in-person meeting) are presented in the Table below.

Table 2b.2.3 Testing. Delphi Panel: Pediatric Psychosis: Timely Inpatient Psychiatric Consultation

	Median score (Scale 1-9)	Mean absolute deviation from median	Agreement status*
Validity	9.0	0.3	Agree
Feasibility	8.0	0.8	Agree

*This is a statistical assessment of whether panelists agreed (A), disagreed (D), or if status was indeterminate (I)

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

PERFORMANCE MEASURE SCORE
EMPIRICAL VALIDITY TESTING
The results of the field test present mixed results of empirical validity.

There were no statistically significant differences between those meeting and those failing the measure in readmissions and ED revisits. The low event rate for these outcome measures may have led to limited power to demonstrate a difference in readmission or ED return visits for patients passing versus failing this quality measure.

In contrast, length of stay was statistically significantly shorter for patient passing the measure. Though the effect size is relatively small (1.5 hours difference), the results support the validity of the measure, providing

evidence that patients with more timely psychiatric consultation have their psychiatric needs more rapidly addressed and are thus able to return home more rapidly.

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—The RAND-UCLA Modified Delphi Method The results from the Delphi panel show strong content validity for this measure, with median validity scores ≥8 (out of 9) following the Delphi panel.

2b3. EXCLUSIONS ANALYSIS

☒ **no exclusions — *skip to section 2b4***

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

NA

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

NA

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

NA

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*) **N/A**

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

As noted in the Submission Item 1b, we performed a field test of the quality measure under consideration. We measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital. Included patients were discharged from one of the three hospitals over the two year period (January 1, 2012-December 31, 2013). The performance scores are presented below in Tables 2b5.2a (performance variation across hospitals) and 2b5.2b (performance variation across socio-demographic characteristics)**.** We tested the difference in performance across the hospitals using an omnibus test for difference, and then performing individual comparisons between each hospitals performance and the performance of the group as a whole. We used Fisher's exact test to assess statistical significance for all comparisons.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?

(e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Table 2b5.2a. Performance Scores for Pediatric Psychosis: Timely Inpatient Psychiatric Consultation						
	Denominator	Numerator	Score	P-value for omnibus test*	Difference from overall mean of others	P-value for difference from overall mean of others*
Hospitals overall	253	214	84.6	0.0002	--	--
Hospital A	81	77	95.1	--	15.4	0.0013
Hospital B	141	108	76.6	--	-18.1	0.0001
Hospital C	31	29	93.6	--	10.2	0.19

*Statistical testing using Fisher's exact test

Table 2b5.2b. Psychiatric consult within 24 hours of admission (Children's Hospitals)						
	N	**%**	**SD**	**OR***	**LCL**	**UCL**
Child gender						
Male	157	84.7	36.1	0.96	0.47	1.96
Female (ref)	95	85.3	35.6			
Child race/ethnicity						
White (ref)	120	80.8	39.5			
Hispanic	6	83.3	40.8	1.19	0.13	10.64
Black	69	88.4	32.3	1.81	0.76	4.30
Other	46	91.3	28.5	2.49	0.81	7.64
Insurance type						
Private (ref)	103	85.4	35.5			
Public/uninsured	149	84.6	36.3	0.93	0.46	1.89
**PMCA category ** **						
Non-chronic (ref)	41	82.9	38.1			
Non-complex chronic	88	86.4	34.5	1.30	0.47	3.60
Complex chronic	93	80.7	39.7	0.86	0.33	2.25

*No performance differences by group were statistically significant. Differences tested using logistic regression.

**PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015).[2] Available only at 2 of the 3 participating hospitals.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

For this pilot test assessing for existing variation in this measure across more than one site, we found that we were able to detect statistically and clinically meaningful differences in hospital performance. Additional information from implementation of the measure at a larger scale, as described in Section 4.1, will assist in assessing variation across a larger group of hospitals.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

N/A

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

NA

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

NA

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

NA

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Missing data likely does not contribute to substantially or meaningfully biased estimates of performance for this measure.

There are two potential areas for missing data: at the level of the administrative claims, which are used for sampling patients, and in the medical abstraction stage.

Administrative Claims

There are two data fields used to identify patients, the diagnosis fields, and the patient age. Patient age is generally considered a reliable field and has minimal missing data.

A primary diagnosis is required for billing, and therefore also is rarely missing. It is known that some providers under-code for mental health diagnoses, which would lead to a risk of under recognition of eligible cases. This may lead to difficulty in capturing reliable estimates of performance at each hospital site, but is less likely to lead to biased estimates. In addition, it is likely that an admitted patient with psychosis is severely symptomatic and will need additional long term services, hence leading to a higher likelihood of a diagnosis being documented.

Medical abstraction

Missing data in the medical abstraction stage is interpreted as the patient not meeting the metric. It would be very unusual for a psychiatric consultation to take place and not to have documentation occur or for the documentation not to be timed and dated, due to medical legal pressure for both of those types of documentation. To the degree that patients are meeting metrics at the site and providers are not documenting this in the medical record (false negative performance scoring), performance measurement (and accompanying internal feedback or public reporting) will likely stimulate improved documentation.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

In the PHIS dataset (used for case finding at Seattle Children's and Cincinnati Children's), age is a required element, and so was not missing for any records for patients from the hospitals with PHIS data. We do not have documentation for how often data was missing from patient medical records regarding patient age at the other hospital, nor regarding missing information on timing of psychiatric consultation for any of the hospitals.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

It is unlikely that missing data contributes to substantial or meaningful biases of performance estimates. See item #2b7.1 for additional discussion of this.

REFERENCES

1. Brook RH. The RAND/UCLA appropriateness method. In: McCormick KA, Moore SR, Siegel RA, eds. *Clinical practice guidelines development:methodology perspectives*. Rockville, MD: Agency for Health Care Policy and Research; 1994.
2. Simon TD, Cawthon ML, Stanford S, et al. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. *Pediatrics.* 2014;133(6):e1647-1654.
3. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy.* 2005;85(3):257-268.
4. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.
5. Lyratzopoulos G, Elliott MN, Barbiere JM, et al. How can health care organizations be reliably compared?: Lessons from a national survey of patient experience. *Med Care.* 2011;49(8):724-733.
6. Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. *N Engl J Med.* 1998;338(26):1888-1896.

7. Shekelle PG, Chassin MR, Park RE. Assessing the predictive ability of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy. *Int J Technol Assess Health Care.* 1998;14(4):707-727.
8. Kravitz RL, Park RE, Kahan JP. Measuring the clinical consistency of panelists' appropriateness ratings: the case of coronary artery bypass surgery. *Health Policy.* 1997;42(2):135-143.
9. Hemingway H, Crook AM, Feder G, et al. Underuse of coronary revascularization procedures in patients considered appropriate candidates for revascularization. *N Engl J Med.* 2001;344 (9):645-654.
10. Selby JV, Fireman BH, Lundstrom RJ, et al. Variation among hospitals in coronaryangiography practices and outcomes after myocardial infarction in a large health maintenance organization. *N Engl J Med.* 1996;335(25):1888-1896.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

Electronic capture of this data could be operationalized through the use of three existing data fields in an electronic medical record and an additional custom field:

Existing fields:

1) Time of admission (T1)
2) Time of electronic note signature (T2)
3) Line of service for provider signing note

Custom field:

4) A field to complete if the consult was completed within 25-48 hours, with programmed options for justified reasons for delay, or None of the above. (R1)

If T2-T1 is =24 hours and line of service=psychiatry, then the measure will have been met for that patient admission.

If T2-T1 is 25-48 hours, then R1 must equal one of the justified reasons for delay.

For hospitals without electronic medical records, the data could not be easily captured in an electronic form.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

In field testing, we abstracted this measure both from paper charts as well as electronic health records. We found that EHR

abstractions were easier due to the structured notes that automatically identified provider names, titles, and departments, facilitating efficient identification of the consultant note.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
No proprietary elements are used in implementing this measure. There are no licenses or fees or other requirements needed to use any aspect of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This measure is part of a set of mental health quality measures the COE4CCN developed as part of the Pediatric Quality Measurement Program, funded by AHRQ, using CHIPRA monies. It has not yet been implemented as the development, validation, and testing were just recently completed. The tools needed to abstract the measures, available online at the website in S.1, are publicly available and non-proprietary, so interested parties can implement them at any time.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
The Children's Hospital Association (CHA) has had representation on the National Advisory Board for COE4CCN since its inception.

CHA has shown great interest in promoting the adoption of inpatient and ED-based measures developed by our Center. The intended audience would be hospital administrators at CHA member hospitals. We would intend to work with CHA to implement these measures over the next several years.
We also intend to publish the development and field testing of these measures in peer reviewed pediatric journals over the next 12 months. Within these publications we will include the URL where the measure data abstraction tool, measure specifications, and abstractor training materials are housed promoting further access to and dissemination of the measures.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not available as not in use for performance improvement.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Credible rationale

The overall goal behind capturing performance results for this measure is to improve timely access to inpatient psychiatric consultation for a severely ill pediatric population. We anticipate that increasing the focus on this process measure will instigate not only improved timeliness, but also enhanced efforts to provide adequate specialty services to these patients.

As experience has borne out, quality measurement efforts can drive improvements in care, whether through increasing focus on an area of care in internal audit and feedback efforts, or through reputational or financial incentive programs (ie, public reporting or pay for performance). We anticipate that the performance results for this measure would drive improvement through similar mechanisms.

Some of the major lessons in quality improvement over the past two decades are that the most effective performance measures are valid, feasible, and consistently specified across requested reports, so that providers do not need to generate data for multiple versions of similar measures. These goals were part of the impetus for the national Pediatric Quality Measures Program that funded our efforts, and provide the underlying rationale for this submission for endorsement.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually

both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** P3_Inpt_Consult_in_24_hours_Appendix_FOR_SUBMISSION-635803524468735341.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Seattle Children´s Research Institute
Co.2 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-
Co.3 **Measure Developer if different from Measure Steward:** Seattle Children´s Research Institute
Co.4 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The COE4CCN convened two expert groups to assist in the development of the Pediatric Mental Health Measures in the Hospital

Setting--the Mental Health Working Group within the COE4CCN and an external panel of experts for the Delphi panel. Please see descriptions of the groups' roles in development as well as member names listed below.

I. Mental Health Working Group: This was a group of pediatric mental health and general pediatrics experts, as well as state Medicaid leadership. Reviewed secondary database analyses of prevalence of common and costly mental health diagnoses. Developed ICD9 code definitions to identify diagnoses of interest. Reviewed and edited the literature reviews conducted by COE4CCN staff. Provided content expertise during development of the detailed measure specifications and data abstraction tool. Participated in the planning and implementation of the field test as well as interpretation of the field test results.

Members of the MHWG:

Naomi S. Bardach, MD, MAS
Assistant Professor of Pediatrics and Health Policy
Department of Pediatrics
Philip R. Lee Institute of Health Policy
University of California San Francisco

Tumaini Ruker Coker, MD, MBA
Assistant Professor of Pediatrics
David Geffen School of Medicine
University of California, Los Angeles
Associate Natural Scientist
RAND, Santa Monica

Glenace Edwall, PsyD, PhD, MPP
Director, Children's Mental Health Division
Minnesota State Health Access Data Assistance Center
Minnesota Department of Human Services

Penny Knapp, MD
Professor Emeritus
Departments of Psychiatry & Pediatrics
University of California Davis

Rita Mangione-Smith, MD, MPH
Professor and Chief | Division of General Pediatrics and Hospital Medicine
University of Washington Department of Pediatrics
Director | Quality of Care Research Fellowship
UW Department of Pediatrics and Seattle Children's Hospital
Investigator | Center for Child Health, Behavior, and Development
Seattle Children´s Research Institute

Michael Murphy, EdD
Associate Professor
Department of Psychology
Harvard Medical School
Staff Psychologist
Department of Child Psychiatry
Massachusetts General Hospital

Laura Marie Prager, MD
Associate Professor of Psychiatry
Department of Child Psychiatry
Massachusetts General Hospital

Laura Richardson, MD, MPH

Professor
Department of Pediatrics and Psychiatry
Division of Adolescent Medicine
University of Washington
Investigator
Center for Child Health, Behavior, and Development
Seattle Children's Research Institute

Bonnie Zima, MD, MPH
Professor-in-Residence
Department of Psychiatry
University of California, Los Angeles
Associate Director
UCLA Health Services Research Center

Delphi panel: Reviewed the literature review and secondary database analyses as prepared by the MHWG and COE staff. Reviewed suggested indicators for face validity and content validity based on the above materials and based on member expertise in the field.

Members of the Delphi panel:

Gary Blau, PhD
Chief, Child, Adolescent and Family Branch,
Center for Mental Health Services (CMHS),
Substance Abuse and Mental Health Services Administration (SAMHSA), Rockville, MD.
Clinical Faculty, Yale Child Study Center, Yale University

Regina Bussing, MD, MSHS
Professor, Division of Child and Adolescent Psychiatry,
Department of Psychiatry, Department of Pediatrics,
and Department of Clinical and Health Psychology,
University of Florida, Gainesville, FL
Director, Florida Outreach Project for Children and Young Adults Who Are Deaf-Blind

Thomas Chun, MD, MPH
Associate Professor, Departments of Emergency Medicine and Pediatrics
Assistant Dean of Admissions
Chair, Admissions Committee
The Alpert Medical School, Brown University
Medical Staff, Department of Pediatric Emergency Medicine
Hasbro Children's Hospital

Sean Ervin, MD, PhD
Assistant Professor in Pediatrics & General Internal Medicine
Hospitalist Medicine
Head of Section- Pediatric Hospital Medicine
Wake Forest University, School of Medicine
Winston-Salem, NC

Doris Lotz, MD, MPH
Medicaid Medical Director
New Hampshire Department of Health and Human Services
Office of Medicaid Business and Policy
Instructor, Geisel School of Medicine at Dartmouth, Department of Psychiatry

Lynn Pedraza, PhD

Executive Director of Family Voices,
Albuquerque, NM

Karen Pierce, MD, DLFAPA, DLFAACAP
Clinical Associate Professor, The Feinberg School of Medicine,
Northwestern University Medical School,
Department of Psychiatry and Behavioral Sciences, Chicago, IL,
President, Illinois Academy of Child Psychiatry

Robert Sege, MD, PhD, FAAP
Professor of Pediatrics, Boston University School of Medicine
Director, Division of Family and Child Advocacy, Boston Medical Center
Core Faculty, Harvard Injury Control Research Center
Core Faculty, Harvard Youth Violence Prevention Center

Gail Slap, MD, MSc
Professor of Pediatrics, Department of Pediatrics,
Professor of Medicine, Department of Medicine,
University of Pennsylvania School of Medicine

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2015

Ad.3 **Month and Year of most recent revision:** 11, 2014

Ad.4 **What is your frequency for review/update of this measure?** Every 6 months

Ad.5 **When is the next scheduled review/update for this measure?** 03, 2014

Ad.6 **Copyright statement:** None

Ad.7 **Disclaimers:** None

Ad.8 **Additional Information/Comments:** None

PSYCHOSIS

Note: There are a number of ICD9 codes that have mapped to the same ICD10 code, and one ICD9 code that mapped to 2 ICD10 codes

ICD9 used in Field test	ICD9 label	ICD10 conversion from CMS GEMS tool	ICD10 label
291.3	alcoh psy dis w hallucin	F10.951	Alcohol use, unspecified with alcohol-induced psychotic disorder with hallucinations
291.5	alcoh psych dis w delus	F10.950	Alcohol use, unspecified with alcohol-induced psychotic disorder with delusions
292.11	drug psych disor w delus	F19.950	Other psychoactive substance use, unspecified with psychoactive substance-induced psychotic disorder with delusions
292.12	drug psy dis w hallucin	F19.951	Other psychoactive substance use, unspecified with psychoactive substance-induced psychotic disorder with hallucinations
293.81	psy dis w delus oth dis	F06.2	Psychotic disorder with delusions due to known physiological condition
293.82	psy dis w halluc oth dis	F06.0	Psychotic disorder with hallucinations due to known physiological condition
295.3	paranoid schizo-unspec	F20.0	Paranoid schizophrenia
295.31	paranoid schizo-subchr	F20.0	Paranoid schizophrenia
295.32	paranoid schizo-chronic	F20.0	Paranoid schizophrenia
295.33	paran schizo-subchr/exac	F20.0	Paranoid schizophrenia
295.34	paran schizo-chr/exacerb	F20.0	Paranoid schizophrenia
295.4	schizophreniform dis nos	F20.81	Schizophreniform disorder
295.41	schizophrenic dis-subchr	F20.81	Schizophreniform disorder
295.42	schizophren dis-chronic	F20.81	Schizophreniform disorder
295.43	schizo dis-subchr/exacer	F20.81	Schizophreniform disorder
295.44	schizophr dis-chr/exacer	F20.81	Schizophreniform disorder
295.7	schizoaffective dis nos	F25.9	Schizoaffective disorder, unspecified
295.71	schizoaffectv dis-subchr	F25.9	Schizoaffective disorder, unspecified
295.72	schizoaffective dis-chr	F25.9	Schizoaffective disorder, unspecified
295.73	schizoaff dis-subch/exac	F25.9	Schizoaffective disorder, unspecified
295.74	schizoafftv dis-chr/exac	F25.9	Schizoaffective disorder, unspecified
295.9	schizophrenia nos-unspec	F20.9	Schizophrenia, unspecified
295.91	schizophrenia nos-subchr	F20.9	Schizophrenia, unspecified
295.92	schizophrenia nos-chr	F20.9	Schizophrenia, unspecified
295.93	schizo nos-subchr/exacer	F20.9	Schizophrenia, unspecified
295.94	schizo nos-chr/exacerb	F20.9	Schizophrenia, unspecified
296.24	depr psychos-sev w psych	F32.3	Major depressive disorder, single episode, severe with psychotic features
296.44	bipol i manic-sev w psy	F31.2	Bipolar disorder, current episode manic severe with psychotic features
297.1	delusional disorder	F22	Delusional disorders

297.2	paraphrenia	F22	Delusional disorders
297.3	shared psychotic disord	F22	Delusional disorders
298.0	react depress psychosis	F32.3	Major depressive disorder, single episode, severe with psychotic features (Note: This is a duplicate, with two ICD10 codes for one ICD9)
298.0	react depress psychosis	F33.3	Major depressive disorder, recurrent, severe with psychotic symptoms (Note: This is a duplicate, with two ICD10 codes for one ICD9)
298.1	excitativ type psychosis	F28	Other psychotic disorder not due to a substance or known physiological condition
298.3	acute paranoid reaction	F23	Brief psychotic disorder
298.4	psychogen paranoid psych	F23	Brief psychotic disorder
298.8	react psychosis nec/nos	F23	Brief psychotic disorder
298.9	psychosis nos	F29	Unspecified psychosis not due to a substance or known physiological condition

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2806
Measure Title: Pediatric Psychosis: Screening for Drugs of Abuse in the Emergency Department
Measure Steward: Seattle Children´s Research Institute
Brief Description of Measure: Percentage of children/adolescents age =5 to =19 years-old seen in the emergency department with psychotic symptoms who are screened for alcohol or drugs of abuse
Developer Rationale: In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.
The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of quality measures related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. Pediatrics 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is an indicator designed to fill this key measurement gap.
The COE4CCN Mental Health Working Group (see item Ad.1 for more details on this group) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group. We found that psychosis was the third most common reason for pediatric mental health hospitalizations (Bardach et al. Pediatrics 2014). Literature reviews were then conducted separately for each of the most common conditions, and one of these reviews focused on children evaluated and treated for psychosis in the ED and inpatient settings. See Evidence form for conceptual model underlying the rationale for the measures.
Based on the literature reviews, we developed a list of draft quality measures to assess the quality of pediatric mental health care in the ED and inpatient settings, including specific measures to assess the quality of care for children presenting with psychotic symptoms. The validity and feasibility of these indicators were then evaluated by an expert panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in 5 hospitals in Washington state, Ohio, and Minnesota. This measure submission presents the results of this development and field testing work.

Numerator Statement: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests.
Denominator Statement: Patients aged =5 to =19 years-old seen in the emergency department with psychotic symptoms.
Denominator Exclusions: No patients were excluded from the target population.

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The evidence for this process measure should demonstrate that the process of checking for drugs of abuse for a patient who presents with psychotic symptoms should improve outcomes and limit missed diagnoses, lack of treatment, and representation to care.

The developer provides the following information for this facility-level process measure:

- The developer cites a 2013 guideline from the American Academy of Child and Adolescent Psychiatry (AACAP): "*Clinical Practice Guideline Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, substance abuse, developmental disabilities, psychosocial stressors, and medical problems. [CS]*
  - There are no neuroimaging, psychological, or laboratory tests that establish a diagnosis of schizophrenia. The medical evaluation focuses on ruling out nonpsychiatric causes of psychosis and establishing baseline laboratory parameters for monitoring medication therapy. ... Toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out."
  - The recommendation carries AACAP's highest grade of clinical standard—i.e., based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials, and/or overwhelming clinical consensus).
  - The guideline does not provide citations for the recommendation, so there is no summary on the quantity, quality, and consistency of the evidence nor a grade. The recommendation's highest grade is derived from overwhemling clinical consensus.
- The developer provides no additional reviews or literature, indicating no studies were identified since AACAP published the guideline in 2013.
- Per the **NQF Algorithm for Evidence,** there is no systematic review (box 3) and no additional empirical evidence submitted (box 7). The Committee's evaluation should focus on whether the rating should be INSUFFICIENT WITH EVIDENCE EXCEPTION or INSUFFICIENT (boxes 10-->12).

Questions for the Committee

- *Are there (OR could there be) performance measures of a related health outcome, OR evidence-based clinical intermediate outcome?*
- *Is there evidence of a systematic opinion (e.g., national/international consensus recommendation) that the benefits of what is being measured outweigh potential harms)?*
- *Does the Steering Committee agree that it is OK (or beneficial) to hold providers accountable in the absence of empirical evidence of benefits to patients?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following information:

- Mental health hospitalizations for pediatrics represented 9.1% of all hospitalizations for children ages >2 in 2009, with psychosis the third most common mental health diagnosis (12.1%).
- Performance gap information was derived from testing the measure using data aggregated over two years from three children's hospitals and two community hospitals. Included patients were discharged from one of the

hospital EDs during the two year measurement period (January 1, 2012-December 31, 2013). The performance scores are presented below:

# of hospitals: 5
# of patients: 257
Mean hospital-level score (0-100 scale): 28.8
95% Confidence interval: 24.5-33.1
Min-Max: 17.8-83.3

- Differences were measured in performance scores by gender, race, insurance type, and chronic disease category (measured using the Pediatric Medical Complexity Algorithm.
- Using linear regression, the developer found chronic disease category was associated with performance, with patients with non-complex chronic conditions more often tested (24.6%, N=67) than children with only an acute condition (15.5%, N=55) or children with a complex chronic condition (16.9%, N=80), with a difference in performance of 9.2 (95% CI 0.1-18.2) compared to patients with acute conditions only.
- The developer noted no other statistically significant differences by patient socio-demographic characteristics from its testing.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Since no disparities were identified during testing, is the Committee aware of evidence that disparities exist in this area of healthcare?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities])*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- No directly applicable evidence available. The guideline has an "out" in it: "when exposure to drugs of abuse cannot otherwise be ruled out" - makes it hard to know what the rate should be.
- I think that the premise is that substance abuse can co-occur with schizophrenia. That is common in the adult population, but the work-up of new onset psychosis in children (especially those in the age group in which schizophrenia is very uncommon) should look for non-psychiatric causes first and there are many classes of drugs that are not drugs of abuse that when either taken in too large doses or ingested by children can result in psychosis. Steroids, ACE inhibitors, stimulant medication etc. can do this. Presentation of psychosis in the ED in children should first rule out medical causes including ingestions or inadvertent overdoses of classes of drugs that can cause psychosis as should other brain pathology. In the ED while the behavior issues around psychosis are the same for schizophrenia and medical causes the risks of harm and death from drug effects is more urgent. This measure not only has no evidence to support it, but it fails to recognize the important medical issues that might cause this symptom. A better measure would be to look for use or ingestion of any drug that might cause psychotic symptoms, not just drugs of abuse looking for the co-occurrence of conditions. The differential diagnosis of new onset psychosis is far wider than psychiatric disorders and ruling out medical causes with different treatments other than antipsychotics is important. I didn't find any guidelines for evaluation of psychosis in children at the ED level.
- Recommendation 3 from AACAP states that screening is indicated when “exposure to drugs of abuse cannot otherwise be ruled out”.
- The recommendation carries the highest grade of clinical standard, overwhelming consensus of best practice
- There is limited evidence to support the use of a drug/alcohol screen for patients with psychotic symptoms. The evidence is based on 2013 guideline from the American Academy of Child and Adolescent Psychiatry (AACAP): Clinical Practice Guideline Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, substance abuse, developmental disabilities, psychosocial stressors, and medical problems. The guideline does

not provide citations for the recommendation, so there is no summary on the quantity, quality, and consistency of the evidence nor a grade. The recommendation's highest grade is derived from clinical consensus."

1b. Performance Gap.

- I am surprised at the low rate of testing found by the developer. Somewhat variable (wide min-max range, but CI not so wide). This seems less than optimal - would be good to have a better understanding of why this is occurring.
- The small number of patients, unclear whether or not it includes kids that presented with psychosis, but didn't have disease, makes it difficult to say much of anything useful about this measure. The sample was too small to outline disparities as it was too small to divide into groups and be statistically significant. Also question whether or not this measure belongs in psychiatry or in emergency medicine with the focus on identifying a cause for the symptoms and treating as indicated (e.g. lupus would require different treatment than drug ingestion which is different than schizophrenia). As well schizophrenia is relatively rare in children especially younger ones.
- There is a performance gap not related to socio-demographic differences
- Not enough information available to tag as disparities sensitive.
    - Measured over a 2 year period at 3 children's hospitals and 2 community hospitals.
    - Hospital mean level score was 28.8

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

The developer provides the following information:

- This is a facility-level measure; higher score = better quality.
- The data sources are administrative claims and electronic health records and paper medical records. The developer provides an attachment for the applicable codes.
- The developer defines the numerator as: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests. The denominator is defined as: Patients 5 to 19 years seen in the emergency department with psychotic symptoms. There are no denominator exclusions, and patients are identified from hospital administrative data.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer provides the following information:

- Empirical testing for reliability was conducted at a critical data elements level and performance measure score level.

- Testing was conducted at five facilities (Seattle Children's Hospital, Cincinnati Children's Hospital, University of Minnesota Children's Hospital, Fairview Ridges Hospital (MN), and Maple Grove Hospital (MN) using 2-year retrospective data (Jan 2012-Dec 2013); N=257 patients.
  - Critical data elements were tested using inter-rater reliability of medical record abstraction.
    - The total population sample size was N=257
    - For this specific measure, however, the sampling N=4 patients—too few to calculate a Kappa. The developer reports, however, 100% agreement.
  - Performance measure score reliability was assessed using the intra-class correlation coefficient (ICC). The ICC assesses the ratio of between site variation and within site variation on performance. Higher ICC implies that the between site variation (signal) is higher than the within site variation (noise)
    - ICCs were computed using STATA SE 13.
    - The developer reports the hospital-level ICC=0.42 (95%CI 0.16-0.73); N=5 hospitals
    - The developer reports that ICCs ≥0.10 indicate that there are meaningful between-site performance differences.
- Per the **NQF Algorithm for Reliability**, empirical testing was performed at the level of the computed performance measure score and so the eligible ratings are HIGH, MODERATE, or LOW (box3-->6)

Questions for the Committee

- *Does the Committee concur with the developer's conclusion that the results demonstrate sufficient reliability so that differences in performance can be identified?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The goal of the measure is to improve outcomes for pediatric patients admitted with psychotic symptoms, which should improve outcomes and limit missed diagnosis, lack of treatment, and representation to care.
- The numerator is: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests. The denominator is: Patients aged 5 to 19 years seen in the emergency department with psychotic symptoms. There were no denominator exclusions. Patients are identified from hospital administrative data.
- The evidence for the specifications provided by the developer centers on an AACAP recommendation that is based on "overwhelming clinical consensus."
- The specifications appear consistent with the AACAP recommendation, which notes, "Toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out."

Question for the Committee

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer provides the following information:

- The developer tested face validity of the performance measure score. (Note, the developer checks testing of critical data elements, but then indicates no empirical testing was done. The material describes the developer's ICD conversion process.)
  - The developer performed systematic face validity assessment (RAND-UCLA Modified Delphi) of whether panelists "would consider providers who adhere more consistently to the quality

measure to be providing higher quality care," which we interpret as face validity assessment at the level of the **computed measure score** (as required by NQF).

- The panelists concluded there was face validity, although other factors were bundled with the assessment.
- Per the **NQF Algorithm for Validity**, when relying only on face validity, the eligible ratings are MODERATE OR LOW (box 4-->5).

Questions for the Committee

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- No exclusions

2b4. Risk adjustment:

- No risk adjustment or risk stratification

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provides the following information:

- The developer tested the difference in performance across the five hospitals using an omnibus test for difference, and then performed individual comparisons between each hospital's performance and the mean of all other hospitals.
- The developer used used ANOVA testing for the omnibus test, and a t-test to assess for individual comparisons between each hospital and the mean of all others.
  - The developer indicates the results detect statistically and clinically meaningful differences in hospital performance.

Question for the Committee

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

- The developer notes is unlikely that missing data contributes to substantial or meaningful biases of performance estimates. The two potential areas for missing data are at the level of the administrative claims and medical abstraction stage. Missing data in the medical abstraction stage are interpreted as the patient not meeting the measure specifications.
  - The developer posits it would be very unusual for a laboratory test (urine or serum) to be sent, processed, and not documented given the regulations around laboratory and quality insurance, as well as the need to be reimbursed for the testing.
  - The developer concludes there is unlikely to be a substantial incidence of false negatives for the measure due to missing data or biased performance results due to differentially missing data.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- A sample of four is too small to make any determinations from. Even if they all measured "drugs of abuse" it doesn't mean that they were measuring the same thing or the same set of drugs. "drugs of abuse" is not definitive and therefore difficult to reproduce without further definition. Unclear if they considered anabolic steroids which can be abused, but are typically not drugs of abuse from a substance abuse standpoint.

- Reliability testing:
    - Critical data elements were tested on only 4 subjects (100% agreement)
    - Performance measure reliability at the hospital level (n-5), ICC = 0.42.

2a2. Reliability testing

- Whether the numerator is drug AND alcohol testing or drug OR alcohol testing is not clear - stated differently in different places.
- Denominator is based on ER diagnoses which seems adequate.
- Drug screens vary in terms of the drugs that are included in the panel. The measure doesn't list the particular drugs that they are referring to except to call them "drugs of abuse" and to talk about co-occurring substance abuse. It would be difficult to know if the same drugs were being measured.
- Looks at whether or not results reported, not whether or not they're used by clinicians. Why is it a composite score (i.e., partial credit if only 1 of the 2 tested) and not "all or none"?
- Clearly defined
- Data sources are administrative claims and electronic health records and paper medical records. Applicable codes are available by developer. Specifications seem appropriate:
    - numerator is: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests.
    - denominator is defined as: Patients 5 to 19 years seen in the emergency department with psychotic symptoms. There are no denominator exclusions, and patients are identified from hospital administrative data.
    - only concern may be with paper charts"

2b1. Validity Specifications

- There is no consideration of the "out" that is provided in the guideline (which is the only evidence supporting the measure).
- While the specifications may be consistent with the evidence, the limitation of toxicology testing to drugs of abuse and the focus on co-occurring mental illness and substance abuse in the documents belie the fact that psychosis may be exposure to a class of drugs not related to abuse and not in fact related to schizophrenia at all.
- Evidence for the specifications is based more on clinical consensus rather than scientific evidence.

2b2. Validity Testing

- No empirical validity testing done. Score from Delphi group acceptable but on the low side.
- The validity of this measure is confounded as it is measured with other factors. The sample size and number of hospitals is small also making conclusions difficult to make. As well it is unclear that this measure improved outcome, function or treatment since they were only looking for co-occurring substance use and not psychosis related to other drugs.
- It looks like 78.6% of the patients in the validation set came from 2 of the 5 hospitals. Is this a broad enough population?
- Face validity measured per developer, not a lot of information given
- Face validity is sufficient

2b3-2b7. Threats to Validity

- Not likely - agree with developer that these are very clear data elements.
- While it would be difficult to lose a lab test, it is unclear that the lab tests would all be the same across the country since toxicology screens differ between regions, hospitals, and labs. Unclear that this constitutes quality care as there are no specifics for what is being tested for, what is considered abnormal, and how the information is being used.
- Two potential areas for missing data are at the level of the administrative claims and medical abstraction stage. Lab tests are typically documented in medical record.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer provides the following information:

- ALL data elements are in defined fields in a combination of electronic sources.
- Data are generated or collected by and used by healthcare personnel during the provision of care.

Questions for the Committee

- *Do you concur that the required data elements are routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Feasible
- Data already collected , feasible to extract from electronic sources
- Electronic records and claims should include such testing however they are unlikely to include the details of the testing (tox screens vary and so may not be measuring the same things).
- Data elements are defined fields in EMR and collected during treatment.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- This measure is not in use. It has not been implemented as the development, validation, and testing were just recently completed.
- Planned use include: Quality Improvement with Benchmarking (external benchmarking to multiple organizations) and quality Improvement (Internal to the specific organization)
- There were no unintended consequences identified during testing.

Questions for the Committee*:*

- *The developer indicates use for benchmarking and quality improvement. NQF endorsement focuses on primarily accountability, and then appropriateness for quality improvement. Is this measure appropriate for accountability purposes?*
- *Can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- No unintended consequences
- This measure is incomplete for the appropriate emergent evaluation of psychosis as it excludes looking for classes of drugs that are not drugs of abuse. It is important to look for co-occurring substance abuse (or psychosis related to drugs of abuse), but that is only part of the equation. Using a measure that doesn't include all of the possibilities gives the impression that this is all that is necessary to provide quality care.
- Not in use yet, would be good for quality improvement/benchmarking.

Criterion 5: Related and Competing Measures

- No related and competing measures

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: **Pediatric Psychosis**: Screening for Drugs of Abuse in the Emergency Department

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the

strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ **Process**: Screening for drugs of abuse for pediatric patients who present to the Emergency Department with symptoms of psychosis.

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

N/A

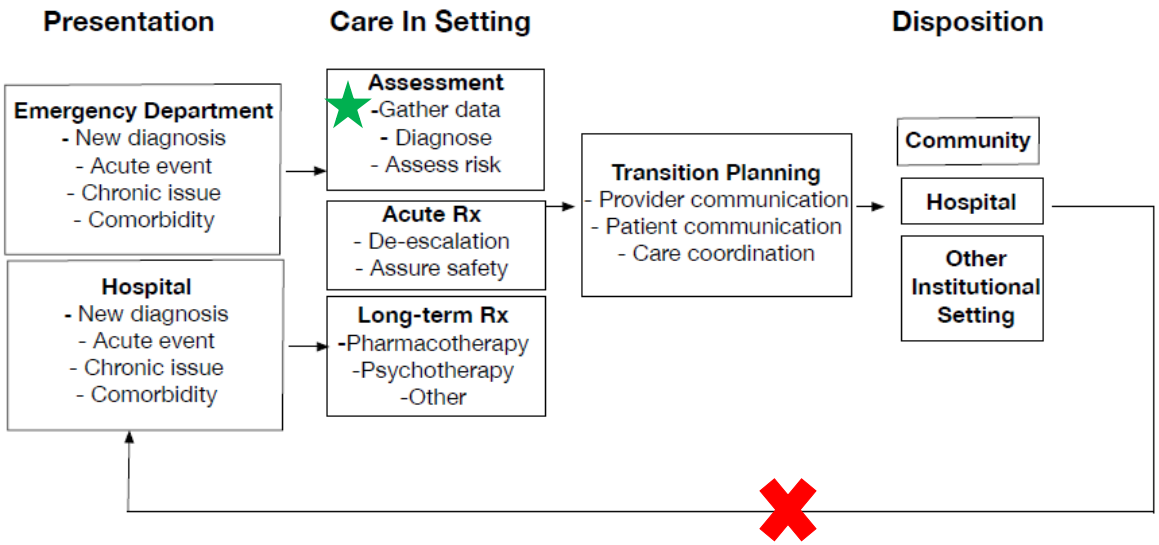
1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

N/A

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.



This diagram depicts the relationship between the care process of interest, marked with a green star, and the target outcomes to prevent (rehospitalizations and re-presentations to the ED), marked with the red X. The proposed measure focuses on whether one element of "Gather data" (Assessment box) was performed. If the process of checking for drugs of abuse for a patient who presents with psychotic symptoms is not performed, this may lead to a missed diagnosis, lack of treatment, and representation to care.

Summary: Overall, there is not extensive empirical literature supporting this process measure, but the benefits likely far outweigh the risks.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

McClellan J, Werry J, Bernet W, Arnold V, Beitchman J, Benson RS, Bukstein O, Kinlan J, Rue D, Shaw J, Kroeger K: Practice parameter for the assessment and treatment of children and adolescents with schizophrenia, Journal of the American Academy of Child and Adolescent Psychiatry 2013, Volume 52, Issue 9, Pages 976–990

http://www.jaacap.com/article/S0890-8567(13)00112-3/fulltext

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, ***substance abuse***, developmental disabilities, psychosocial stressors, and medical problems. [CS]

Youth with suspected schizophrenia require a thorough psychiatric and medical evaluation, including the assessment for common comorbid conditions, such as substance abuse or cognitive delays. When present, active psychotic symptoms are generally prioritized as the main target for treatment. Comorbid conditions, such as substance abuse, may respond better to treatment once acute symptoms of schizophrenia are stabilized. However, any life-threatening symptoms, such as suicidal behavior or severe aggressive behaviors, must be prioritized in the treatment plan.

There are no neuroimaging, psychological, or laboratory tests that establish a diagnosis of schizophrenia. **The medical evaluation focuses on ruling out nonpsychiatric causes of psychosis** and establishing baseline laboratory parameters for monitoring medication therapy. More extensive evaluation is indicated for atypical presentations, such as a gross deterioration in cognitive and motor abilities, focal neurologic symptoms, or delirium.

Assessments are obtained based on specific medical indications, e.g., neuroimaging studies when neurologic symptoms are present or an electroencephalogram for a clinical history suggestive of seizures**. Toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out**. Genetic testing is indicated if there are associated dysmorphic or syndromic features. Similarly, tests to rule out specific syndromes or diseases (e.g., amino acid screens for inborn errors of metabolism, ceruloplasmin for Wilson disease, porphobilinogen for acute intermittent porphyria) are indicated for clinical presentations suggestive of the specific syndrome in question. Broad screening for rare medical conditions is not likely to be informative in individuals with psychosis who do not present with other neurologic or medical concerns.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

The AACAP guidelines granted this their highest grading:

•Clinical Standard [CS] is applied to recommendations that are based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials) and/or overwhelming clinical consensus

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

•Clinical Guideline [CG] is applied to recommendations that are based on strong empirical evidence (e.g., nonrandomized controlled trials, cohort studies, case-control studies) and/or strong clinical consensus

•Clinical Option [OP] is applied to recommendations that are based on emerging empirical evidence (e.g., uncontrolled trials or case series/reports) or clinical opinion, but lack strong empirical evidence and/or strong clinical consensus

•Not Endorsed [NE] is applied to practices that are known to be ineffective or contraindicated

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, *substance abuse*, developmental disabilities, psychosocial stressors, and medical problems. [CS]

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The AACAP guideline does not provide citations for the Recommendation and so there is no grade assigned for the quality of the quoted evidence to support the Recommendation. The specific endorsement of drugs of abuse screening within Recommendation 3 is therefore not supported with citations of evidence. Nevertheless, the guidelines granted the Recommendation overall the highest grading of Clinical Standard [CS] (defined below). Thus, this recommendation is bolstered by overwhelming clinical consensus.

"Clinical Standard [CS] is applied to recommendations that are based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials) and/or overwhelming clinical consensus"

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

•Clinical Guideline [CG] is applied to recommendations that are based on strong empirical evidence (e.g., nonrandomized controlled trials, cohort studies, case-control studies) and/or strong clinical consensus

•Clinical Option [OP] is applied to recommendations that are based on emerging empirical evidence (e.g., uncontrolled trials or case series/reports) or clinical opinion, but lack strong empirical evidence and/or strong clinical consensus

•Not Endorsed [NE] is applied to practices that are known to be ineffective or contraindicated

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

NA

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

NA

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

NA

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

NA

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

NA

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

No studies providing new evidence to support this quality measure were identified since the publishing of the AACAP guideline in 2013.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
P2_Screen_for_Tox_evidence_attachment_2015_09_30_FOR_SUBMISSION.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.
The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of quality measures related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. Pediatrics 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is an indicator designed to fill this key measurement gap.
The COE4CCN Mental Health Working Group (see item Ad.1 for more details on this group) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group. We found that psychosis was the third most common reason for pediatric mental health hospitalizations (Bardach et al. Pediatrics 2014). Literature reviews were then conducted separately for each of the most common conditions, and one of these reviews focused on children evaluated and treated for psychosis in the ED and inpatient settings. See Evidence form for conceptual model underlying the rationale for the measures.
Based on the literature reviews, we developed a list of draft quality measures to assess the quality of pediatric mental health care in the ED and inpatient settings, including specific measures to assess the quality of care for children presenting with psychotic symptoms. The validity and feasibility of these indicators were then evaluated by an expert panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in 5 hospitals in Washington state, Ohio, and Minnesota. This measure submission presents the results of this development and field testing work.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In a field test of this quality measure, performed as part of the funded development work, we measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital and from two community hospitals in Minnesota, Fairview Ridges Hospital and Maple Grove Hospital. Included patients were discharged from one of the hospital EDs during the two year measurement period (January 1, 2012-December 31, 2013). The performance scores are presented below.

# of hospitals: 5
# of patients: 257
Mean hospital-level score (0-100 scale): 28.8
95% Confidence interval: 24.5-33.1
Min-Max: 17.8-83.3

See Testing form, item 2b.5.2a for data on individual hospital performance.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In the field testing described above, we measured differences in performance scores by gender, race, insurance type, and chronic disease category (measured using the Pediatric Medical Complexity Algorithm—Simon et al. Pediatrics 2015). Chronic disease category was associated with performance, with patients with non-complex chronic conditions more often tested (24.6%, n=67) than children with only an acute condition (15.5%, n=55) or children with a complex chronic condition (16.9%, n=80), with a difference in performance of 9.2 (95% CI 0.1-18.2) compared to patients with acute conditions only. The confidence interval and statistical testing were generated using linear regression.

There were no other statistically significant differences by patient socio-demographic characteristics in our testing. Please see Testing form, item 2b.5.2b for data.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Psychosis in pediatric patients is a high priority aspect of healthcare, with substantial inpatient utilization and high severity of illness, in addition to a number of associated costs to the healthcare system and to patients and families. Mental health hospitalizations for pediatrics represented 9.1% of all hospitalizations for children ages >2 in 2009, with psychosis the third most common mental health diagnosis (12.1%), after depression (44.1%) and bipolar disorder (18.1%).1 A significant increase in the diagnosis of psychotic disorders from 8.3 to 12.0 percent of hospital discharges was found in a national survey of inpatient mental health services for children and adolescents from 1999 to 2000.2 Specific predictors of poor long term outcomes include more than two inpatient-treated episodes of schizophrenia3 and a longer duration of first inpatient treatment.3 Lay et al.3 found that 12 years after their initial diagnoses of schizophrenia only 17% of adolescents had not been readmitted for further inpatient treatment, and there was a median of 4 subsequent inpatient-treated episodes. Similarly, Fleischhaker et al.4 found an average of 3 readmissions for 40% of patients in a 10-year follow-up for adolescent-onset schizophrenia.
Children and adolescents with a diagnosis of a psychotic disorder face a number of challenges medically, socially, and developmentally. Several studies found a high risk of educational and/or occupational impairment for patients with early-onset schizophrenia.3,4
A number of costs have been associated with early-onset psychosis for the medical system as well as the patient and family. Length of stay for inpatients with psychosis has been found to typically be longer than for other mental health diagnoses.5 In addition, in a comparison of mental health versus non-mental health ED visits from 2001-2008, patients with a mental health diagnosis had fewer referrals to outpatient care5 and a higher number of inpatient admissions.5 Long-term studies of patients with early-onset psychosis

have found that as adults, most were financially dependent on family or receiving public assistance.3,4
In the proposed measure, we specifically focus on the issue of comorbid substance abuse in this population. The American Academy of Child and Adolescent Psychiatry (AACAP) recommends that youth with suspected schizophrenia require a thorough psychiatric and medical evaluation, including the assessment for common comorbid conditions, such as substance abuse or cognitive delays,6 specifying that toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out.6 Comorbid substance abuse is common in patients with psychosis7-9 and can lead to decreased access of psychiatric services,10,11 while also leading to potentially avoidable healthcare utilization.6,11 Accurately diagnosing co-morbid substance abuse, or accurately diagnosing substance abuse presenting with psychotic symptoms, is an essential first step to appropriate management, referral, and obtaining access to services to address the substance abuse.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Bardach NS, Coker TR, Zima BT, et al. Common and costly hospitalizations for pediatric mental health disorders. Pediatrics. 2014;133(4):602-609.
2. Case BG, Olfson M, Marcus SC, Siegel C. Trends in the inpatient mental health treatment of children and adolescents in US community hospitals between 1990 and 2000. Arch Gen Psychiatry. 2007;64(1):89-96.
3. Lay B, Blanz B, Hartmann M, Schmidt MH. The psychosocial outcome of adolescent-onset schizophrenia: a 12-year followup. Schizophr Bull. 2000;26(4):801-816.
4. Fleischhaker C, Schulz E, Tepper K, Martin M, Hennighausen K, Remschmidt H. Long-Term Course of Adolescent Schizophrenia. Schizophrenia Bulletin. 2005;31(3):769-780.
5. Case SD, Case BG, Olfson M, Linakis JG, Laska EM. Length of stay of pediatric mental health emergency department visits in the United States. J Am Acad Child Adolesc Psychiatry. 2011;50(11):1110-1119.
6. McClellan J, Stock S. Practice parameter for the assessment and treatment of children and adolescents with schizophrenia. Journal of the American Academy of Child & Adolescent Psychiatry. 2013;52(9):976-990.
7. Hsiao R, McClellan J. Substance abuse in early onset psychotic disorders. Journal of Dual Diagnosis. 2008;4(1):87-99.
8. Cannon TD, Cadenhead K, Cornblatt B, et al. Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America. Arch Gen Psychiatry. 2008;65(1):28-37.
9. Regier DA, Farmer ME, Rae DS, et al. Comorbidity of mental disorders with alcohol and other drug abuse: Results from the epidemiologic catchment area (eca) study. JAMA. 1990;264(19):2511-2518.
10. Dyck DG, Hendryx MS, Short RA, Voss WD, McFarlane WR. Service use among patients with schizophrenia in psychoeducational multiple-family group treatment. Psychiatr Serv. 2002;53(6):749-754.
11. Schooler NR, Keith SJ, Severe JB, et al. Relapse and rehospitalization during maintenance treatment of schizophrenia. The effects of dose reduction and family treatment. Arch Gen Psychiatry. 1997;54(5):453-463.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Behavioral Health : Alcohol, Substance Use/Abuse, Behavioral Health : Screening, Behavioral Health : Serious Mental Illness, Mental Health, Mental Health : Alcohol, Substance Use/Abuse, Mental Health : Serious Mental Illness

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Measure specifications can be found at the following URL under the heading: "Mental Health Measures":
http://www.seattlechildrens.org/research/child-health-behavior-and-development/mangione-smith-lab/measurement-tools/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PSYCHOSIS_ICD9_and_ICD10_Codes_for_Denominator_Identification_SUBMITTED-635803493103736421.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
24 month period of data, retrospectively collected. We propose using 24 months due to the low prevalence of the condition. This is the period used in the field testing of the measure.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients passing the quality measure are identified during medical record abstraction using the guidelines below. The item numbers match the "Medical Records Abstraction Tool Guidelines" under "Mental Health Measures"provided on the website in S.1. This language is also in the "Medical Records Electronic Abstraction and Scoring Tool" on the website.
11. Urine Drug Screening /Serum Alcohol Screening – [Module: Psychosis, ED care] This item applies to children and adolescents presenting with psychotic symptoms who were admitted to the marker ED. Indicate if the patient had a urine drug screen and/or serum alcohol screen while in the ED. The alcohol test will be a separate test from the drug tests. The drug test must be comprehensive in that it tests for multiple types of illicit drugs. Do NOT give credit for tests that include results of just a single drug. Drug screens commonly include tests for benzodiazepines, barbiturates, methamphetamine, cocaine, methadone, opiates, tetrahydrocannabinol, etc.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients aged =5 to =19 years-old seen in the emergency department with psychotic symptoms.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Cases are identified from hospital administrative data.

Patients aged =5-=19 years-old

Patients have at least one of the following ICD9 codes for psychosis, as a primary or secondary diagnosis: 291.3, 291.5, 292.11, 292.12, 293.81, 293.82, 295.30, 295.31, 295.32, 295.33, 295.34, 295.40, 295.41, 295.42, 294.43, 295.44, 295.70, 295.71, 295.72, 295.73, 295.74, 295.90, 295.91, 295.92, 295.93, 295.94, 296.24, 296.44, 297.1, 297.2, 297.3, 298.X
These codes were chosen by Members of the COE4CCN Mental Health Working Group (see Ad.1) co-chaired by Psychiatric Health Services Researchers Drs. Michael Murphy and Bonnie Zima.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
No patients were excluded from the target population.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1. Identify eligible population at hospital using administrative data. N=total population
Step 2. Assess patient chart for indicator status. Pass (A=1) if documentation present of urine drug testing or both urine drug testing and serum alcohol testing. Pass (B=1) if documentation present of serum alcohol testing or both urine drug testing and serum alcohol testing.
Step 3. Calculate Patient score= 100*(A+B)/2. Results=0, 50, 100
Step 4. Calculate hospital score=Sum(Patient score)/N

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation*

Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A. Given the low prevalence of the condition, the measured group is the entire population of eligible patients.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
There are two potential areas for missing data: at the level of the administrative claims, which are used for sampling patients, and during medical abstraction.

Administrative Claims
There are two data fields used to identify eligible patients, the diagnosis fields and the patient age. If either is missing the case is deleted.

Medical abstraction
Missing data in the medical abstraction stage is interpreted as the patient not meeting the metric.
Please see item 2b7.1 in the testing form for additional discussion of the handling of missing data.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data collection tool is publicly available on the website in S.1. and also attached in the Appendix materials.
Title: "Medical Record Measure Electronic Abstraction and Scoring Tool" under "Mental Health Measures"

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Emergency Medical Services/Ambulance, Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
P2_Testing_for_Tox_Testing_Attachment_2015_10_13_SUBMITTED.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: **Pediatric Psychosis**: Screening for Drugs of Abuse in the Emergency Department
Date of Submission: 9/30/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient

preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

As described in the submission form, the validity and feasibility of the COE4CCN pediatric mental health measures were evaluated by an expert panel using the RAND-University of California, Los Angeles (UCLA) modified Delphi method.[1]

Detailed measure specifications were developed for the endorsed pediatric mental health quality measures. These specifications were then used to develop an electronic excel macro data collection tool for use with medical records data. The tool has automated scoring capability and is available on the website listed in item S.1. Abstraction and scoring guidelines are provided as an appendix to this submission.

Field Testing of the Delphi Panel Endorsed Pediatric Mental Health Quality Measures

Three tertiary care children's hospitals and two community hospitals participated in the field test of the emergency department (ED) ***Pediatric Psychosis*** Mental Health quality measures. For each hospital, two research nurses were trained to use the medical record abstraction tool and the companion abstraction tool guidelines. For training purposes, the nurses abstracted several sample charts targeting psychosis. Their abstractions were compared to gold-standard abstractions previously completed by the developer of the measure specifications. Abstractors were considered fully trained when they could reliably abstract the gold-standard medical records.

Case Selection

Cases for the field test were selected using International Classification of Diseases 9th Revision Clinical Modification (ICD-9) codes for psychosis from administrative databases from each hospital for discharges occurring between January 1st,2012 and December 31st, 2013 (see Appendix for a list of ICD-9 codes used to select cases for abstraction).

The final sample goal for psychosis was a total of 100 cases selected from the two larger hospitals and 35 from the three smaller hospitals, with 25% replacement cases in order to have adequate sample after patients were excluded during the medical record abstraction phase. Because of limited sample sizes at each hospital for psychosis, all eligible patients were included in the final sample. See **Table 2b5.1** for sample sizes in each hospital.

Medical Record Abstractions

For each hospital, two trained nurse abstractors were each assigned half of the case sample for psychosis. Data for each case were entered by the nurses into the electronic Pediatric Mental Health abstraction tool and both the raw data and auto-generated measure scores were uploaded to a central research database for further analysis.

At the two larger tertiary care hospitals, each nurse abstracted Pediatric Psychosis measures from 14 additional charts that were randomly selected from the other nurse's sample to facilitate assessment of inter-rater reliability (see inter-rater reliability testing results in **2a2.3** below). The 14 charts were among a total of 60 (10% sample) pulled for inter-rater reliability testing of quality measures we developed and tested across three different mental health diagnoses (psychosis, danger to self/suicidality, and substance abuse).

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**

☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Two existing administrative datasets were used to sample patients using the ICD9 codes.

The Pediatric Health Information System (PHIS) database was used to sample the medical records from two of the children's hospitals. This is a comparative pediatric database, and includes clinical and resource utilization data for inpatient, ambulatory surgery, emergency department and observation unit patient encounters for 45 children's hospitals. (More information about PHIS is available at: https://www.childrenshospitals.org/Programs-and-Services/Data-Analytics-and-Research/Pediatric-Health-Information-System)

The hospital administrative discharge databases were used to sample the medical records from the other hospitals.

1.3. What are the dates of the data used in testing? January 1, 2012-December 31st, 2013

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Three tertiary care children's hospitals and two community hospitals were included in the field test, from Washington state, Ohio, and Minnesota. The children's hospitals were: Seattle Children's Hospital, Cincinnati

Children's Hospital, and University of Minnesota Children's Hospital; the two community hospitals were in Minnesota: Fairview Ridges Hospital and Maple Grove Hospital.
These hospitals were selected as they are all member organizations of the COE4CCN multi-stakeholder consortium of organizations that took part in the Center's measure development activities.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Table 1.6 Testing: Sociodemographic Characteristics of Patients Eligible for Measurement with Pediatric Psychosis: Screening for Drugs of Abuse in the Emergency Department **(N=257)**

	N	**%**
Child gender		
Male	150	58
Female	98	38
Missing	9	4
Child race/ethnicity		
Hispanic	3	1
White	134	52
Black	76	30
Other	32	12
Missing	12	5
Insurance type		
Public	133	52
Private	106	41
Uninsured	9	4
Missing	9	4
PMCA category*		
Non-chronic condition	55	27
Non-complex chronic condition	67	33
Complex chronic condition	80	40

* PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015).[2] Available only at 2 of the 3 participating hospitals.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

NA

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

To measure patient-level sociodemongraphic variables, we used patient gender, race, ethnicity, insurance type, and chronic disease status. These variables were derived from the administrative claims data from each participating hospital. Chronic disease status was captured using the Pediatric Medical Complexity Algorithm (PMCA), which categorizes pediatric inpatients using diagnostic ICD9 codes as having an acute medical condition only (non-chronic condition), a non-complex chronic condition, or a complex chronic condition.[2] Retrospective claims data needed to run PMCA were only available from 2 of the 5 field test hospitals.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Critical data elements used in the measure were tested for inter-rater reliability of medical record abstraction. Reliability was measured using the prevalence adjusted bias adjusted kappa (PABAK) statistic for patient eligibility for measurement, and for the patient score for the quality measure. Kappa is a statistic that captures the proportion of agreement beyond that expected by chance, that is, the *achieved* beyond-chance agreement as a portion of the *possible* beyond-chance agreement.[3] PABAK is a measure of inter-rater reliability that adjusts the magnitude of the kappa statistic to take account of the influences of high or low prevalence and of inter-rater differences in assessment of prevalence. The PABAK statistic adjusts for high or low prevalence and is what we used in our calculations of inter-rater reliability.

Performance measure score was assessed for reliability across performance sites using the intra-class correlation coefficient (ICC). The ICC assesses the ratio of between site variation and within site variation on performance. Higher ICC implies that the between site variation (signal) is higher than the within site variation (noise). ICCs were computed using STATA SE 13.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

There are two stages of medical record abstraction for which we tested inter-rater reliability for all Pediatric Mental Health Measures: patient eligibility for the measure; and patient score for the quality measure. For this measure, because there were no medical record exclusions, we did not measure patient eligibility kappas, since there were no abstractions for that stage.

The specific measure addressed in this submission was one of 6 psychosis measures included in the field test as part of the broader COE4CCN Pediatric Mental Health Measures in the Hospital Setting Project.

Across all 6 psychosis measures tested in the field, 120 records were sampled and abstracted by both nurse abstractors.
Kappa for patient score for all 6 psychosis measures (n=98 eligible patient charts):
0.62.
PABAK for patient score for all 6 psychosis measures (n=98 eligible patient charts):
0.71.

For the specific submitted measure, only a very small subset (n=4) of the randomly sampled charts were eligible. There were too few patients eligible for this measure to calculate kappa. Instead, we present the percent agreement.
Percent agreement for patient scores on the quality measure under consideration:
100%

Performance measure score:
We performed ICC testing for performance variation at the level of the hospital, since that is the intended level of measurement. However, despite adequate sample size at the patient level within each site (see Table 2b5.1 below), the number of higher level clusters in our field test is limited to the 5 participating hospitals. Future measurement across a larger number of participating hospitals will give more generalizable estimations of ICC for this measure.
Hospital-level ICC=0.42 (95%CI 0.16-0.73). N=5 hospitals

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Critical data elements: Interpretation of Kappas is generally cited as follows[3,4]: ≤0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect.
Hence, inter-rater reliability for psychosis measures was substantial. For the specific submitted measure, percent agreement was perfect.

Performance measure score: Hospital level ICC based on the five hospitals is relatively high. ICCs ≥0.10 are considered relatively high.[5] Hence, the ICCs indicate that there are meaningful between-site performance differences.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☐ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)itself*

CRITICAL DATA ELEMENTS

ICD10 CONVERSION (no testing performed)

1. Statement of intent for the selection of ICD-10 codes:
    a. **The goal is to convert this measure to a new code set, fully consistent with the intent of the original measure.**
2. Excel spreadsheet with original ICD-9 codes from the Field test and the ICD9-ICD10 conversion table is attached at S2.b
3. Description of the process used to identify ICD-10 codes, including:
    a. Experts who assisted in the process:
        i. Bonnie Zima (co-chair Mental Health Working Group, see Ad.1)
        ii. Michael Murphy (co-chair Mental Health Working Group, see Ad.1)
    b. Name of the tool used to identify/map to ICD-10 codes:
        i. Transformation was based on the Centers for Medicaid and Medicare Services Gems tool.
    c. Stakeholder input was obtained from the COE4CCN Mental Health Multi-stakeholder Working Group. See below.

Psychosis ICD9 to ICD10 Conversion: Stakeholder Comments

A) Researcher and practitioner stakeholder #1:

"Psychosis - F44.89 - I usually think of dissociative disorders and conversion as not being delusional or psychotic. They are more loss of function than hallucinations, etc. So, I am not sure that this code belongs."

Response: consultation with stakeholder #3 and then deleted this code.

B) Researcher and practitioner stakeholder #2:

"I read all the new ICD 10 dx for both psychosis and substance abuse and they all seemed appropriate. They also all seemed to correspond pretty well to their ICD 9 antecedents. I am signing off on these lists. I think that the codes make sense."

Response: none needed

C) Researcher and practitioner stakeholder #3:

"re: Psychosis - F44.89, agree with [stakeholder #1] re: conversion is a somatoform disorder. Would delete."

"re: Psychosis - F44.89, I've honestly never heard of the dx "reactive confusion" and it's not in either the DSM 5 or DSM IVR. Thus I agree with [stakeholder #1]. I also wonder whether during this exercise we are getting caught up with a more historical shift within the DSM to align with the ICD…."

Response: Deleted F44.89

D) State Medicaid office stakeholder #4:

"The mental health folks in my agency are ahead of the rest of us as they have created crosswalks that make sense for our programs. Basically the codes are being based off of the DSM-5. The DSM-5 diagnoses lists both ICD-9 and ICD-10 codes with the diagnoses."

Response: Because we went through the DSM for psychosis and chose specific ICD9s for the field testing, and there is a consistent 1:1 match with ICD9 and ICD10, we decided to keep the crosswalk for ICD9-ICD10 for psychosis.

PERFORMANCE MEASURE SCORE

EMPIRICAL VALIDITY TESTING

We did not validate this measure empirically against another measure or health outcome, due to consensus of the COE4CCN Mental Health Working Group that this is a measure of technical quality and is only one of many factors expected to ultimately influence outcomes. This measure focuses on accurate diagnosis and assessment of comorbidities which should result in more appropriate treatment and ultimately lead to beneficial changes in utilization or other directly measurable effects on health outcomes. That said, by itself, the measure was judged to be too narrow and distal from such outcomes to hypothesize a direct effect that might be tested.

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—The RAND-UCLA Modified Delphi Method

The content validity of the group of quality measures developed in the COE4CCN Pediatric Mental Health measures effort, which included the psychosis measure proposed, was established using the RAND-UCLA Modified Delphi Method. The process began with the nomination of 10 individuals by 8 stakeholder organizations including the American Academy of Child and Adolescent Psychiatry, the AAP Committee on Pediatric Emergency Medicine, the AAP Task Force on Mental Health, the Medicaid Medical Directors Learning Network, the AAP Section on Hospitalist Medicine, Family Voices, the Society for Adolescent Medicine, and the Substance Abuse and Mental Health Services Administration. Nine of the nominees agreed to be members of our multi-stakeholder Delphi panel. All panelists were people deemed by the nominating organizations to have substantial expertise and/or experience related to child mental health (see Ad.1 for a list of panel members). The panel read the psychosis literature review written by project staff and reviewed and scored each proposed quality measure on validity. This method is a well-established, structured approach to measure evaluation that involves two rounds of independent panel member scoring, with group discussion in between.[1] After reviewing literature review and draft psychosis quality measures, panel members were asked to rate each measure's validity on a scale from 1 (low) to 9 (high). Validity was assessed by considering whether there was adequate scientific evidence or expert consensus to support its link to better outcomes; whether there would be health benefits associated with receiving measure-specified care; whether they would consider providers who adhere more consistently to the quality measure to be providing higher quality care; and whether adherence to the measure is under the control of health care providers and/or systems. The Delphi method has been found to be reliable and to have content, construct and predictive validity.[6-10] For a quality measure or measure component to move to the next stage of measure development, it had to have a median validity score $\geq$ 7 (1-9 scale) and be scored without disagreement based on the mean absolute deviation from the median after the second round of scoring. This process ensures that only measures widely judged to be valid moved forward into measure specification. See **Table 2b.2.3** for Delphi panel scores on the measure for this submission.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

CRITICAL DATA ELEMENTS

ICD10 CONVERSION (no testing performed)

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—The RAND-UCLA Modified Delphi Method

The scores for this measure from the 9 members of the panel after round 2 of Delphi scoring (scoring done after discussions at the in-person meeting) are presented in the Table below.

Table 2b.2.3 Testing. Delphi panel: Pediatric Psychosis: **Screening for Drugs of Abuse in the Emergency Department**

	Median score	Mean absolute deviation from median	Agreement status*
Drug Screening (Urine)			
Validity	8.0	0.8	Agree
Feasibility	9.0	0.4	Agree
Alcohol screening (serum)			
Validity	7.0	1.3	Agree
Feasibility	9.0	0.4	Agree

*This is a statistical assessment of whether panelists agreed (A), disagreed (D), or if status was indeterminate (I)

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

PERFORMANCE MEASURE SCORE
SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—DELPHI PANEL
The results from the Delphi panel show strong content validity for this measure, with median validity scores ≥7 (out of 9) following the Delphi panel.

2b3. EXCLUSIONS ANALYSIS
NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified

(*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

As noted in the Submission Item 1b, we performed a field test of the quality measure under consideration. We measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital and from two community hospitals in Minnesota, Fairview Ridges Hospital and Maple Grove Hospital. Included patients were discharged from one of the hospitals over the two year period (January 1, 2012-December 31, 2013). The performance scores are presented below in Tables 2b5.2a (performance variation across hospitals) and 2b5.2b (performance variation across socio-demographic characteristics)**.** We tested the difference in performance across the hospitals using an omnibus test for difference, and then performing individual comparisons between each hospital's performance and the mean of all other hospitals. We used ANOVA testing (4df) for the omnibus test, and a t-test to assess for individual comparisons between each hospital and the mean of all others.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Table 2b5.2a. Performance Scores for Pediatric Psychosis: Screening for Drugs of Abuse in the Emergency Department

	Eligible patients	Hospital-level Score, Mean (95% CI)	P-value for omnibus test*	Difference from mean of all others	P-value for difference from overall mean**
Hospitals overall	257	28.8 (24.5-33.1)	<0.0001	--	--
Hospital A	36	25.0 (14.7-35.3)	--	-4.4	0.48
Hospital B	166	17.8 (14.1-21.4)	--	-31.1	<0.0001
Hospital C	18	83.3 (66.3-100.4)	--	58.6	<0.0001
Hospital D	22	65.9 (47.3-84.5)	--	40.6	<0.0001
Hospital E	15	40.0 (18.6-61.4)	--	11.9	0.20

*Testing performed using ANOVA (4df)
**Testing performed using t-test

Table 2b5.2b. Socio-Demographic Group Scores for Pediatric Psychosis: Screening for Drugs of Abuse in the Emergency Department

	N	Mean	SD	Difference	LCL	UCL
Child gender						
Female (ref)	98	27.0	32.2			
Male	150	29.3	36.3	2.3	-6.6	11.2
Child race/ethnicity						
White (ref)	134	28.0	34.4			

Hispanic	3	16.7	28.9	-11.3	-50.0	27.4
Black	76	21.7	28.7	-6.3	-15.8	3.2
Other	32	40.6	41.0	12.6	-0.4	25.7
Insurance type						
Private (ref)	106	30.7	36.9			
Public/uninsured	142	26.8	33.0	-3.9	-12.7	4.9
PMCA category**						
Non-chronic (ref)	55	15.5	25.2			
Non-complex chronic	67	24.6	26.6	9.2*	0.1	18.2
Complex chronic	80	16.9	23.8	1.4	-7.3	10.1

*p<0.05. Differences tested using linear regression.
**PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015). Includes data from 2 children's hospitals only

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

For this pilot test assessing for existing variation in this measure across more than one site, we found that we were able to detect statistically and clinically meaningful differences in hospital performance. Additional information from implementation of the measure at a larger scale, as described in Section 4.1, will assist in assessing variation across a larger group of hospitals.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Missing data likely does not contribute to substantially or meaningfully biased estimates of performance for this quality measure.

There are two potential areas for missing data: at the level of the administrative claims, which are used for sampling patients, and in the medical abstraction stage.

Administrative Claims
There are two data fields used to identify eligible patients, the diagnosis fields and the patient age. Patient age is generally considered a reliable field and has minimal missing data.

A primary diagnosis is required for billing, and therefore also is rarely missing. It is known that some providers under-code for mental health diagnoses, which would lead to a risk of under recognition of eligible cases. This may lead to difficulty in capturing reliable estimates of performance at each hospital site, but is less likely to lead to biased estimates.

Medical abstraction
Missing data in the medical abstraction stage is interpreted as the patient not meeting the metric. It would be very unusual for a laboratory test (urine or serum) to be sent, processed, and not documented, due to regulation around laboratory reporting and quality assurance, as well as the financial imperative to bill and be reimbursed for the testing. Hence, we believe it is reasonable to assume that if these data elements are missing from the health record, then the process of care was not performed. Such cases are scored as not having passed the quality measure. It is unlikely that there is a substantial incidence of false negatives due to missing data, or of biased estimates due to differentially missing data.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

It was not possible to determine how often the data described above were missing. For administrative data, if a child had a diagnosis of psychosis, but this was not coded for the encounter, there would be no way to know this other than to abstract all charts for children in the eligible age range who had ED visits during the measurement timeframe to assess the frequency with which this diagnosis is documented in the record but not coded for in billing data. This approach would not be logistically feasible. For laboratory data in medical records, we believe the true rate of missing data for tests that were actually performed would be exceedingly rare for the reasons we have outlined under section 2b7.1. There would be no way to assess whether a missing lab value, where there is no evidence in the medical record of either a lab order or test result, was secondary to not doing the test versus the order and/or test result not being recorded.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

It is unlikely that missing data contributes to substantial or meaningful biases of performance estimates.

REFERENCES

1. Brook RH. The RAND/UCLA appropriateness method. In: McCormick KA, Moore SR, Siegel RA, eds. *Clinical practice guidelines development:methodology perspectives*. Rockville, MD: Agency for Health Care Policy and Research; 1994.
2. Simon TD, Cawthon ML, Stanford S, et al. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. *Pediatrics.* 2014;133(6):e1647-1654.
3. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy.* 2005;85(3):257-268.
4. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.
5. Lyratzopoulos G, Elliott MN, Barbiere JM, et al. How can health care organizations be reliably compared?: Lessons from a national survey of patient experience. *Med Care.* 2011;49(8):724-733.
6. Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. *N Engl J Med.* 1998;338(26):1888-1896.
7. Shekelle PG, Chassin MR, Park RE. Assessing the predictive ability of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy. *Int J Technol Assess Health Care.* 1998;14(4):707-727.
8. Kravitz RL, Park RE, Kahan JP. Measuring the clinical consistency of panelists' appropriateness ratings: the case of coronary artery bypass surgery. *Health Policy.* 1997;42(2):135-143.
9. Hemingway H, Crook AM, Feder G, et al. Underuse of coronary revascularization procedures in patients considered appropriate candidates for revascularization. *N Engl J Med.* 2001;344 (9):645-654.
10. Selby JV, Fireman BH, Lundstrom RJ, et al. Variation among hospitals in coronaryangiography practices and outcomes after myocardial infarction in a large health maintenance organization. *N Engl J Med.* 1996;335(25):1888-1896.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

In development of measure specifications using sample records from the field test hospitals, we found that it was important to specify the types of laboratory tests that might be sent to test for alcohol and drugs. We document this in the data collection tool for review during abstraction, using the following language:

"Indicate if the patient had a urine or serum toxicology screen for alcohol and drugs. The alcohol test will be a separate test from the drug tests. The drug test must be comprehensive in that it tests for multiple types of illicit drugs. Do NOT give credit for tests that include results of just a single drug. Drug screens commonly include tests for benzodiazepines, barbiturates, methamphetamine, cocaine, methadone, opiates, tetrahydrocannabinol, etc."

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

No proprietary elements are used in implementing this measure. There are no licenses or fees or other requirements needed to use any aspect of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This measure is part of a set of mental health quality measures the COE4CCN developed as part of the Pediatric Quality Measurement Program, funded by AHRQ, using CHIPRA monies. It has not yet been implemented as the development, validation, and testing were just recently completed. The tools needed to abstract the measures are publicly available and non-proprietary, so interested parties can implement them at any time.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Children's Hospital Association (CHA) has had representation on the National Advisory Board for COECCN since its inception. CHA has shown great interest in promoting the adoption of inpatient and ED-based measures developed by our Center. The intended audience would be hospital administrators at CHA member hospitals. We would intend to work with CHA to implement these measures over the next several years.

We also intend to publish the development and field testing of these measures in peer reviewed pediatric journals over the next 12

months. Within these publications we will include the URL where the measure data abstraction tool, measure specifications, and abstractor training materials are housed promoting further access to and dissemination of the measures.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

Credible rationale

The overall goal behind capturing performance results for this measure is to optimize appropriate diagnosis in a high-risk population. The danger of misdiagnosis is two-fold. On the one hand, patients with mental illness have a high incidence of co-morbid substance abuse disorders; on the other hand intoxication with drugs of abuse or alcohol, or a mixture, may present as psychotic symptoms. Treatment of psychosis without additionally treating co-morbid substance abuse can contribute to delayed and forgone treatment for a serious mental illness. Preventing this delayed or forgone treatment has the potential to improve care and long-term outcomes for a vulnerable population, given the evidence that earlier treatment can ameliorate the severity of illness for early onset schizophrenia (see Evidence form).

As experience has borne out, quality measurement efforts can drive improvements in care, whether through increasing focus on an area of care in internal audit and feedback efforts, or through reputational or financial incentive programs (e.g., CMS' public reporting or value-based purchasing programs). We anticipate that the performance results for this measure would drive improvement through similar mechanisms.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

There were no unintended consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** P2_Screen_for_Tox_Appendix_FOR_SUBMISSION-635803523158179295.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Seattle Children´s Research Institute
Co.2 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-
Co.3 **Measure Developer if different from Measure Steward:** Seattle Children´s Research Institute
Co.4 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The COE4CCN convened two expert groups to assist in the development of the Pediatric Mental Health Measures in the Hospital Setting--the Mental Health Working Group within the COE4CCN and an external panel of experts for the Delphi panel. Please see descriptions of the groups' roles in development as well as member names listed below.
I. Mental Health Working Group: This was a group of pediatric mental health and general pediatrics experts, as well as state Medicaid leadership. Reviewed secondary database analyses of prevalence of common and costly mental health diagnoses. Developed ICD9 code definitions to identify diagnoses of interest. Reviewed and edited the literature reviews conducted by COE4CCN staff. Provided content expertise during development of the detailed measure specifications and data abstraction tool. Participated in the planning and implementation of the field test as well as interpretation of the field test results.

Members of the MHWG:

Naomi S. Bardach, MD, MAS
Assistant Professor of Pediatrics and Health Policy
Department of Pediatrics
Philip R. Lee Institute of Health Policy
University of California San Francisco

Tumaini Ruker Coker, MD, MBA
Assistant Professor of Pediatrics
David Geffen School of Medicine
University of California, Los Angeles
Associate Natural Scientist
RAND, Santa Monica

Glenace Edwall, PsyD, PhD, MPP
Director, Children's Mental Health Division
Minnesota State Health Access Data Assistance Center
Minnesota Department of Human Services

Penny Knapp, MD
Professor Emeritus
Departments of Psychiatry & Pediatrics
University of California Davis

Rita Mangione-Smith, MD, MPH
Professor and Chief | Division of General Pediatrics and Hospital Medicine
University of Washington Department of Pediatrics
Director | Quality of Care Research Fellowship
UW Department of Pediatrics and Seattle Children's Hospital
Investigator | Center for Child Health, Behavior, and Development
Seattle Children´s Research Institute

Michael Murphy, EdD
Associate Professor
Department of Psychology
Harvard Medical School
Staff Psychologist
Department of Child Psychiatry
Massachusetts General Hospital

Laura Marie Prager, MD
Associate Professor of Psychiatry
Department of Child Psychiatry
Massachusetts General Hospital

Laura Richardson, MD, MPH
Professor
Department of Pediatrics and Psychiatry
Division of Adolescent Medicine
University of Washington
Investigator
Center for Child Health, Behavior, and Development
Seattle Children's Research Institute

Bonnie Zima, MD, MPH
Professor-in-Residence
Department of Psychiatry
University of California, Los Angeles
Associate Director
UCLA Health Services Research Center

Delphi panel: Reviewed the literature review and secondary database analyses as prepared by the MHWG and COE staff. Reviewed suggested indicators for face validity and content validity based on the above materials and based on member expertise in the field.

Members of the Delphi panel:

Gary Blau, PhD
Chief, Child, Adolescent and Family Branch,
Center for Mental Health Services (CMHS),
Substance Abuse and Mental Health Services Administration (SAMHSA), Rockville, MD.
Clinical Faculty, Yale Child Study Center, Yale University

Regina Bussing, MD, MSHS
Professor, Division of Child and Adolescent Psychiatry,
Department of Psychiatry, Department of Pediatrics,
and Department of Clinical and Health Psychology,
University of Florida, Gainesville, FL
Director, Florida Outreach Project for Children and Young Adults Who Are Deaf-Blind

Thomas Chun, MD, MPH
Associate Professor, Departments of Emergency Medicine and Pediatrics
Assistant Dean of Admissions
Chair, Admissions Committee
The Alpert Medical School, Brown University
Medical Staff, Department of Pediatric Emergency Medicine
Hasbro Children's Hospital

Sean Ervin, MD, PhD
Assistant Professor in Pediatrics & General Internal Medicine
Hospitalist Medicine
Head of Section- Pediatric Hospital Medicine
Wake Forest University, School of Medicine
Winston-Salem, NC

Doris Lotz, MD, MPH
Medicaid Medical Director
New Hampshire Department of Health and Human Services
Office of Medicaid Business and Policy
Instructor, Geisel School of Medicine at Dartmouth, Department of Psychiatry

Lynn Pedraza, PhD
Executive Director of Family Voices,
Albuquerque, NM

Karen Pierce, MD, DLFAPA, DLFAACAP
Clinical Associate Professor, The Feinberg School of Medicine,
Northwestern University Medical School,
Department of Psychiatry and Behavioral Sciences, Chicago, IL,
President, Illinois Academy of Child Psychiatry

Robert Sege, MD, PhD, FAAP
Professor of Pediatrics, Boston University School of Medicine
Director, Division of Family and Child Advocacy, Boston Medical Center
Core Faculty, Harvard Injury Control Research Center
Core Faculty, Harvard Youth Violence Prevention Center

Gail Slap, MD, MSc
Professor of Pediatrics, Department of Pediatrics,
Professor of Medicine, Department of Medicine,
University of Pennsylvania School of Medicine

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:**

Ad.3 **Month and Year of most recent revision:**

Ad.4 **What is your frequency for review/update of this measure?**

Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

PSYCHOSIS

Note: There are a number of ICD9 codes that have mapped to the same ICD10 code, and one ICD9 code that mapped to 2 ICD10 codes

ICD9 used in Field test	ICD9 label	ICD10 conversion from CMS GEMS tool	ICD10 label
291.3	alcoh psy dis w hallucin	F10.951	Alcohol use, unspecified with alcohol-induced psychotic disorder with hallucinations
291.5	alcoh psych dis w delus	F10.950	Alcohol use, unspecified with alcohol-induced psychotic disorder with delusions
292.11	drug psych disor w delus	F19.950	Other psychoactive substance use, unspecified with psychoactive substance-induced psychotic disorder with delusions
292.12	drug psy dis w hallucin	F19.951	Other psychoactive substance use, unspecified with psychoactive substance-induced psychotic disorder with hallucinations
293.81	psy dis w delus oth dis	F06.2	Psychotic disorder with delusions due to known physiological condition
293.82	psy dis w halluc oth dis	F06.0	Psychotic disorder with hallucinations due to known physiological condition
295.3	paranoid schizo-unspec	F20.0	Paranoid schizophrenia
295.31	paranoid schizo-subchr	F20.0	Paranoid schizophrenia
295.32	paranoid schizo-chronic	F20.0	Paranoid schizophrenia
295.33	paran schizo-subchr/exac	F20.0	Paranoid schizophrenia
295.34	paran schizo-chr/exacerb	F20.0	Paranoid schizophrenia
295.4	schizophreniform dis nos	F20.81	Schizophreniform disorder
295.41	schizophrenic dis-subchr	F20.81	Schizophreniform disorder
295.42	schizophren dis-chronic	F20.81	Schizophreniform disorder
295.43	schizo dis-subchr/exacer	F20.81	Schizophreniform disorder
295.44	schizophr dis-chr/exacer	F20.81	Schizophreniform disorder
295.7	schizoaffective dis nos	F25.9	Schizoaffective disorder, unspecified
295.71	schizoaffectv dis-subchr	F25.9	Schizoaffective disorder, unspecified
295.72	schizoaffective dis-chr	F25.9	Schizoaffective disorder, unspecified
295.73	schizoaff dis-subch/exac	F25.9	Schizoaffective disorder, unspecified
295.74	schizoafftv dis-chr/exac	F25.9	Schizoaffective disorder, unspecified
295.9	schizophrenia nos-unspec	F20.9	Schizophrenia, unspecified
295.91	schizophrenia nos-subchr	F20.9	Schizophrenia, unspecified
295.92	schizophrenia nos-chr	F20.9	Schizophrenia, unspecified
295.93	schizo nos-subchr/exacer	F20.9	Schizophrenia, unspecified
295.94	schizo nos-chr/exacerb	F20.9	Schizophrenia, unspecified
296.24	depr psychos-sev w psych	F32.3	Major depressive disorder, single episode, severe with psychotic features
296.44	bipol i manic-sev w psy	F31.2	Bipolar disorder, current episode manic severe with psychotic features
297.1	delusional disorder	F22	Delusional disorders

297.2	paraphrenia	F22	Delusional disorders
297.3	shared psychotic disord	F22	Delusional disorders
298.0	react depress psychosis	F32.3	Major depressive disorder, single episode, severe with psychotic features (Note: This is a duplicate, with two ICD10 codes for one ICD9)
298.0	react depress psychosis	F33.3	Major depressive disorder, recurrent, severe with psychotic symptoms (Note: This is a duplicate, with two ICD10 codes for one ICD9)
298.1	excitativ type psychosis	F28	Other psychotic disorder not due to a substance or known physiological condition
298.3	acute paranoid reaction	F23	Brief psychotic disorder
298.4	psychogen paranoid psych	F23	Brief psychotic disorder
298.8	react psychosis nec/nos	F23	Brief psychotic disorder
298.9	psychosis nos	F29	Unspecified psychosis not due to a substance or known physiological condition

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2807
Measure Title: Pediatric Danger to Self: Discharge Communication with Outpatient Provider
Measure Steward: Seattle Children´s Research Institute
Brief Description of Measure: Percentage of children/adolescents age >=5 to <=19 years-old admitted to the hospital with dangerous self-harm or suicidality, should have documentation in the hospital record of discussion between the hospital provider and the patient´s outpatient provider regarding the plan for follow-up (discussion can be by phone or email).
Developer Rationale: In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.

The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of quality measures related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is designed to address part of this key measurement gap.

The COE4CCN Mental Health Working Group (see item Ad.1) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group (Bardach et al. Pediatrics 2014). We found that depression was the most common reason for pediatric mental health hospitalizations, followed by bipolar disorder and psychosis. Danger to self and suicidality is cross-cutting across all three diseases, and, as an indicator of severity of illness, is crucial to address with high quality care in the ED and inpatient settings. Thus, we chose to focus on this as an area of measurement. See Evidence form for conceptual model underlying the rationale for the measure. A literature review was then conducted on this topic area as part of the COE4CCN work.

Based on this literature review, we developed a suggested list of draft quality measures to assess the quality of pediatric mental health care in the hospital setting for children with suicidality, or "danger to self". The validity and feasibility of these draft quality measures were then evaluated by a multi-stakeholder panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in hospitals in Washington state, Ohio, and Minnesota. This proposal presents the results of this development and validation work.

The rationale for the focus of this measure is supported by more general evidence related to hospital-to-home transitions indicating that "warm hand-offs" between inpatient and outpatient providers are associated with higher attendance at planned follow-up visits, fewer readmissions, and fewer return ED visits (Desai et al. 2015; see Item 1c.3-1c.4 below and Evidence form for additional evidence). The goal of this measure is to maximize the chances that the child/adolescent will get needed outpatient mental healthcare and will be less likely to contemplate or attempt suicide again.

Numerator Statement: Children/adolescents admitted to the hospital for dangerous self-harm or suicidality should have documentation in the hospital record of discussion between the hospital provider and the patient´s outpatient provider regarding the plan for follow-up (discussion can be by phone or email) prior to discharge.
Denominator Statement: Patients aged >=5 to <=19 years-old admitted to the hospital with a discharge diagnosis of danger to self or suicidality.
Denominator Exclusions: Patients are excluded if they are transferred to an acute or non-acute inpatient facility, left against medical

advice (AMA) or eloped. They are also excluded if the hospital provider is also the post-discharge provider or post-discharge follow-up is arranged to occur at the marker hospital's own outpatient psychiatric clinic.

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this process measure:

- The developer links measuring processes of care to reducing re-presentation with danger to self/suicidality. Thus, evidence for this process should demonstrate that improved communication will ensure continued access to needed treatment for severely ill patients, which leads to the desired outcomes of improved adherence to care and reduced risk of recurrence of active suicidal or self-harm behavior.
- The measure derives from a guideline of the American Academy of Child & Adolescent Psychiatry, which in turn relies on a recommendation from the National Institute for Health and Care Excellence (NICE). The developer reported that there were no cited trials to support the recommendation and that the cited recommendation was made as an expert consensus statement, not one that assessed the quantity, quality, and consistency of evidence. The grade assigned was "Very low quality: We are very uncertain about the estimate."
- The developer also conducted its own literature review examining processes and structures of care related to transitions between sites of care, generally. The developer provided information on two studies that focused on the communication between inpatient and outpatient providers, generally, that demonstrated improved outcomes; the developer notes bundled interventions were assessed, not the single intervention of discussion between the hospital provider and the patient´s outpatient provider regarding the plan for follow-up, this measure's focus. The developer reported that both studies met quality of evidence ratings of 1 (systematic review) based on the OCEBM 2011 Levels of Evidence.
- For the NICE recommendation, per NQF's Evidence Algorithm this is process measure with a systematic review based on expert opinion (box 3→box 7). For the developer's additional evidence review, the developer cites two studies that were systematic reviews (graded) but were general, not precisely, to the measure's focus (i.e., bundled interventions and not specific to the pediatric population or patients at risk of dangerous self-harm or suicidality). The developer also states it graded all applicable references, but details on the quantity, quality, and consistency for the cited studies and the additional literature are not provided.

Questions for the Committee

- *Is the relationship of this process measure to patient outcomes clear? If so, how strong is the evidence for this relationship (i.e., directly applicable to the process of care being measured)?*
- *Given the information submitted, how should the developer's literature review be assessed—as a systematic review or as empirical evidence submitted (box 2 vs box 7)? Or, is the INSUFFICIENT WITH EXCEPTION rating*

appropriate because there is no literature specific to the measure (i.e., is related, but not specific) but the Committee believes it is beneficial to hold providers accountable for performance in the absence of specific empirical evidence? Alternatively, the Committee may judge that INSUFFICIENT information has been provided about the literature review.

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer provided performance results for this measure using data aggregated (N=177) over two years from three children's hospitals (i.e., Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital). The mean performance score was 20.5% across the three children's hospital.
- The developer noted that no statistically significant differences were found in disparities data for population group (i.e., gender, race, insurance type, and chronic disease category).

Question for the Committee

- *Is there a gap in care that warrants a national performance measure?*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- Very low quality evidence. Note that a systematic review of low quality evidence does not constitute high quality evidence. This is a measure of something that seems like it should improve care but there is little evidence to support it.
- While the literature cited was more general than this specific measure focus, the concept of communication between the outpatient and inpatient treating healthcare personnel has improved care in other realms. The likelihood of harm is small although it is not clear that this will prevent suicide or suicidal thoughts. Coordination of care to the outpatient world is likely better than no coordination and the potential of being lost to follow-up. Access to mental health services is one of the approaches to teen suicide prevention.
- The evidence support transition bundles of which communication is only 1 part. Evidence does not seem to suggest impact on utilization of emergency or inpatient services. Transfer of information does not necessarily mean that the receiver is acting appropriately.
- Measure was generated from a guideline of the American Academy of Child & Adolescent Psychiatry, which in turn relies on a recommendation from the National Institute for Health and Care Excellence (NICE). The developer reported that there were no cited trials to support the recommendation and that the cited recommendation was made as an expert consensus statement, not one that assessed the quantity, quality, and consistency of evidence. The grade assigned was "Very low quality".
- Process measure has evidence is low quality and systematic review bundled this process with several others. The "warm handoff" is insufficiently studied.
- Two more general studies were found that focused on the communication between inpatient and outpatient providers (not specific to the measure), generally, that demonstrated improved outcomes; both studies met quality of evidence ratings of 1 (systematic review) based on the OCEBM 2011 Levels of Evidence.

1b. Performance Gap.

- Extremely low performance - which suggests either that this is not something that is perceived as an important measure of care quality OR that the important parameter (communication) is being accomplished in some other way.
- A rate of 20% for communication between the inpatient healthcare provider and the outpatient healthcare provider that will be taking over care is really low and would give a lot of room for improvement. The N is small, but the success rate is really low.
- Not certain about gaps in care--the adage, "if it's not documented it's not done" may apply legally, but not necessarily in real life.

- There is a performance gap , but no difference between populations
- No statistically significant differences were found in disparities data for population group. Performance results were provided for this measure using data aggregated (N=177) over two years from three children's hospitals. The mean performance score was 20.5% across the three children's hospital.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

The developer provided the following information:

- This measure is specified at the facility level of analysis. It uses data from administrative claims, electronic health records, and/or paper medical records.
- This measure captures children ages 5-19 years admitted to the hospital with dangerous self-harm or suicidality that should have documentation in the hospital record of discussion between the hospital provider and the patients outpatient provider regarding the plan for follow up. Patients were excluded if they are transferred to an acute or non-acute inpatient facility, left against medical advice, and if the hospital provider is also the post-discharge provider or post-discharge follow-up is arranged at the hospital's outpatient psychiatric clinic.
- ICD-9 and ICD-10 codes to identify patients admitted to the hospital for dangerous self-harm or suicidality (the denominator) are provided. The developer notes in the supplement excel file that the ICD-10 codes were identified through use of the ICD-10-CM Table of Drugs and Chemicals located at: http://www.tacomacc.edu/UserFiles/Servers/Server_6/File/him//HIM240/ICD10Cmcodebook/icd10cm_drug_2011.pdf.
- Information needed for the numerator must be obtained from the medical record. Patients admitted with dangerous self-harm or suicidality are identified during medical record abstraction using the Mental Health Medical Records Measures Data Abstraction Tool Guidelines and Medical Records Measures Scoring Specifications. The developer provided these guidelines and scoring specifications in the appendices for this measure.
- The calculation algorithm is provided. In this algorithm, the developer suggests using 100*n (the numerator population) divided by N (hospital's eligible target denominator population using administrative claims data) minus e (the patients excluded based on medical record abstraction) to calculate the score (100*n/(N-e)).
- This measure is not stratified or risk-adjusted.

Questions for the Committee

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer provides the following information:

- Empirical reliability testing for this measure was conducted at the data element level for the measure as specified.
- The developer tested the measure using data from 117 medical records obtained from three children's hospitals. The timeframe for testing was January 1, 2012-December 31, 2013. The developer states that testing was conducted using

administrative claims, paper records, and electronic records and notes that the Pediatric Health Information System (PHIS) database was used to sample the medical records from two of the children's hospital and the hospital administrative discharge database was used to sample the medical records from the third.

- Reliability testing was conducted both at the critical data element level (i.e., inter-rater reliability) and performance score level (i.e., intra-class correlation coefficient).
- At the critical data element level, reliability was assessed on 40 charts using the prevalence adjusted bias adjusted kappa (PABAK) statistics for *patient eligibility* for measurement and the *patient score* for the quality measure.
    - Results for the IRR for assessment of patient eligibility were Kappa=0.80; PABAK=0.85. The developer notes this is considered generally perfect.
    - The sample of cases was too small to calculate a Kappa or results for the patient score. The developer instead provided the percent agreement between abstractors regarding patient score for this measure, which was 88%.
- For reliability at the computed performance measure score, the developer performed ICC testing at the hospital level (the intended Level of Analysis).
    - The ICC for N=3 hospitals was 0.34 (95%CI 0.03-0.92).
    - The developer notes the ICC is "relatively high" (based on the literature), and hence there are meaningful performance differences between the sites.
- Per the NQF Algorithm for Reliability, testing was conducted at both the computed performance score and the patient-level data elements and so the rating options are HIGH, MODERATE, LOW, INSUFFICIENT.

Questions for the Committee

- *Is the test sample for data element reliability adequate to generalize for widespread implementation? MODERATE, LOW, OR INSUFFICIENT rating?*
- *Are the methodology and test sample for computed performance score reliability adequate to generalize for widespread implementation? HIGH, MODERATE, LOW, OR INSUFFICIENT?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The measure specifications are related, but not specific to, the evidence presented in criteria 1a.

Question for the Committee

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- The developer conducted both empirical validity testing and systematic assessment of face validity of performance measure score for this measure.
- Empirical validity testing was used to assess the quality measure and the validation metrics (i.e., 30-day readmissions and 30-day ED revisits).
    - There were no statistically significant differences between those meeting and those failing the measure in readmissions (OR=1.00) and ED revisits (OR=1.01).
    - The developers note the relatively low sample size of eligible patients may have led to limited power to demonstrate a difference in readmission or ED return visits for patients passing versus failing this quality measure.
- The developer performed systematic face validity assessment (RAND-UCLA Modified Delphi) of "whether panelists would consider providers who adhere more consistently to the quality measure to be providing higher quality care,"

which we interpret as face validity assessment at the level of the **computed measure score** (as required by NQF).

- Per the NQF Algorithm for Validity, for empirical testing at the level of computed performance score, the highest rating option is HIGH. For face validity at the measure score level, the highest rating option is MODERATE.

Questions for the Committee

- *Do the results from empirical testing demonstrate sufficient validity so that conclusions about quality can be made? If not, do the results from the face validity assessment demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer states in its testing document that there are no exclusions. Accordingly, the developer does not address the required information related to the method of testing the impact of the exclusions, the statistical results from such testing, nor the interpretation of the results.
- Elsewhere, however, the developer notes: Patients are excluded if they are transferred to an acute or non-acute inpatient facility, left against medical advice (AMA) or eloped. They are also excluded if the hospital provider is also the post-discharge provider or post-discharge follow-up is arranged to occur at the marker hospital's own outpatient psychiatric clinic.

Questions for the Committee

- *Are the exclusions described (though not on the testing form) appropriate? If so, the Committee should request additional information required by the NQF guidance.*
- *Are the exclusions consistent with the evidence?*
- *Are the exclusions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment:

- This measure is not risk-adjusted.

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provides the following information:

- Omnibus testing for difference in performance across all three hospitals found statistically significant differences among the three test sites, with a p-value of 0.002.

Question for the Committee

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

- The developer states that there are two potential areas for missing data: a t the level of the administrative claims (i.e., and in the medical abstraction stage.
- The developer posits that the claims data are unlikely to be a source of missing data.
- The developer notes that at the medical abstraction state, missing data are interpreted as the provider not meeting the metric and that, to the degree that the metric is being met but not being documented (false negative performance), performance measurement will stimulate improved documentation.

Questions for the Committee

- *Does the Committee concur with the developers' conclusions regarding missing data?*
- *Does the Committee view the potential for false negative performance significant? Is this a threat to the measure's validity (i.e., to be an indicator of quality)?*

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- Exclusions seem to be large in number, limiting any effect that this measure might have. The actual measure numerator is for a 24 hour period that ends 24 hours before actual discharge - noting that time of discharge varies for many different reasons. This seems arbitrary and possibly not useful. Communication that occurs shortly before discharge might be even more useful.
- The sample is small and the reliability data could have been generated with the entire sample without a lot more effort in order to provide more data.
- Documentation--who is responsible for forwarding information/plan to next provider? Is this different at different sites? If not MD, then who is responsible for documenting?
- I am concerned that this measure can be consistently implemented.
- Data is pulled from administrative claims, electronic health records, and/or paper medical records. Relevant ICD 9 and ICD 10 codes are provided. Patients admitted with dangerous self-harm or suicidality are identified during medical record abstraction using the Mental Health Medical Records Measures Data Abstraction Tool Guidelines and Medical Records Measures Scoring Specifications. The developer provided these guidelines and scoring specifications in the appendices for this measure; Calculation algorithm is provided. Question whether the data will easily be extracted from medical record."

2a2. Reliability testing

- Concerns that if there is a shared EMR between the hospital and the follow up institution, an email or phone call is not needed in the same way that it is not needed if the follow up is in the same institution (an exclusion criterion)
- The data elements appear to be clear as is the population.
- Not convinced that the measure will be reliably documented
- Tested measure from 117 medical records obtained from 3 children's hospitals over 2 years. Reliability testing happened at the critical data element level (inter-rater reliability) and performance score level (intra-class correlation coefficient). Results demonstrate reliability.

2b1. Validity Specifications

- There is no evidence for it to be inconsistent with.
- While the measure is not exactly the same the concept of provider communication when handing off care in general has been shown to provide higher quality care and is probably relevant here.
- Specifications are related but not specific to the evidence.

2b2. Validity Testing

- Readmissions and ER revisits are not affected, OR 1.0 with very narrow CIs. Can confidently say improving this measure won't affect these outcomes. Also concerned with the statement that the measure might lead to better documentation, which the developer would like to have in order to validate the measure. Doesn't seem a strong reason to institute a measure.
- Readmission to the hospital and ED may be so infrequent with this group that with such a small sample it is not reflected here. Other measures of quality of care might be a better measure such as in person follow-up post discharge, evidence of continuing treatment, etc. That said communication between providers is usually considered to provide higher quality care than the absence.
- Conducted both empirical validity testing and systematic assessment of face validity of performance measure score for this measure. Small sample size, not adequate to determine validity.

2b3-2b7. Threats to Validity

- Developer is inconsistent re exclusions. Not sure why transfers to other facilities would be excluded, while discharge to Day Hospital would not. Not sure that variation in this measure would really reflect variation in

quality.

- Some of the exclusions are concerning. Even if the child is followed within the provider system of the hospital, the inpatient and outpatient practitioners are not always the same and communication may not occur unless it is deliberate. One cannot assume that it will be done by records. If a patient at risk for self-harm or suicide leaves AMA, then it might be appropriate to inform their outpatient provider so that they can contact them and make sure that they are safe. Transfer to an acute or nonacute facility should include a conversation between the sending and the receiving practitioners and that could be measured in the same manner as the conversation with the outpatient provider it is still a handoff. Documentation quality and missing documentation of phone calls might be an issue affecting rate of contact.
- Exclusions for same practitioner: group practices, different inpatient vs outpatient practitioners in the same group--not clear how these are handled.
- Two potential areas for missing data: at the level of the administrative claims and in the medical abstraction stage.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer reports:

- The data source for this measure is claims, paper medical records, or electronic medical records.
- No data elements are in defined fields in electronic sources.
- The developer notes the chart abstraction component requires use of the tool and guidelines it developed.

Questions for the Committee

- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Documentation of phone call or email, especially if the email is outside of the EHR system, seems likely to be difficult and unreliable.
- Documentation of calls related to patient care should be standard, but isn't always which is a weakness of this approach. Using a homegrown data collection tool means that it is not easily applicable to other sites.
- Could make this electronic, discreet data.
- Feasibility is low. Difficult and expensive to extract and not convinced that missing data would accurately indicate the communication did not occur
- Data source for this measure is claims, paper medical records, or electronic medical records. No data elements are in defined fields in electronic sources however there is a tool used for chart abstraction with guidelines available.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

The developer reports:

- The measure is not currently in use, however the developer described how it plans to use the measure for quality improvement and quality improvement with benchmarking. [NQF emphasizes endorsement for accountability, though measure also can be used for quality improvement.]
- The developer stated that performance results for this measure (optimize transitions of care for a high-risk population) will ultimately drive improvement.
- The developer stated that no unintended consequences were identified during testing of this measure.

Questions for the Committee

- *The developer indicates quality improvement and benchmarking to drive improvement. Will public reporting of performance results further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- This measure adds a dimension to the existing NQF follow-up for a high profile, more fragile population of children being discharged from psychiatric hospitalizations. The combination of these two measures would help connect the dots for this population for whom more frequent or soon follow-up may be appropriate.
- The measure would drive improvement of communication or documentation of communication?
- Not being used now but plans to use it for quality improvement/benchmarking. Has the potential to drive improvement.

Criterion 5: Related and Competing Measures

- 0576 : Follow-Up After Hospitalization for Mental Illness (FUH) is an NQF-endorsed measure. This NQF-endorsed measure reports two rates: percentage of discharges for which the patient received follow-up within 7 days and within 30 days of discharge. A broad range of diagnoses are included.
- Both measures focus on the transition from inpatient to outpatient care, however this new measure focuses on a narrower population (danger to self or suicidality) and different process (communication re: follow-up care).

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Pediatric Danger to Self: Discharge Communication to Outpatient Provider

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Children/adolescents admitted to the hospital with dangerous self-harm or suicidality should have documentation of a discussion between the hospital provider and the patient's outpatient provider regarding the plan for follow-up.

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

NA

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

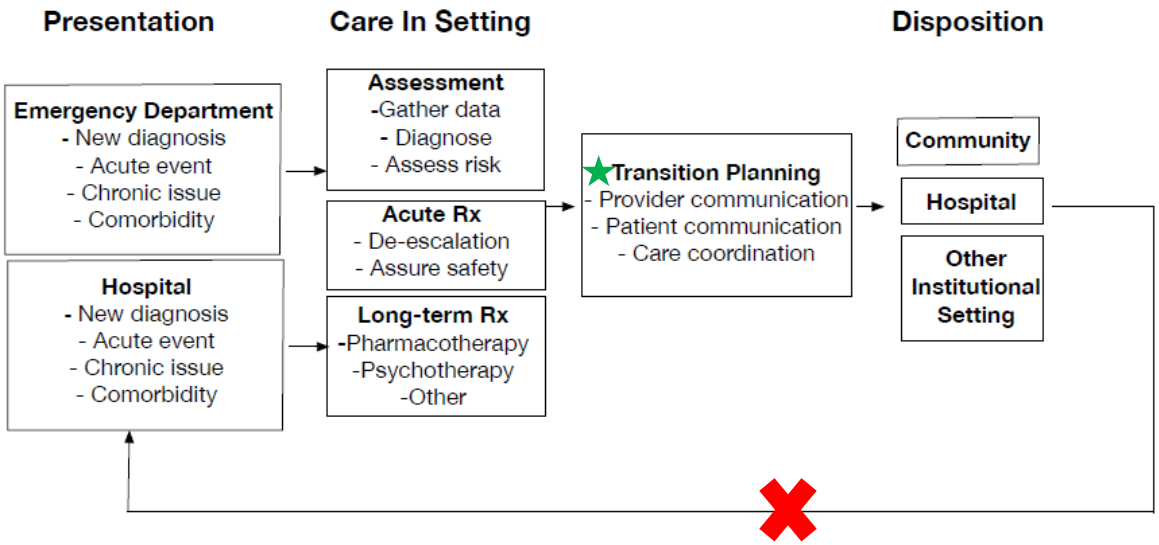
NA

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The evidence supporting this measure derives from the AACAP guidelines on treatment of suicidality.



The figure above depicts the conceptual model describing how measured processes of care might reduce re-presentation with danger to self/suicidality. The green star marks the care process that the measure assesses, which is to improve the transition back to the community through better communication and development of a shared plan of care with the outpatient provider. The red X marks the pathway back to the undesirable outcome we hypothesize would be mitigated if measure performance is optimal. We hypothesize that improved communication will ensure continued access to needed treatment for a severely ill patient. This will potentially improve their adherence to care and decrease the risk of a recurrence of active suicidal or self-harm behavior.

The evidence below supports the conceptual model. The evidence we present combines consensus opinion, from the National Institute for Health and Care Excellence (NICE) guideline on psychiatric care for adolescents and children with schizophrenia (No. 155),[1,2] with published literature on discharge transitions for chronically ill patients. We included the literature on transitions for other chronically ill populations because, though there is a large body of research on suicide and self-harm in children and adolescents , it focuses on characterizing who is at risk for self-harm and suicide, with very little systematic research on the optimal quality of care for children and adolescents who self-harm.[3-7] This is likely due to the rarity and potential morbidity of suicide, which makes evaluation methods such as randomized control trials difficult to implement and evaluate.[8,9] Suicide also can be spontaneous[10,11] and difficult to predict, with many people committing suicide without first presenting with suicidality.[12] This makes gathering data on patients prior to a suicide attempt difficult, if not impossible. However, because attempting suicide is a leading predictor of later suicide attempts,[13-16] quality measures focusing on optimal care for this population as they transition to the outpatient setting are important. The underlying assumption in extrapolating from the literature on transitions in other diagnoses is that mental health conditions associated with self-harm and suicidality are chronic illnesses, presenting acutely, similar to other illnesses in the hospital transitions literature (e.g., heart failure), in that they are chronic and require care coordination,[17,18] and so we turned to the transitions literature to inform our measure development work in this area.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Psychosis and schizophrenia in children and young people: Recognition and management URL: http://www.nice.org.uk/guidance/CG155

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

"1.3.6 Develop a care plan with the parents or carers of younger children, or jointly with the young person and their parents or carers, as soon as possible, and:

- include activities that promote physical health and social inclusion, especially education, but also employment, volunteering and other occupations such as leisure activities
- provide support to help the child or young person and their parents or carers realize the plan
- give an up-to-date written copy of the care plan to the young person and their parents or carers if the young person agrees to this; give a copy of the care plan to the parents or carers of younger children; agree a suitable time to review it
- send a copy to the primary healthcare professional who made the referral

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Strength of recommendations

Some recommendations can be made with more certainty than others. The Guideline Development Group makes a recommendation based on the trade-off between the benefits and harms of an intervention, taking into account the quality of the underpinning evidence.

The wording used in the recommendations in this guideline denotes the certainty with which the recommendation is made (the strength of the recommendation).

2. Interventions that should (or should not) be used – a 'strong' recommendation

'Offer' (and similar words such as 'refer' or 'advise') indicate that guideline authors were confident that, for the vast majority of patients, an intervention will do more good than harm, and be cost effective. Similar forms of words (for example, 'Do not offer…') indicate confidence that an intervention will not be of benefit for most patients.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1. Interventions that must (or must not) be used

The words 'must' or 'must not' are usually only used if there is a legal duty to apply the recommendation. Occasionally 'must' (or 'must not') indicates a situation in which the consequences of not following the recommendation could be extremely serious or potentially life threatening.

3. Interventions that could be used

The word 'consider' indicates confidence that an intervention will do more good than harm for most patients, and be cost effective, but other options may be similarly cost effective. The choice of intervention, and whether or not to have the intervention at all, is more likely to depend on the patient's values and preferences than for a strong recommendation, and so the healthcare professional should spend more time considering and discussing the options with the patient.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**
Same citation as for grading recommendations in 1a.4.1.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The quoted recommendation from the NICE guideline falls under the topic of: "Access to and the delivery of services, and the experience of care", and the description for this topic in the NICE guideline section titled "From Evidence to Recommendations" notes the dearth of published evidence in this area (page 108 of the guidelines cited in 1a.4.1).

The strength of the cited NICE guideline recommendation (item 1a4.3) was based on expert consensus.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

"Very low quality: We are very uncertain about the estimate."

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

● High quality: Further research is very unlikely to change our confi dence in the estimate of effect.

● Moderate quality: Further research is likely to have an important impact on our confi dence in the estimate of effect and may change the estimate.

● Low quality: Further research is very likely to have an important impact on our confi dence in the estimate of effect and is likely to change the estimate.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: May 1 2012-September 25, 2014

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

There were no cited trials to support the recommendation. After reviewing the narrative summary of the evidence, including prior NICE Guidelines regarding treatment of children and adults with schizophrenia, the cited recommendation were made as an expert consensus statement (page 100 of the cited Guideline in 1a.4.1).

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

"Very low quality: We are very uncertain about the estimate."

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

"Very low quality: We are very uncertain about the estimate."

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

No harms were studied.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

An updated evidence review for the NICE guideline cited above was published in March 2015 and did not find additional relevant studies (URL: https://www.nice.org.uk/guidance/cg155/evidence/cg155-psychosis-and-schizophrenia-in-children-and-young-people-evidence-update2).

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in addition to being tasked with mental health measure development, was tasked with developing measures related to the quality of pediatric hospital-to-home transitions.

The COE4CCN Transitions Working Group, who were charged with developing the Center's hospital-to-home transition quality measures, was a multi-stakeholder group of researchers, providers, patient advocates, payers and federal agency representatives with expertise in and/or experience with care transitions. Members of the working group conducted a targeted literature review to inform quality measure development. The literature review pertaining to this measure set reviewed studies examining processes and structures of care related to transitions between sites of care in pediatrics. The findings of this review were shared with the Mental Health Working Group as part of the larger COE4CCN effort, and select relevant articles are presented here. The full transitions literature review was published in 2015.[19] The methods used to assess the evidence for the full literature are briefly summarized below.

In order to identify peer-reviewed literature examining the relationship between transition processes of care and outcomes we conducted a search using Medline, CINAHL, Cochrane Library, Embase, Web of Science and Psych INFO. The following search terms were used in various combinations: transition(s), transitional, handoff (s), handover(s), discharge, discharge planning, transfer(s), patient discharge, early patient discharge, patient discharge education, outpatient, home, clinic, emergency room or department, hospital, inpatient, intensive care unit (ICU), rehabilitation facility, skilled nursing facility, rest home, long term care, primary care, general practitioner, and specialist. Searches were specified such that articles would not be editorials or comments. Inclusion criteria specified that articles must be written in the English language, published between 2001 and 2012, and included all ages. Though we are focused on populations between 0 and 19 years of age, we felt it was likely that much of the literature relevant to this search may be in older age populations.

In total, when duplicates were removed the search returned 3707 articles. We screened titles and abstracts and removed articles that focused only on medication reconciliation and those specific to clinical care and clinical outcomes rather than the process of transition between sites of care. Articles retained appeared to contain elements of the transition process as independent variables and measures of health or utilization as outcome variables. We also included patients' or caregivers' reported experiences of care as an outcome variable. One hundred and sixty seven articles remained for full text review. An additional 21 articles were added from the reference lists of selected articles and relevant reviews.

The strength of evidence was formally rated for each study according to the University of Oxford's Centre for Evidence-Based Medicine (OCEBM) levels of evidence.[20]

Overall findings for transitions literature review

The literature review identified two systematic reviews and 7 studies on multi-component bundles of interventions, but no studies that focused on single component interventions. One systematic review[21] examined 36 randomized controlled trials (RCTs) focused on interventions to improve patient handovers from the hospital to primary care mainly in elderly populations. Thirty-four studies included multicomponent interventions using a comprehensive model.[21] Of the 36 RCT's twenty-five (69.4%) studies had statistically significant effects in favor of the intervention group in one or more outcomes. Effective interventions included but were not limited to discharge planning, and shared involvement in follow-up by hospital and community care providers. The studies identified looked at many components of the transition process simultaneously. Therefore, it was not possible to isolate how specific aspects of the transition process were related to specific outcomes of interest including Emergency Department (ED) use, re-hospitalization, hospital bed days, length of stay, cost, completion of outpatient workups, adherence to recommended care, identifiable accountable provider, quality of life, family preparedness for discharge, patient function, unmet needs and satisfaction.

Quality of evidence was based on the OCEBM 2011 Levels of Evidence: 1 = systematic review; 2 = randomized trial; 3 = cohort study; 4 = case-control study; 5 = mechanism-based reasoning.

1a.8.2. Provide the citation and summary for each piece of evidence

Of the studies in the review described above, we focused on the literature on communication between inpatient and outpatient providers. We cite two studies below; both met quality of evidence ratings of 1 based on the OCEBM 2011 Levels of Evidence.[20] Both were conducted in adult and elderly populations. They support the specific process of care in the proposed measure, though both studies assess it as part of a bundled intervention.

22. Balaban RB, Weissman JS, Samuel PA, Woolhandler S. Redefining and redesigning hospital discharge to enhance patient care: a randomized controlled study. *J Gen Intern Med.* 2008;23(8):1228-1233.

Balaban et al[22] conducted an RCT examining the effectiveness of a 4-step discharge intervention compared to existing hospital practice and protocol. The intervention included 1) a user- friendly discharge form; 2) electronic transfer of the patient discharge form to a nurse at PCP's office; 3) telephone contact by the primary care nurse to the patient and 4) PCP review and modification of the discharge transfer form. The transfer of the patient discharge form to a nurse at the PCP office and PCP review and modification of the discharge transfer form are relevant to the proposed measure on communication between inpatient and outpatient providers for pediatric patients with danger to self or suicidality. One hundred ninety six patients enrolled in the trial. Undesirable outcomes measured included no outpatient follow-up within 21 days, readmission within 31 days, ED visit within 31 days and failure by the PCP to complete a recommended work-up. Patients in the intervention group (n=47) had fewer incomplete recommended outpatient work-ups (11.5% in the intervention group compared to 31% in the concurrent and historical controls) and higher outpatient follow-up rates when compared to concurrent (n=49) and historical controls (n=100). There were no significant differences in readmissions or ED visits within 31 days.

23. Finn KM, Heffner R, Chang Y, et al. Improving the discharge process by embedding a discharge facilitator in a resident team. *J Hosp Med.* 2011;6(9):494-500.

Finn et al.[23] conducted an RCT where a nurse practitioner was randomly assigned to one of five resident medical teams to complete discharge paperwork, arrange follow-up appointments and prescriptions, communicate discharge plans with the outpatient nurse and PCP, and answer questions from discharged patients. The component of communication of discharge plans with the outpatient nurse and PCP is relevant to the proposed measures. Patients in the intervention group were significantly more likely to have their discharge summary completed before their first follow-up appointment. There was no effect on 30-day readmissions or ED visits. Patients cared for on the intervention team had more follow-up appointments scheduled at the time of

discharge and attended the scheduled appointments within 2 weeks of discharge more often than control team patients. Though both groups of patients reported similar rates of questions after discharge, intervention group patients could better identify whom to call with questions and were more satisfied with the discharge process.

Thus, the available evidence suggests that communication of that plan to the PCP at discharge is associated with improved timeliness of care plan completion, and better attendance at follow-up visits and completion of recommended work-ups in the outpatient setting, and is not associated with changes in utilization. This evidence informed our measure development as we looked for processes of care to capture the quality of care transitions in the conceptual model in Section **1a.3**.

REFERENCES

1. Kendall T, Hollis C, Stafford M, Taylor C. Recognition and management of psychosis and schizophrenia in children and young people: summary of NICE guidance. *BMJ.* 2013;346.
2. Psychosis and schizophrenia in children and young people: Recognition and management. CG 155. 2013; http://www.nice.org.uk/guidance/cg155/resources/guidance-psychosis-and-schizophrenia-in-children-and-young-people-pdf. Accessed 2013.
3. AACAP. Practice parameter for the assessment and treatment of children and adolescents with suicidal behavior. *J. Am. Acad. Child Adolesc. Psychiatry.* 2001;40(7 SUPPLEMENT).
4. Ougrin D, Tranah T, Leigh E, Taylor L, Asarnow JR. Practitioner review: Self-harm in adolescents. *J Child Psychol Psychiatry.* 2012;53(4):337-350.
5. Hawton K, Saunders KEA, O'Connor RC. Self-harm and suicide in adolescents. *Lancet.* 2012;379(9834):2373-2382.
6. Nock MK. Future Directions for the Study of Suicide and Self-Injury. *Journal of Clinical Child & Adolescent Psychology.* 2012;41(2):255-259.
7. Newton AS, Hamm MP, Bethell J, et al. Pediatric suicide-related presentations: a systematic review of mental health care in the emergency department. *Annals of Emergency Medicine.* 2010;56(6):649-659.
8. Cohen J. Statistical approaches to suicidal risk factor analysis. *Annals of the New York Academy of Sciences.* 1986;487:37-41.
9. Goldstein RB, Black DW, Nasrallah A, Winokur G. The prediction of suicide: Sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1,906 patients with affective disorders. *Archives of General Psychiatry.* 1991;48:418-422.
10. Simon RI. Imminent suicide: The illusion of short-term prediction. *Suicide and Life-Threatening Behavior.* 2006;36(3):296-301.
11. Brown LK, Overholser J, Spirito A, Fritz GK. The correlates of planning in adolescent suicide attempts. *J Am Acad Child Adolesc Psychiatry.* 1991;30(1):95-99.
12. Isometsa ET, Heikkinen ME, Marttunen MJ, Henriksson MM, Aro HM, Lonnqvist JK. The last appointment before suicide: Is suicide intent communicated? *American Journal of Psychiatry.* 1995;152(6):919-922.
13. Tishler CL, Reiss NS, Rhodes AR. Suicidal behavior in children younger than twelve: a diagnostic challenge for emergency department personnel. *Acad Emerg Med.* 2007;14(9):810-818.
14. Kennedy SP, Baraff LJ, Suddath RL, Asarnow JR. Emergency department management of suicidal adolescents. *Annals of Emergerncy Medicine.* 2004;43(4):452-460.
15. Goldston DB, Daniel SS, Reboussin DM, Reboussin BA, Frazier PH, Kelley AE. Suicide attempts among formerly hospitalized adolescents: A prospective naturalistic study of risk during the first 5 years after discharge. *Journal of the American Academy of Child & Adolescent Psychiatry.* 1999;38(6):660-671.
16. Joiner TE, Rudd MD, Rouleau MR, Wagner KD. Parameters of suicide crises vary as a function of previous suicide attempts in youth inpatients. *Journal of the American Academy of Child & Adolescent Psychiatry.* 2000;39(7):876-880.
17. Simon TD, Cawthon ML, Stanford S, et al. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. *Pediatrics.* 2014;133(6):e1647-1654.

18. American Academy of C, Adolescent P. Practice parameter for the assessment and treatment of children and adolescents with suicidal behavior. American Academy of Child and Adolescent Psychiatry. *J Am Acad Child Adolesc Psychiatry.* 2001;40(7 Suppl):24S-51S.
19. Desai AD, Popalisky J, Simon TD, Mangione-Smith RM. The effectiveness of family-centered transition processes from hospital settings to home: a review of the literature. *Hospital pediatrics.* 2015;5(4):219-231.
20. OCEBM Levels of Evidence Working Group. Oxford Centre for Evidence-Based Medicine. The Oxford 2011 Levels of Evidence. http://www.cebm.net/index.aspx?o=5653.
21. Hesselink G, Schoonhoven L, Barach P, et al. Improving patient handovers from hospital to primary care: a systematic review. *Ann Intern Med.* 2012;157(6):417-428.
22. Balaban RB, Weissman JS, Samuel PA, Woolhandler S. Redefining and redesigning hospital discharge to enhance patient care: a randomized controlled study. *J Gen Intern Med.* 2008;23(8):1228-1233.
23. Finn KM, Heffner R, Chang Y, et al. Improving the discharge process by embedding a discharge facilitator in a resident team. *J Hosp Med.* 2011;6(9):494-500.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
DSD9_Inpt_Transition_Communication_Evidence_Form_2015_10_07.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.

The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of quality measures related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is designed to address part of this key measurement gap.

The COE4CCN Mental Health Working Group (see item Ad.1) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group (Bardach et al. Pediatrics 2014). We found that depression was the most common reason for pediatric mental health hospitalizations, followed by bipolar disorder and psychosis. Danger to self and suicidality is cross-cutting across all three diseases, and, as an indicator of severity of illness, is crucial to address with high quality care in the ED and inpatient settings. Thus, we chose to focus on this as an area of measurement. See Evidence form for conceptual model underlying the rationale for the measure. A literature review was then conducted on this topic area as part of the COE4CCN work.

Based on this literature review, we developed a suggested list of draft quality measures to assess the quality of pediatric mental health care in the hospital setting for children with suicidality, or "danger to self". The validity and feasibility of these draft quality measures were then evaluated by a multi-stakeholder panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in hospitals in Washington state, Ohio, and Minnesota. This proposal presents the results of this development and validation work.

The rationale for the focus of this measure is supported by more general evidence related to hospital-to-home transitions indicating that "warm hand-offs" between inpatient and outpatient providers are associated with higher attendance at planned follow-up visits, fewer readmissions, and fewer return ED visits (Desai et al. 2015; see Item 1c.3-1c.4 below and Evidence form for additional evidence). The goal of this measure is to maximize the chances that the child/adolescent will get needed outpatient mental healthcare and will be less likely to contemplate or attempt suicide again.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In a field test of the measure, performed as part of the funded development work, we measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital. Included patients were discharged from one of the hospitals over the two year period (January 1, 2012-

December 31, 2013). The performance scores are presented below.

# of hospitals: 3
# of patients: 177
Mean (SD): 20.5% (15.6)
Min-Max: 0%-38.0%
IQR: N/A

See Testing form, item 2b.5.2a for data on individual hospital performance.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In the field testing described above, we measured differences in performance scores by gender, race, insurance type, and chronic disease category (measured using the Pediatric Medical Complexity Algorithm—Simon et al. Pediatrics 2015).

We did not find any statistically significant differences in performance across groups. Please see Testing form, item 2b.5.2b for the results of these analyses.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
A leading cause of morbidity/mortality, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
When the COE4CCN Mental Health Working group convened, it first undertook two tasks: reviewing the existing measures in pediatric mental health; and conducting a secondary analysis of national and state-based data to identify the most common mental health diagnoses resulting in ED use and hospitalization in the pediatric age group. We found that there is a dearth of rigorously developed quality measures in pediatric mental health.1 We also found that mental health hospitalizations for pediatrics represented 9.1% of all hospitalizations for children ages >2 in 2009, with depression the most common mental health diagnosis (44.1% of pediatric mental health hospitalizations), followed by bipolar disorder (18.1%) and psychosis (12.1%).2 Mental health hospitalizations were costly, with annual national charges of $1.33 billion for depression hospitalizations alone in 2009, and a combined cost of $2.6 billion for the top three diagnoses.

As we planned the literature reviews to identify the best evidence on management of these conditions and guide measure development, we chose to focus one review on the cross-cutting diagnosis of danger to self (e.g., suicide or self-harm). We chose danger to self for several reasons: the most common ED and inpatient hospital presentation for depression is self-harm or suicidality;3 suicidality is not limited to depressed youth, hence, recommendations regarding suicidality apply more broadly than the population of pediatric patients with major depressive disorder; and suicidality and self-harm are severe manifestations of mental

illness, with a high burden to families as well as a high risk of subsequent mortality.4-7

Prevalence

According to the CDC, suicide is the third-leading cause of death among youths aged 10 to 24, with 4,600 young people taking their own lives each year.8 The estimated rate of suicide attempts among school-aged youths in the US was 8% in 2011, with almost 16% of youths reporting seriously considering suicide, and about 13% making a plan to commit suicide.9 Furthermore, in 2011 2.4% of US youths made a suicide attempt that was serious enough to be treated by a doctor or nurse,9 and about 157,000 children and adolescents presented to the ED with self-inflicted injuries.8 Rates of suicidality differ by demographic characteristics.8 Female adolescents are more likely to attempt suicide than males, but males are more likely to commit suicide. Older adolescents are more likely to commit suicide than are younger adolescents, and White youth are more likely than are Black or Latino youth to commit suicide.

A recent study examining adolescents with suicidal thoughts, plans, and attempts found that over 89% of those with suicidal thoughts, 93% of those with suicide plans, and 96% of those who attempted suicide had a diagnosable mental health disorder.10 Multiple studies have found that depression10-13 and bipolar disorder14 are significant risk factors for self-harm and suicide in children and adolescents. It is estimated that almost 57% of children and adolescents who experience suicidal thoughts, 70% of those with suicide plans, and 76% of those who attempt suicide have a diagnosable depressive disorder.10 Thus, optimizing care for patients who present with suicidality or "danger to self" can lead to cross-cutting improvements in inpatient care for highly prevalent mental health diagnoses and for the most severely ill children and adolescents.

Lastly, mental health assessments of youths who self-harm may also help identify those at increased risk for future self-harm behaviors. Adolescents who present with self-harm and who score high on depression are more likely to repeat their self-harm behavior.15 A measure, like the one we have developed, focused on documentation of a successful hand-off between the inpatient and outpatient settings for youth admitted with dangerous self-harm or suicidality may enhance the likelihood of them accessing outpatient mental health services after hospital discharge16-19 which could ultimately prevent a subsequent suicide, a devastating potential consequence of ongoing untreated mental illness.

1c.4. Citations for data demonstrating high priority provided in 1a.3

1. Zima BT, Murphy JM, Scholle SH, et al. National quality measures for child mental health care: background, progress, and next steps. Pediatrics. 2013;131 Suppl 1:S38-49.
2. Bardach NS, Coker TR, Zima BT, et al. Common and costly hospitalizations for pediatric mental health disorders. Pediatrics. 2014;133(4):602-609.
3. AACAP. Practice parameter for the assessment and treatment of children and adolescents with depressive disorders. J. Am. Acad. Child Adolesc. Psychiatry. 2007;46(11).
4. Tishler CL, Reiss NS, Rhodes AR. Suicidal behavior in children younger than twelve: a diagnostic challenge for emergency department personnel. Acad Emerg Med. 2007;14(9):810-818.
5. Kennedy SP, Baraff LJ, Suddath RL, Asarnow JR. Emergency department management of suicidal adolescents. Annals of Emergerncy Medicine. 2004;43(4):452-460.
6. Goldston DB, Daniel SS, Reboussin DM, Reboussin BA, Frazier PH, Kelley AE. Suicide attempts among formerly hospitalized adolescents: A prospective naturalistic study of risk during the first 5 years after discharge. Journal of the American Academy of Child & Adolescent Psychiatry. 1999;38(6):660-671.
7. Joiner TE, Rudd MD, Rouleau MR, Wagner KD. Parameters of suicide crises vary as a function of previous suicide attempts in youth inpatients. Journal of the American Academy of Child & Adolescent Psychiatry. 2000;39(7):876-880.
8. CDC. Youth Suicide. 2012; http://www.cdc.gov/violenceprevention/pub/youth_suicide.html. Accessed Oct 18, 2012.
9. CDC. Youth Risk Behavior Surveillance--United States 2011. MMWR Surveillance Summary 2011. 2012;61:1-162.
10. Nock MK, Green J, Hwang I, et al. Prevalence, correlates, and treatment of lifetime suicidal behavior among adolescents: Results from the national comorbidity survey replication adolescent supplement. JAMA Psychiatry. 2013:1-11.
11. Lewinsohn PM, Rohde P, Seeley JR. Major depressive disorder in older adolescents: Prevalence, risk factors, and clinical implications. Clinical Psychology Review. 1998;18(7):765-794.
12. Lewinsohn PM, Rohde P, Seeley JR. Psychosocial characteristics of adolescents with a history of suicide attempt. Journal of the American Academy of Child & Adolescent Psychiatry. 1993;32(1):60-68.
13. Nock MK, Kazdin AE. Examination of affective, cognitive, and behavioral factors and suicide-related outcomes in children and young adolescents. Journal of Clinical Child and Adolescent Psychology. 2002;31(1):48-58.
14. Guile JM, Brunelle J, Consoli A, Bodeau N, Cohen D. Bipolar disorder type I and suicide attempts in adolescence: Data from a follow-up study. Adolescent Psychiatry. 2012;2(1):88.
15. Hawton K, Kingsbury S, Steinhardt K, James A, Fagg J. Repetition of deliberate self-harm by adolescents: The role of psychological factors. Journal of Adolescence. 1999;22(3):369-378.

16. Preen DB, Bailey BE, Wright A, et al. Effects of a multidisciplinary, post-discharge continuance of care intervention on quality of life, discharge satisfaction, and hospital length of stay: a randomized controlled trial. Int J Qual Health Care. 2005;17(1):43-51.
17. Balaban RB, Weissman JS, Samuel PA, Woolhandler S. Redefining and redesigning hospital discharge to enhance patient care: a randomized controlled study. J Gen Intern Med. 2008;23(8):1228-1233.
18. Harrison MB, Browne GB, Roberts J, Tugwell P, Gafni A, Graham ID. Quality of life of individuals with heart failure: a randomized trial of the effectiveness of two models of hospital-to-home transition. Med Care. 2002;40(4):271-282.
19. Finn KM, Heffner R, Chang Y, et al. Improving the discharge process by embedding a discharge facilitator in a resident team. J Hosp Med. 2011;6(9):494-500.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Behavioral Health : Serious Mental Illness, Behavioral Health : Suicide, Mental Health, Mental Health : Serious Mental Illness, Mental Health : Suicide

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access, Care Coordination, Care Coordination : Readmissions, Safety : Readmissions

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Measure specifications can be found at the following URL under the heading: "Mental Health Measures":
http://www.seattlechildrens.org/research/child-health-behavior-and-development/mangione-smith-lab/measurement-tools/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** DANGER_TO_SELF_ICD9_and_ICD10_for_Denominator_Identification_SUBMITTED.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Children/adolescents admitted to the hospital for dangerous self-harm or suicidality should have documentation in the hospital

record of discussion between the hospital provider and the patient´s outpatient provider regarding the plan for follow-up (discussion can be by phone or email) prior to discharge.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
24 month period of data, retrospectively collected. We propose using 24 months due to the low prevalence of the condition. This was the period used in the field testing of the measure.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients passing the quality measure are identified during medical record abstraction using the guidelines below (See"Medical Records Abstraction Tool Guidelines" under "Mental Health Measures" provided on the website in S.1.) This language is also in the "Medical Records Electronic Abstraction and Scoring Tool" on the website in S.1.

Follow-up MD – [Module: Dangerous self-harm/suicidal ideation, inpatient care] Indicate "1" if at the time of discharge, the patient had a designated primary care provider (PCP) or psychiatrist who would manage the patient's care post-discharge. Even patients with no known provider at the time of hospital admission should have been referred to a follow-up provider who was a PCP or a psychiatrist at the time of discharge. Indicate "2" if there is no follow-up provider identified.

Follow-up MD: SI Plan - [Module: Dangerous self-harm/suicidal ideation, inpatient care] Indicate "1" if the hospital provider communicated (by telephone or email) with the follow-up provider (PCP or psychiatrist) during the time window of 24 hours prior to discharge to 48 hours after discharge. The window of time is computed based on the discharge date and time and is displayed within the question text in the data collection tool. The purpose of this communication is to be sure a safe transition is in place, as this item applies only to patients hospitalized for self-harm/suicidal ideation. Select response "2" if the hospital provider is also the follow-up outpatient provider OR if outpatient care has been arranged to be continued in the marker hospital´s own psychiatric outpatient clinic. The latter arrangement is considered to be an adequate communication of the safety plan for the patient. If you cannot verify that there was any communication between the hospital provider and the follow-up PCP/psychiatrist AND there is no same-institution psychiatric clinic follow-up arranged, select response "3" (Neither of the above/No data).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients aged >=5 to <=19 years-old admitted to the hospital with a discharge diagnosis of danger to self or suicidality.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Cases are identified from hospital administrative data using the field for patient age and any diagnosis fields (primary or subsequent).

Patients aged >=5 to <=19 years

Patients have at least one of the following ICD9 codes for suicidal ideation as a primary or other discharge diagnosis:
e950-e959, V62.84

These codes were chosen by Members of the COE4CCN Mental Health Working Group (see Ad.1) co-chaired by Psychiatric Health Services Researchers Drs. Michael Murphy and Bonnie Zima.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients are excluded if they are transferred to an acute or non-acute inpatient facility, left against medical advice (AMA) or eloped. They are also excluded if the hospital provider is also the post-discharge provider or post-discharge follow-up is arranged to occur at the marker hospital's own outpatient psychiatric clinic.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Denominator exclusions are made using the following information obtained during medical abstraction (see Item S.18 for scoring using this information):
Discharge Disposition – [Module: Dangerous self-harm/suicidal ideation, inpatient care] Indicate the patient's disposition at discharge. If the patient was transferred to an acute or non-acute inpatient facility other than the marker hospital, select response "1" on the abstraction tool. This case will be excluded since care continued at that institution. Response "2" on the abstraction tool includes patients who left AMA or who eloped. Response"3" on the abstraction tool is for patients who were discharged to some sort of holding facility such as jail, juvenile detention, or other holding placement. Response "4" on the abstraction tool is for patients who were discharged to half- or partial-hospitalization. The definition of half- or partial-hospitalization varies among sites, but in general indicates an arrangement where the patient is at home at night, but in a therapeutic environment during the day. Response "5" on the abstraction tool is for patients who were discharged to home, which includes a foster home or other group homelike arrangement.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. N= The hospital's eligible target denominator population using administrative claims data

2. n= The numerator population, the cases meeting the target process

The numerator is the sum of those cases with a Pass from the denominator, calculated using the results from the data abstracted in Item S.6 above:
Score = Pass =1 if Follow-up MD: SI Plan = 1 (communication within specified time window)
Score = Fail =0 if Follow-up MD = 2 (no follow-up PCP or psychiatrist identified by inpatient team by the time of discharge).

Score = Fail =0 if Follow-up MD: SI plan = 3 (no communication within time window)

3. e= The patients excluded based on medical record abstraction (Item S.11)
Patients are excluded from the denominator of the measure if they are transferred to an inpatient facility or left the hospital against medical advice or eloped (Discharge Disposition = "1" or "2"). They are also excluded if the hospital provider is also the post-discharge provider or post-discharge follow-up is arranged to occur at the marker hospital's own outpatient psychiatric clinic (Follow-up MD: SI plan = "2").

Patients are eligible for the measure (included in the denominator if the abstractor selects values "3", "4", or "5" on the abstraction tool (discharged to jail, juvenile detention or other holding placement, half- or partial-hospitalization, or home) and the post-discharge provider is not the hospital provider or marker hospital outpatient psychiatric clinic (Follow-up MD: SI plan is not equal to "2").

4. Calculate the score: 100*n/(N-e)

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A. Given the low prevalence of the condition, the measured group is the entire population of eligible patients.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data collection tool is publicly available on the website in S.1. under "Mental Health Measures."
Title: "Medical Record Measure Electronic Abstraction and Scoring Tool"

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health/Psychiatric : Inpatient, Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
DSD9_Inpt_Transition_Communication_Testing_2015_10_13_SUBMITTED.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Pediatric Danger to Self: Discharge Communication with Outpatient Provider
Date of Submission: 10/8/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ **Process**
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient

preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

As described in the submission form, the validity and feasibility of the COE4CCN pediatric mental health measures were evaluated by a multi-stakeholder panel using the RAND-University of California, Los Angeles (UCLA) modified Delphi method.[1]

Detailed measure specifications were developed for the Delphi panel endorsed pediatric mental health quality measures. These specifications were then used to develop an electronic excel macro data collection tool for use with medical records data. The tool has automated scoring capability and is available on the website listed in item S.1. Abstraction and scoring guidelines are provided as an appendix to this submission.

Field Testing of the Delphi Panel Endorsed Pediatric Mental Health Quality Measures

Three tertiary care children's hospitals participated in the field test of the ***Pediatric Danger to Self/Suicidality*** Mental Health quality measures. For each hospital, two research nurses were trained to use the medical record abstraction tool and the companion abstraction tool guidelines. For training purposes, the nurses abstracted excerpts from several sample charts targeting the abstraction content for the mental health conditions and including both Emergency Department (ED) and inpatient care. Their abstractions were compared to gold-standard abstractions previously completed by the developer of the measure specifications. Abstractors were considered fully trained when the trainer observed that they could reliably abstract the applicable gold-standard medical record excerpts.

Case Selection

Cases for the field test were selected using International Classification of Diseases 9th Revision Clinical Modification (ICD-9) codes for danger to self from administrative databases from each hospital for discharges occurring between January 1st,2012 and December 31st, 2013 (see Appendix for a list of ICD-9 codes used to select cases for abstraction).

The final sample goal for danger to self/suicidality was a total of 100 cases selected from the two larger hospitals and 30 from the smaller hospital, with 25% replacement cases in order to have adequate sample after patients were excluded during the medical record abstraction phase. Because of limited available sample sizes at each hospital for Danger to Self/Suicidality, all eligible patients were included in the final sample. See **Table 2b5.1** for sample sizes in each hospital.

Medical Record Abstractions

At each hospital, the two trained nurse abstractors were each assigned half of the case sample for Danger to Self/Suicidality. Data for each case were entered by the nurses into the electronic abstraction tool and both the raw data and auto-generated quality measure scores were uploaded to a central research database for further analysis.

At the two larger tertiary care hospitals, each nurse abstracted pediatric Danger to Self/Suicidality measures from 40 additional charts that were randomly selected from the other nurse's sample to facilitate assessment of inter-rater reliability (see inter-rater reliability testing results in **2a2.3** below). The 40 charts were among a total of 120 (10% sample) pulled for inter-rater reliability testing of quality measures we developed and tested across three different mental health diagnoses (psychosis, danger to self/suicidality, and substance abuse).

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From:	Measure Tested with Data From:

(*must be consistent with data sources entered in S.23*)	
☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Two existing administrative datasets were used to sample patients using the ICD9 codes.

The Pediatric Health Information System (PHIS) database was used to sample the medical records from two of the children's hospitals. This is a comparative pediatric database, and includes clinical and resource utilization data for inpatient, ambulatory surgery, emergency department and observation unit patient encounters for 45 children's hospitals. (More information about PHIS is available at: https://www.childrenshospitals.org/Programs-and-Services/Data-Analytics-and-Research/Pediatric-Health-Information-System)

The hospital administrative discharge database was used to sample the medical records from the third field test hospital.

1.3. What are the dates of the data used in testing? January 1, 2012-December 31st, 2013

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Three hospitals that admit children in the targeted age range were included in the field test. All three are stand-alone children's hospitals. They are located in Washington state (Seattle Children's Hospital), Minnesota (University of Minnesota Children's Hospital), and Ohio (Cincinnati Children's Hospital). All have dedicated inpatient psychiatric units.

These hospitals were selected as they are all member organizations of the COE4CCN multi-stakeholder consortium of organizations that took part in the Center's measure development activities.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Table 1.6: Sociodemographic Characteristics of Patients Eligible for Measurement with Pediatric Danger to Self: Discharge Communication with Outpatient Provider (N=177)

	N	%
Child gender		
Male	56	32
Female	121	68
Missing	0	0
Child race/ethnicity		
Hispanic	7	4
White	121	68
Black	25	14
Other	21	12
Missing	3	2
Insurance type		
Public	85	48
Private	86	49
Uninsured	6	3
Missing	0	0
PMCA category*		
Non-chronic condition	7	4
Non-complex chronic condition	94	60
Complex chronic condition	55	35
Missing	0	0

* PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015).[2] Available only at 2 of the 3 participating hospitals.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

N/A

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

To measure patient-level sociodemongraphic variables, we used patient gender, race, ethnicity, insurance type, and chronic disease status. These variables were derived from the administrative claims data from each participating hospital. Chronic disease status was captured using the Pediatric Medical Complexity Algorithm (PMCA), which categorizes pediatric inpatients using diagnostic ICD9 codes as having an acute medical condition only (non-chronic condition), a non-complex chronic condition, or a complex chronic condition.[2] Retrospective claims data needed to run PMCA were only available from 2 of the field test hospitals.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Critical data elements used in the measure were tested for inter-rater reliability of medical record abstraction. Reliability was measured using the prevalence adjusted bias adjusted kappa (PABAK) statistic for patient eligibility for measurement, and for the patient score for the quality measure. Kappa is a statistic that captures the proportion of agreement beyond that expected by chance, that is, the *achieved* beyond-chance agreement as a portion of the *possible* beyond-chance agreement.[3] PABAK is a measure of inter-rater reliability that adjusts the magnitude of the kappa statistic to take into account the influences of high or low prevalence and of inter-rater differences in assessment of prevalence. The PABAK statistic adjusts for high or low prevalence and is what we used in our calculations of inter-rater reliability.

Performance measure score was assessed for reliability across performance sites using the intra-class correlation coefficient (ICC). The ICC assesses the ratio of between site variation and within site variation on performance. Higher ICC implies that the between site variation (signal) is higher than the within site variation (noise). ICCs were computed using STATA SE 13.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g*., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Critical data elements:

The specific measure addressed in this submission was one of a group of danger to self/suicidality measures included in the field test as part of the broader COE4CCN Pediatric Mental Health Measures in the Hospital Setting development effort.

There are two stages of medical record abstraction for which we tested inter-rater reliability for danger to self/suicidality measures: *patient eligibility for the measure;* and *patient score for the quality measure*.

Sampling of charts

Patient eligibility: 40 charts for IRR assessment for danger to self/suicidality measures were sampled and tested for IRR for assessing patient eligibility.

Patient score: IRR for quality measure patient score was only calculated for the **subset** of the sampled charts of patients deemed eligible for the specific quality measure.

IRR results

Patient eligibility

Kappa for patient eligibility (n=40 charts):

0.80

PABAK for patient eligibility (n=40 charts):

0.85

Patient score

There was only a small subset (n=8) of the randomly sampled reliability charts that were eligible specifically for the **Pediatric Danger to Self: Discharge Communication with Outpatient Provider** measure. This sample of cases was too small to calculate a kappa for patient score. Instead, we present the percent agreement between abstractors regarding patient score for this measure.

Percent agreement for patient scores on the quality measure under consideration:

88% (7 of 8)

Performance measure score reliability:

We performed ICC testing for performance variation at the level of the hospital, since that is the intended level of measurement. However, despite adequate sample size at the patient level within each site (see Table 2b5.1 below), the number of higher level clusters in our field test is limited to the 3 participating hospitals. Future measurement across a larger number of participating hospitals will give more generalizable estimations of ICC for this measure.

Hospital-level ICC (N=3 hospitals):

0.34 (95%CI 0.03-0.92)

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Critical data elements: Interpretation of Kappas is generally cited as follows[3,4]: ≤0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect.

Hence, inter-rater reliability for eligibility for this measure was almost perfect. Percent agreement was very high.

Performance measure score: Hospital level ICC based on the three hospitals is relatively high. ICCs ≥0.10 are considered relatively high.[5] Hence, the ICCs indicate that there are meaningful between-site performance differences.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

CRITICAL DATA ELEMENTS

ICD10 CONVERSION (no testing performed)

1. Statement of intent for the selection of ICD-10 codes:
    a. **The goal is to convert this measure to a new code set, fully consistent with the intent of the original measure.**
2. Excel spreadsheet with original ICD-9 codes from the Field test and the ICD9-ICD10 conversion table is attached at S.2b
3. Description of the process used to identify ICD-10 codes, including:
    a. Experts who assisted in the process:
        i. Bonnie Zima (co-chair Mental Health Working Group, see Ad.1)
        ii. Michael Murphy (co-chair Mental Health Working Group, see Ad.1)
    b. Name of the tool used to identify/map to ICD-10 codes:
        i. Transformation was based on the Centers for Medicaid and Medicare Services Gems tool.
    c. Stakeholder input was obtained from the COE4CCN Mental Health Multi-stakeholder Working Group. See below.

Danger to Self/Suicidality ICD9 to ICD10: Stakeholder Comments

A) Researcher and practitioner stakeholder #1:

I am questioning the Y92 codes - I may not understand them but they don't look like they specify self-harm, just location when injured.

Response: See response to Stakeholder #3 below. Deleted these codes.

For the extra codes, I am not sure about late effect - mainly because it could be years after the attempt.

Response: See response to Stakeholder #3 below. Deleted these codes

B) Researcher and practitioner stakeholder #2:

"I read all the new ICD 10 dx for both psychosis and substance abuse and they all seemed appropriate.

They also all seemed to correspond pretty well to their ICD 9 antecedents.
However, the former ICD9 Dx e950x-e953.x are marked 'no Dx' with the note that they are covered by other ICD9 Dx (and corresponding ICD 10) but these are not shown.

It would be important to verify that this is so because these codes cover self poisoning using various substances including drug overdoses which I imagine will be one of the most common methods

So there would undoubtedly be kids in our data set whose ICD 9 Dx would not have an ICD 10 code unless the other criteria were given someplace else. So I am saying that it would make sense to check these."

Response: compilation of all causes of suicidality from ICD10, under the Suicidality chapter

Stakeholder #2 Comment after subsequent compilation of lists for all causes of suicidality in ICD10:

"I looked over the new spreadsheet [of codes for Suicidality from ICD10] and the Table of Drugs and Chemicals.

I am signing off on these lists. I think that the codes make sense.

I also think that most of the things that cause harm are included in the Table of Drugs and Chemicals (169 pages and about 7500 drugs and chemicals!)."

C) Researcher and practitioner stakeholder #3:

"I am assuming that the Y92 codes are the specific information re: mode of the suicide? If so, this is way too much information and I would suggest we just use SI yes/no."

Response: Because ICD10 is deigned to capture detail, we will miss many codes if we do not include the more detailed ones. However, we will not include Y92 codes based on other stakeholder input above, and based on the intended use of Y92 codes: They are supposed to always be used with an external cause code, and indicate only the specific place where the external cause occurred. We are not including it, since it should not increase sensitivity if everyone uses it with an external cause code, and it will decrease specificity if people use it incorrectly—e.g., if they don't code for the external cause, and we pick up the case based on this code alone, and it turns out that it was not intentional self-harm, we will potentially capture people who are not eligible for the measure.

"Please clarify the operational definition for "late effect"—at first glance it seemed to me that this person likely had a primary dx of a traumatic injury and then later someone also documented the etiology. To determine the operational definition, we'd have to look at the code book to be sure ….."

Response: We went to the code book to better understand. It clarifies three different options for the last place of the 7 digit code, that indicates whether it is the initial visit, a subsequent visit, or a visit related to sequelae. This code only includes that third category, which is why Laura was concerned that it could refer to a sitatuion years out from an event. So we will not include, since again, it will potentially include patients in the eligible population for the measure which are likely to not be relevant for inclusion, if they are admitted for something unrelated to mental health, which is a sequelae of their attempt at self harm in the past. The goal is to get a population that was hospitalized for suicide or self harm, to make sure that there is a good hand-off to the outpatient setting.

"For both late effect and mode of SI, I wondered whether only someone who tended to report more comprehensively would add this level of detail in their reporting?"

Response: Probably, but ICD10 is so new that only time will tell how people end up using the codes. We are making our best set of codes based on the description of how the codes are supposed to be used.

D) State Medicaid office stakeholder #4:

"The mental health folks in my agency are ahead of the rest of us as they have created crosswalks that make sense for our programs. Basically the codes are being based off of the DSM-5. The DSM-5 diagnoses lists both ICD-9 and ICD-10 codes with the diagnoses."

Response: Since Danger to Self does not have a specific DSM chapter, we used the ICD10 chapter and used the crosswalk from CMS GEMs for the ICD-9 codes we used in the field test.

PERFORMANCE MEASURE SCORE

EMPIRICAL VALIDITY TESTING

We assessed the patient-level relationship between meeting the quality measure and three utilization outcomes that, per our conceptual model, were outcomes of interests and which we hypothesized a priori might have a relationship with the measure.

Multivariable regression was used to assess the independent relationship between meeting the measure and the validation metric of interest, independent of other confounders. Covariates were chosen based on face validity (gender and insurance type) and based on empirical evidence that they were associated with both the quality measure and the validation metric (admitting hospital, and child race/ethnicity).

Validation Metrics:

30 day readmission to the hospital (measured as readmission within 30 days of discharge, to the same hospital, since we did not have data on readmissions to other hospitals). (logistic model)

30-day return ED visit (measured as return visit within 30 days of discharge, to the same hospital, since we did not have data on readmissions to other hospitals). (logistic model)

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF FACE VALIDITY—The RAND-UCLA Modified Delphi Method

The face validity of the group of quality measures developed in the COE4CCN Pediatric Mental Health measures effort, which included the danger to self/suicidality measure proposed, was established using the RAND-UCLA Modified Delphi Method. The process began with the nomination of 10 individuals by 8 stakeholder organizations including the American Academy of Child and Adolescent Psychiatry, the American Academy of Pediatrics (AAP) Committee on Pediatric Emergency Medicine, the AAP Task Force on Mental Health, the Medicaid Medical Directors Learning Network, the AAP Section on Hospitalist Medicine, Family Voices, the Society for Adolescent Health and Medicine, and the Substance Abuse and Mental Health Services Administration. Nine of the nominees agreed to be members of our multi-stakeholder Delphi panel. All panelists were people deemed by the nominating organizations to have substantial expertise and/or experience related to child mental health (see Ad.1 for a list of panel members). The panel read the danger to self/suicidality literature review written by project staff and reviewed and scored each proposed quality measure on validity. This method is a well-established, structured approach to measure evaluation that involves two rounds of independent panel member scoring, with group discussion in between.[1] After reviewing the literature review and draft danger to self/suicidality quality measures, panel members were asked to rate each measure's validity on a scale from 1 (low) to 9 (high). Validity was assessed by considering whether there was adequate scientific evidence or expert consensus to support the measure's link to better outcomes; whether there would be health benefits associated with receiving measure-specified care; whether panelists would consider providers who adhere more consistently to the quality measure to be providing higher quality care; and whether adherence to the measure is under the control of health care providers and/or systems. The Delphi method has been found to be reliable and to have content, construct and predictive validity.[6-10] For a quality measure or measure component to move to the next stage of measure development, it had to have a median validity score $\geq$ 7 (1-9 scale) and be scored without disagreement based on the mean absolute deviation from the median after the

second round of scoring. This process ensures that only measures widely judged to be valid moved forward into measure specification. See **Table 2b.2.3** for Delphi panel scores on the measure for this submission.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

CRITICAL DATA ELEMENTS
ICD10 CONVERSION (no testing performed)

PERFORMANCE MEASURE SCORE
EMPIRICAL VALIDITY TESTING

Table 2b2.3. Validation Metrics for Pediatric Danger to Self: Discharge Communication with Outpatient Provider (N=177)

	Met measure (n=48)	Did not meet measure (n=129)	Adjusted OR (95% CI)*	p-value
30-day readmissions, n (%)	7/48 (14.6%)	13/129 (10.1%)	1.00 (0.99-1.02)	0.45
30-day ED (Orevisits, n (%)	5/48 (10.4%)	8/129 (6.2%)	1.01 (1.00-1.02)	0.19

*Adjusted for hospital, race/ethnicity, gender, and insurance type, modeled using logistic regression.

PERFORMANCE MEASURE SCORE
SYSTEMATIC ASSESSMENT OF FACE VALIDITY—The RAND-UCLA Modified Delphi Method: The scores for this measure from the 9 members of the panel after round 2 of Delphi scoring (scoring done after discussions at the in-person meeting) are presented in the Table below.

Table 2b2.3: Delphi scores for Pediatric Danger to Self: Discharge Communication with Outpatient Provider

	Median score	Mean absolute deviation from median	Agreement status*
Validity	9.0	0.7	Agree
Feasibility	9.0	0.7	Agree

*This is a statistical assessment of whether panelists agreed (A), disagreed (D), or if level of agreement was indeterminate (I)

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

PERFORMANCE MEASURE SCORE
EMPIRICAL VALIDITY TESTING
There were no statistically significant differences between those meeting and those failing the measure in readmissions and ED revisits. The relatively low sample size of eligible patients for this measure may have led to limited power to demonstrate a difference in readmission or ED return visits for patients passing versus failing this quality measure.

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF FACE VALIDITY—The RAND-UCLA Modified Delphi Method: The results from the Delphi panel show strong face validity for this measure, with the highest possible median validity scores (9 out of 9) following round 2 of the Delphi panel scoring.

2b3. EXCLUSIONS ANALYSIS

☒ **no exclusions — *skip to section 2b4***

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

As noted in the Submission Item 1b, we performed a field test of the quality measure under consideration. We measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital. Included patients were discharged from one of the three hospitals over the two year period (January 1, 2012-December 31, 2013). The performance scores are presented below in Tables 2b5.2a (performance variation across hospitals) and Table 2b5.2b (performance variation across socio-demographic characteristics)**.**

We tested the difference in performance across the hospitals using an omnibus test for difference, and then performing individual comparisons between each hospital's performance and the performance of the group as a whole. We used Fisher's exact test to assess statistical significance for all comparisons.

Performance variation across sociodemographic characteristics was assessed using logistic regression. Performance results in Table 2b5.2b are presented for each characteristic—each characteristic was modeled without adjusting for other covariates.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Omnibus testing for difference in performance across all three hospitals found statistically significant difference, with a p-value of 0.002.

Table 2b5.2a. Hospital Performance Scores for Pediatric Danger to Self: Discharge Communication with Outpatient Provider

	Denominator	Numerator	Score	Difference from overall mean of others*	P-value for difference from overall mean of others**
Hospital A	79	30	38.0	19.6	0.004
Hospital B	77	18	23.4	-6.6	0.40
Hospital C	21	0	0.0	-30.8	0.0012

* Each hospital's performance was compared to the pooled mean of the other two hospitals.
**Statistical testing using Fisher's exact test.

Table 2b5.2b. Performance Scores by Socio-demographic Characteristics for Pediatric Danger to Self: Discharge Communication with Outpatient Provider

	N	%	SD	OR*	LCL	UCL
Child gender						
Male	56	30.4	46.4	1.3	0.6	2.6
Female (ref)	121	25.6	43.8			
Child race/ethnicity						
White (ref)	121	28.1	45.1			
Hispanic	7	14.3	37.8	0.4	0.1	3.7
Black	25	32.0	47.6	1.2	0.5	3.1
Other	21	23.8	43.6	0.7	0.2	2.0
Insurance type						
Private (ref)	86	27.9	45.1			
Public/uninsured	91	26.4	44.3	0.9	0.5	1.8
**PMCA category ** **						
Non-chronic (ref)	7	14.3	37.8			
Non-complex chronic	94	28.7	45.5	2.4	0.3	21.0
Complex chronic	55	36.4	48.6	3.4	0.4	30.5

*No performance differences by group were statistically significant. Differences tested using logistic regression, without adjusting for other covariates.
**PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015).[2] Available only at 2 of the 3 participating hospitals.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

For this pilot field test assessing for existing variation in this measure across more than one site, we found that we were able to detect statistically and clinically meaningful differences in hospital performance. Additional information from implementation of the measure at a larger scale, as described in Section 4.1, will assist in assessing variation across a larger group of hospitals.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Missing data likely does not contribute to substantially or meaningfully biased estimates of performance for this measure.

There are two potential areas for missing data: at the level of the administrative claims, which are used for sampling patients, and in the medical abstraction stage.

Administrative Claims

There are two data fields used to identify patients, the diagnosis fields, and the patient age. Patient age is generally considered a reliable field and has minimal missing data.

A primary diagnosis is required for billing, and therefore also is rarely missing. It is known that some providers under-code for mental health diagnoses, which would lead to a risk of under recognition of eligible cases. This may lead to difficulty in capturing reliable estimates of performance at each hospital site, but is less likely to lead to biased estimates. In addition, it is likely that an admitted patient with danger to self/suicidality will need additional long term services, hence leading to a higher likelihood of a diagnosis being documented.

Medical abstraction

Missing data in the medical abstraction stage is interpreted as the patient not meeting the metric. To the degree that patients are meeting metrics at the site and providers are not documenting this in the medical record (false negative performance scoring), performance measurement (and accompanying internal feedback or public reporting) will likely stimulate improved documentation. This improved documentation will allow more valid assessments of the relationship of the process of care assessed by this measure, i.e., inpatient-to-outpatient provider communication regarding the follow-up plan and patient outcomes (e.g. decreased readmissions, increased completion of outpatient follow-up appointments).

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

In the PHIS dataset (used for case finding at Seattle Children's and Cincinnati Children's), age is a required element, and so was not missing for any records for patients from the hospitals with PHIS data. We do not have documentation for how often data was missing from patient medical records regarding patient age at the other hospital.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

It is unlikely that missing data contributes to substantial or meaningful biases of performance estimates. See item #2b7.1 for additional discussion of this.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

Electronic capture of these data could be operationalized through the use of a structured data field in an electronic medical record to indicate whether discussion between the hospital provider and the patient´s outpatient provider regarding the plan for follow-up (discussion can be by phone or email) occurred prior to discharge.

Use of this structured field would need to be validated in future testing, e.g., validated through caregiver report.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

<u>IF a PRO-PM</u>, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

In field testing, we found that it was important to specify the potential cases regarding whether the follow-up MD was the same as the treating psychiatrist, in which case we would not expect documentation of a communication to occur. We document this in the data collection tool for review during abstraction, using the following language:

"Select response 2 if the hospital provider is also the follow-up outpatient provider OR if outpatient care has been arranged to be continued in the marker hospital's own psychiatric outpatient clinic. The latter arrangement is considered to be an adequate

communication of the safety plan for the patient."

These patients were excluded from scoring.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
No proprietary elements are used in implementing this measure. There are no licenses or fees or other requirements needed to use any aspect of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

NA

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This measure is part of a set of mental health quality measures the COE4CCN developed as part of the Pediatric Quality Measurement Program, funded by AHRQ, using CHIPRA monies. It has not yet been implemented as the development and validation were just recently completed. The tools needed to abstract the measures, available online at the website in S.1, are publicly available and non-proprietary, so interested parties can implement them at any time.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Children's Hospital Association (CHA) has had representation on the National Advisory Board for COE4CCN since its inception. CHA has shown great interest in promoting the adoption of inpatient and ED-based measures developed by our Center. The intended audience would be hospital administrators at CHA member hospitals. We would intend to work with CHA to implement these measures over the next several years.

We also intend to publish the development and field testing of these measures in peer reviewed pediatric journals over the next 12 months. Within these publications we will include the URL where the measure data abstraction tool, measure specifications, and abstractor guidelines are housed promoting further access to and dissemination of the measures.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Credible rationale

The overall goal behind capturing performance results for this measure is to optimize transitions of care for a high-risk population. As noted above in section 1c.3, adolescents who present with self-harm are more likely to repeat their self-harm behavior. The goal in focusing on the transition from the highly monitored inpatient setting to the outpatient setting is to optimize the chance of preventing a future attempt at self-harm. Adequate communication between inpatient and outpatient providers is a key element to a successful transition (see Evidence form).

As experience has borne out, quality measurement efforts can drive improvements in care, whether through increasing focus on an area of care in internal audit and feedback efforts, or through reputational or financial incentive programs (e.g., CMS' public reporting or value-based purchasing programs). We anticipate that the performance results for this measure would drive improvement through similar mechanisms.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0576 : Follow-Up After Hospitalization for Mental Illness (FUH)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Measure #0576 focuses on the population of pediatric patients with any mental health diagnosis and assesses whether they had follow-up appointments within 7 and 30 days after hospitalization. Though #0576 and the proposed measure both focus on the transition from inpatient to outpatient care, the proposed measure focuses on a different process to support a successful transition. In addition, this measure has a more specific measure population – one that is at particularly high risk if successful follow-up doesn't occur after hospital discharge.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** DSD9_Inpt_Transition_Appendix_FOR_SUBMISSION.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Seattle Children´s Research Institute
Co.2 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-
Co.3 **Measure Developer if different from Measure Steward:** Seattle Children´s Research Institute
Co.4 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role

in measure development.

The COE4CCN convened two groups to assist in the development of the Pediatric Mental Health Measures in the Hospital Setting--the COE4CCN Mental Health Working Group and an external, multi-stakeholder Delphi panel. Please see descriptions of the groups' roles in development as well as member names listed below.

I. Mental Health Working Group: This was a group of pediatric mental health and general pediatric experts, as well as state Medicaid leadership. This group reviewed secondary database analyses examining the prevalence of common and costly mental health diagnoses; developed ICD9 code definitions to identify diagnoses of interest; reviewed and edited the literature reviews conducted by COE4CCN staff; provided content expertise during development of the detailed measure specifications and data abstraction tool; and participated in the planning and implementation of the field test as well as interpretation of the field test results; developed ICD10 code set for ICD9 to ICD10 conversion.

Members of the MHWG:

Naomi S. Bardach, MD, MAS
Assistant Professor of Pediatrics and Health Policy
Department of Pediatrics
Philip R. Lee Institute of Health Policy
University of California San Francisco

Tumaini Ruker Coker, MD, MBA
Assistant Professor of Pediatrics
David Geffen School of Medicine
University of California, Los Angeles
Associate Natural Scientist
RAND, Santa Monica

Glenace Edwall, PsyD, PhD, MPP
Director, Children's Mental Health Division
Minnesota State Health Access Data Assistance Center
Minnesota Department of Human Services

Penny Knapp, MD
Professor Emeritus
Departments of Psychiatry & Pediatrics
University of California Davis

Rita Mangione-Smith, MD, MPH
Professor and Chief | Division of General Pediatrics and Hospital Medicine
University of Washington Department of Pediatrics
Director | Quality of Care Research Fellowship
UW Department of Pediatrics and Seattle Children's Hospital
Investigator | Center for Child Health, Behavior, and Development
Seattle Children´s Research Institute

Michael Murphy, EdD
Associate Professor
Department of Psychology
Harvard Medical School
Staff Psychologist
Department of Child Psychiatry
Massachusetts General Hospital

Laura Marie Prager, MD
Associate Professor of Psychiatry
Department of Child Psychiatry
Massachusetts General Hospital

Laura Richardson, MD, MPH
Professor
Departments of Pediatrics and Psychiatry
Division of Adolescent Medicine
University of Washington
Investigator
Center for Child Health, Behavior, and Development
Seattle Children's Research Institute

Bonnie Zima, MD, MPH
Professor-in-Residence
Department of Psychiatry
University of California, Los Angeles
Associate Director
UCLA Health Services Research Center

II. Delphi panel: Reviewed the literature review and secondary database analyses as prepared by the MHWG and COE staff. Reviewed suggested quality measures for face validity and content validity based on the above materials and based on member expertise in the field.

Members of the Delphi panel:

Gary Blau, PhD
Chief, Child, Adolescent and Family Branch,
Center for Mental Health Services (CMHS),
Substance Abuse and Mental Health Services Administration (SAMHSA), Rockville, MD.
Clinical Faculty, Yale Child Study Center, Yale University

Regina Bussing, MD, MSHS
Professor, Division of Child and Adolescent Psychiatry,
Department of Psychiatry, Department of Pediatrics,
and Department of Clinical and Health Psychology,
University of Florida, Gainesville, FL
Director, Florida Outreach Project for Children and Young Adults Who Are Deaf-Blind

Thomas Chun, MD, MPH
Associate Professor, Departments of Emergency Medicine and Pediatrics
Assistant Dean of Admissions
Chair, Admissions Committee
The Alpert Medical School, Brown University
Medical Staff, Department of Pediatric Emergency Medicine
Hasbro Children's Hospital

Sean Ervin, MD, PhD
Assistant Professor in Pediatrics & General Internal Medicine
Hospitalist Medicine
Head of Section- Pediatric Hospital Medicine
Wake Forest University, School of Medicine
Winston-Salem, NC

Doris Lotz, MD, MPH
Medicaid Medical Director
New Hampshire Department of Health and Human Services
Office of Medicaid Business and Policy
Instructor, Geisel School of Medicine at Dartmouth, Department of Psychiatry

Lynn Pedraza, PhD
Executive Director of Family Voices,
Albuquerque, NM

Karen Pierce, MD, DLFAPA, DLFAACAP
Clinical Associate Professor, The Feinberg School of Medicine,
Northwestern University Medical School,
Department of Psychiatry and Behavioral Sciences, Chicago, IL,
President, Illinois Academy of Child Psychiatry

Robert Sege, MD, PhD, FAAP
Professor of Pediatrics, Boston University School of Medicine
Director, Division of Family and Child Advocacy, Boston Medical Center
Core Faculty, Harvard Injury Control Research Center
Core Faculty, Harvard Youth Violence Prevention Center

Gail Slap, MD, MSc
Professor of Pediatrics, Department of Pediatrics,
Professor of Medicine, Department of Medicine,
University of Pennsylvania School of Medicine

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2015

Ad.3 **Month and Year of most recent revision:** 11, 2014

Ad.4 **What is your frequency for review/update of this measure?** Every 6 months

Ad.5 **When is the next scheduled review/update for this measure?** 04, 2016

Ad.6 **Copyright statement:**

Ad.7 **Disclaimers:** none

Ad.8 **Additional Information/Comments:** none

Note: Include all ICD10 codes in this sheet, directly crosswalked from original ICD9 code set for field test, and also include all ICD10 codes for suicidality referenced in the tab "ICD10 codes for suicidality"
Note: Some have NoDx, corresponding to no ICD10 (see Notes), and some have more than one ICD10.

Root ICD9 Code	ICD9 used in Field Test	ICD9 label	ICD10 conversion from CMS GEMS tool	ICD10 label
V62.84	V62.84	suicidal ideation	R45851	Suicidal ideations
E950.X	E950	suic/self-pois w sol/liq*	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.0	poison-analgesics	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.1	poison-barbiturates	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.2	poison-sedat/hypnotic	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.3	poison-psychotropic agt	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.4	poison-drug/medicin nec	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.5	poison-drug/medicin nos	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.6	poison-agricult agent	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.7	poison-corrosiv/caustic	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.8	poison-arsenic	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E950.9	poison-solid/liquid nec	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
E951.X	E951	poison-utility gas*	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E951.0	poison-piped gas	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E951.1	poison-gas in container	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E951.8	poison-utility gas nec	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
E952.X	E952	poison-gas/vapor nec*	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E952.0	poison-exhaust gas	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E952.1	poison-co nec	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E952.8	poison-gas/vapor nec	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E952.9	poison-gas/vapor nos	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
E953.X	E953	injury-strangul/suffoc*	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E953.0	injury-hanging	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E953.1	injury-suff w plas bag	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E953.8	injury-strang/suff nec	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
	E953.9	injury-strang/suff nos	NoDx	See next tab (ICD10 codes for suicidality, and Table of Substances) (Note: SKIP, see next tab)
E954	E954	injury-submersion	X71.8XXA	Other intentional self-harm by drowning and submersion, initial encounter
E954	E954	injury-submersion	X71.9XXA	Intentional self-harm by drowning and submersion, unspecified, initial encounter
E955.0	E955.0	injury-handgun	X72.XXXA	Intentional self-harm by handgun discharge, initial encounter
E955.1	E955.1	injury-shotgun	X73.0XXA	Intentional self-harm by shotgun discharge, initial encounter

E955.2	E955.2	injury-hunting rifle	X73.1XXA	Intentional self-harm by hunting rifle discharge, initial encounter
E955.3	E955.3	injury-military firearm	X73.2XXA	Intentional self-harm by machine gun discharge, initial encounter
E955.4	E955.4	injury-firearm nec	X73.9XXA	Intentional self-harm by unspecified larger firearm discharge, initial encounter
E955.5	E955.5	injury-explosives	X75.XXXA	Intentional self-harm by explosive material, initial encounter
E955.6	E955.6	self inflict acc-air gun	X74.01XA	Intentional self-harm by airgun, initial encounter
E955.7	E955.7	self inj-paintball gun	X74.02XA	Intentional self-harm by paintball gun, initial encounter
E955.9	E955.9	injury-firearm/expl nos	X74.9XXA	Intentional self-harm by unspecified firearm discharge, initial encounter
E956	E956	injury-cut instrument	X78.9XXA	Intentional self-harm by unspecified sharp object, initial encounter
E957.0	E957	inju-jump from hi place*	X80.XXXA	Intentional self-harm by jumping from a high place, initial encounter
E957.1	E957.1	injury-jump fm struc nec	X80.XXXA	Intentional self-harm by jumping from a high place, initial encounter
E957.2	E957.2	injury-jump fm natur sit	X80.XXXA	Intentional self-harm by jumping from a high place, initial encounter
E957.9	E957.9	injury-jump nec	X80.XXXA	Intentional self-harm by jumping from a high place, initial encounter
E958.0	E958.0	injury-moving object	X81.8XXA	Intentional self-harm by jumping or lying in front of other moving object, initial encounter
E958.1	E958.1	injury-burn, fire	X76.XXXA	Intentional self-harm by smoke, fire and flames, initial encounter
E958.2	E958.2	injury-scald	X77.2XXA	Intentional self-harm by other hot fluids, initial encounter
E958.3	E958.3	injury-extreme cold	X83.2XXA	Intentional self-harm by exposure to extremes of cold, initial encounter
E958.4	E958.4	injury-electrocution	X83.1XXA	Intentional self-harm by electrocution, initial encounter
E958.5	E958.5	injury-motor veh crash	X82.8XXA	Other intentional self-harm by crashing of motor vehicle, initial encounter
E958.6	E958.6	injury-aircraft crash	X83.0XXA	Intentional self-harm by crashing of aircraft, initial encounter
E958.7	E958.7	injury-caustic substance	X83.8XXA	Intentional self-harm by other specified means, initial encounter
E958.8	E958.8	injury-nec	X83.8XXA	Intentional self-harm by other specified means, initial encounter
E958.9	E958.9	injury-nos	X83.8XXA	Intentional self-harm by other specified means, initial encounter

Codes for suicidality from ICD10. Include all for ICD10 conversion for Danger to Self Indicator

http://www.tacomacc.edu/UserFiles/Servers/Server_6/File/him//HIM240/ICD10Cmcodebook/icd10cm_drug_2011.pdf

Codes are listed in column E except for those for specific drugs and chemicals, which are available en block in the Table of Drugs and Chemicals

Level 1 descriptor	Level 2 descriptor	Level 3 descriptor	Level 4 descriptor	ICD10 code
Suicide, suicidal (attempted) (by)				X83.8
	blunt object			X79
	burning, burns			X76
		hot object		X77.9
			fluid NEC	X77.2

			household appliance	X77.3
			specified NEC	X77.8
			steam	X77.0
			tap water	X77.1
			vapors	X77.0
	caustic substance			. (Note: See Table of Drugs and Chemicals, column "Poisoning, Intentional Self Harm")
	cold, extreme			X83.2
	collision of motor vehicle with			.
		motor vehicle		X82.0
		specified NEC		X82.8
		train		X82.1
		tree		X82.2
	crashing of aircraft			X83.0
	cut (any part of body)			X78.9
	cutting or piercing instrument			X78.9
		dagger		X78.2
		glass		X78.0
		knife		X78.1
		specified NEC		X78.8
		sword		X78.2
	drowning (in)			X71.9
		bathtub		X71.0
		natural water		X71.3
		specified NEC		X71.8
		swimming pool		X71.1
			following fall	X71.2
	electrocution			X83.1
	explosive (s) (material)			X75
	fire, flames			X76
	firearm			X74.9
		airgun		X74.01
		handgun		X72
		hunting rifle		X73.1
		larger		X73.9

			specified NEC	X73.8
		machine gun		X73.2
		shotgun		X73.0
		specified NEC		X74.8
	hanging			X83.8
	hot object —see Suicide, burning, hot object			.
	jumping			.
		before moving object		X81.8
			motor vehicle	X81.0
			subway train	X81.1
			train	X81.1
		from high place		X80
	lying before moving object, train, vehicle			X81.8
	poisoning —see Table of Drugs and Chemicals			. (Note: See Table of Drugs and Chemicals, column "Poisoning, Intentional Self Harm")
	puncture (any part of body) —see Suicide, cutting or piercing instrument			.
	scald —see Suicide, burning, hot object			.
	sharp object (any) —see Suicide, cutting or piercing instrument			.
	shooting —see Suicide, firearm			.
	specified means NEC			X83.8
	stab (any part of body) —see Suicide, cutting or piercing instrument			.
	steam, hot vapors			X77.0
	strangulation			X83.8
	submersion —see Suicide, drowning			X83.8
	suffocation			.
	wound NEC			X83.8

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2815
Measure Title: CAPQuaM PQMP Mental Health Follow Up Measure Timeliness 1: Delayed coordination of care following mental health discharge
Measure Steward: University Hospitals Cleveland Medical Center
Brief Description of Measure: This measure describes the presence or absence of delay in follow up visits with mental health and primary care clinicians following hospital discharge of a child with a primary mental health diagnosis or from a mental health facility.
Developer Rationale: Pediatric mental health hospitalizations have increased 24% during 2007 – 2010. In total, US children spent 1,721,765 days in hospitals for mental health care in 2012. Recent estimates put the cost of mental health hospitalization of children at $11.6 billion between 2006 and 2011. This analysis also found mental health admissions were higher among black and white children compared with Hispanic children, and were more common for children with public insurance than private or no insurance. Additionally, children who are admitted to the hospital for a mental health condition are very likely to meet criteria for Children with a special health care need.

Follow-up is a key component of the optimal management of any number of medical conditions, but is especially critical for children with mental health diagnoses. Timely follow up with both primary care clinicians and mental health practitioners after a hospital discharge are imperative to deliver the best outcomes. According to a guideline developed by the American Academy of Child and Adolescent Psychiatry (AACAP) and the American Psychiatric Association (APA) (1997), there is a need for regular and timely assessments and documentation of the patient´s response to all treatments. When considering follow up care, the literature makes the distinction between coordination across systems of care (in this case the primary care system and the mental health system), and continuity (in this case within the mental health system)."

Therefore, this measure looks at continuity of care as a component of coordination of care and assess follow up appointments both within and outside of the mental healthcare system. This measure reflects one aspect of coordination of care following the discharge of a child who has been hospitalized for a primary diagnosis that is specified as being a mental health diagnosis. This measure describes key attributes regarding the timeliness of follow up after a mental health discharge for children. Specifically, this measure set looks at the failure to establish timely follow up care subsequent to the day of discharge in both the primary care and mental healthcare systems. Stratification by type of failure enhances the granularity of the measure and provides clear data to support improvement initiatives. The measure is specified to be reported as an aggregate for the included age groups (Birth-21 years, with 19-21 optional) and also stratified by age group (Birth–5 years, 6-11 years, 12-18 years, 19-21 years (optional).

With a better understanding of follow up patterns after hospitalization for a mental health condition, health care organizations and policy makers can develop better informed services, health policy and planning for children with mental health conditions. Coordination of care is an emerging interest.

References:
American Academy of Child and Adolescent Psychiatry, American Psychiatric Association. Criteria for short-term treatment of acute psychiatric illness. 1997.

National Committee for Quality Assurance (NCQA). HEDIS 2015: Healthcare Effectiveness Data and Information Set. Vol. 1, narrative. Washington (DC): National Committee for Quality Assurance (NCQA); 2014. various p.

Numerator Statement: Whether or not follow up visits to a primary care clinician or a behavioral health clinician were delayed past 30 days after discharge from a qualifying hospitalization.
Denominator Statement: Hospital discharges of children from birth through their 21st birthday (0-21) discharged from an inpatient

visit in a mental health facility or from any facility with a primary mental health diagnosis.
Denominator Exclusions: Children who are not continuously enrolled in any a program reporting data available to the reporting or accountability entity for at least 180 days following the date of discharge.

Children who are re-admitted to any hospital on the day of discharge.

Measure Type: Process
Data Source: Administrative claims
Level of Analysis: Facility, Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : National, Population : Regional, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

- This is a process measure. Evidence for this measure should demonstrate that delays in mental health follow-up and in primary care follow-up after a child has been discharged from a mental health hospitalization leads to undesired health outcomes (or, alternatively, that prompt follow-up leads to desired outcomes). Preferably this evidence would be derived from a systematic review and grading of the quantity, quality, and consistency of the body of evidence examining the linkage between delayed mental health and primary care follow-up and those undesired health outcomes (or, alternatively, the linkage between prompt mental health and primary care follow-up and desired health outcomes).
- The undesired outcome(s) that would be avoided by prompt follow-up (or, alternatively, the desired outcomes that would be experienced due to prompt follow-up) was not specified by the developer in the diagram included in section 1a.3. However, in section 1a.8.2, the developer notes that continuity (i.e., mental health follow-up) impacts utilization and patient attitudes and cites one study that shows that "*A transition care-coordinator paradigm improves medical health constructs and can in fact save lives in medical settings*". The developer also refers to a 1997 guideline recommendation regarding the need for regular and timely assessments.
- The evidence for this measure is not based on a systematic review and grading of the empirical evidence. Instead, the developer conducted its own literature review, which was informed by parent focus groups and expert panelists who provided input on the development of the measure. The majority of the evidence summarized by the developer addresses that follow-up rates are modifiable; gaps in follow-up care; types of interventions; predictors of continuity versus the relationship of follow-up to improvement in the undesired outcomes. No evidence appears to be presented regarding the specific timeframe of 30 days for follow-up.
- Per NQF's Evidence Algorithm for a process measure with no systematic review, the highest eligible rating is MODERATE if the quality of evidence indicates a **high certainty** that benefits clearly outweigh undesirable effect. Without a systematic review, the evidence should be **high-moderate quality** and indicate **substantial net benefit**.

Questions for the Committee

- *Is the relationship of this measure to patient outcomes presented reasonable?*
- *Is the evidence directly applicable to the process of care being measured?*
- *How strong is the evidence for this relationship? Is the evidence high-moderate quality and is there substantial net benefit to justify a MODERATE rating?*

- *The measure specifies follow-up within 30 days, but no evidence appears to be summarized that specifically supports this timeframe versus others. Is the timeframe reasonable? Does the Committee wish to explore this further with the developer?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- Using data from the New York State Medicaid Managed Care program (year not stated), the developer reports that among children with mental health discharge, 7.52% had a delayed primary care follow-up, 23.37% had a delayed mental care follow-up, and 44.47% had delayed follow-up for both mental health and primary care.
- The developer also presents data indicating differences in timing of follow-up according to race/ethnicity (although the statistics shown are for 7 days, 21 days, and 60 days, but not for 30 days, as specified in the measure). The developer states its data indicate "convincing evidence that performance in this population differs by race and ethnicity."
- The developer reports it has analyzed socioeconomic status based on poverty in the home county of each child and that better performance was found in more wealthy counties, although no specific data were provided.
- The developer specified an approach to rurality/urbanicity in the home county of each child and reports better performance in large urban vs. small urbans vs. rural counties, although no specific data were provided.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- It is unclear what the causal pathway is that links evidence to improved outcomes from the "continuity" espoused by this measure. I can see very little actual indirect, let alone direct, evidence that follow-up as noted (while making sense from perhaps a general gestalt) is linked to positive outcomes. This said, I am not sure the authors really knew how to fill out the forms and link information concerning suicide, readmission, and resistance to treatment to follow-up.
- Evidence provided for this process measure is not based on systemic review and the empirical evidence is not graded. The developer used a parent focus group and expert panel to guide an independent literature review. Using the algorithm, the evidence presented would be moderate-low. Although the relationship of the measure to the patient outcomes (undesired outcomes avoided by timely follow-up) is reasonable more evidence is needed.

1b. Performance Gap.

- While there are differences, and appear to be important disparity differences, they are not summarized in an organized fashion.
- Data analysis was performed using New York State Medicaid Managed Care data from an unidentified year. The data does demonstrate a disparity showing that mental health admissions were higher in black and Caucasian children as compared to Hispanic children. The developer states that there is race/ethnicity differences in follow-up, but no specific data is presented. Further information is needed to determine if there is a gap that warrants a national performance measure. Also, data analysis of a broader payer mix would be beneficial.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- This measure is specified at the facility, health plan, and integrated health system levels of analysis, as well as at the population level (e.g., state), for use with administrative claims data. A lower score indicates higher quality.
- This measure captures children ages 0-21 years discharged from an inpatient mental health facility or from any hospital with a mental health primary diagnosis who do not have a mental health follow-up visit within 30 days or a primary care follow-up visit within 30 days during a one-year measurement period. Exclusions include patients who are re-admitted on the day of discharge and those who are not continuously enrolled for at least 180 days following discharge.
- Codes to identify patients discharged from a mental health hospitalization (ICD-9 codes, ICD-10 codes, place of service codes, revenue codes, CPT codes, and HCPCS codes) are provided (although all are not described in detail). Codes to identify patients with mental health or primary care follow-up (CPT codes, HCPCS codes) are provided. Providers considered "acceptable" for follow-up are described, but codes are not provided. Discharge status—used to identify exceptions—are described but codes are not provided.
- Both ICD-9 and ICD-10 codes are provided in an excel spreadsheet, and a description of the process used to identify ICD-10 codes is included.
- The calculation algorithm is detailed and should allow for consistent calculation of the measure.
- Overall results for this measure reflect delayed follow-up for mental health OR for primary care. However, the developer indicates results for should also be stratified to indicate delayed mental health follow-up only, delayed primary care follow-up only, or delay of both mental health and primary care follow-up.
- The developer also has encouraged stratification of measure results according to age group, race/ethnicity, urban/rural status, county poverty level, insurance type, and benefit type. To facilitate this stratification, the developer has included instructions on how to stratify. This stratification is meant to illuminate possible disparities in care and is not meant to serve as a method of risk-adjustment.

Questions for the Committee :

- *Can this measure be used at the facility (i.e., hospital) level?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

- It does not appear that empirical reliability testing was conducted. The developer states that " *Our findings and our standard errors suggest high signal to noise ratio of the measure as well as sensitivity to small differences in the specifications*". However, no information about the method of testing or the actual results are presented.
    - NQF guidance indicates that reliability testing can be conducted at the performance score level or at the data element level, or both.
        - Testing at the performance score level (for example, through a signal-to-noise analysis) will indicate whether there is enough variation across the measured entities (e.g., hospitals, states) , over and above that caused by random measurement error, to be able to distinguish among the measured entities. Although the developer refers to the standard errors of the measure, NQF does not consider an assessment of standard errors to be adequate reliability testing at the measure score level.
        - Testing at the data element level, when assessed in the same population in the same time period (e.g., through an analysis of inter-rater reliability), will indicate that it is possible to consistently collect data. Note that NQF guidance also indicates that if data element validity is demonstrated, reliability testing at the data element level is not required. In its discussion of validity, the developer does not provide any empirical data demonstrating data element validity.
    - Per the NQF Algorithm to evaluate the information provided by the developer for reliability: Only descriptive statistics computed (box 2) → no empirical validity testing data at the data element level).

Questions for the Committee:

- *Has the developer demonstrated through empirical reliability testing that differences in performance across*

measured entities (e.g., hospitals, states) can be identified?

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- In the evidence provided, there is no evidence related to the threshold (e.g., 7 days, 30 days, 60 days, etc.) to determine delayed follow-up. In section 2b2.2, the developer notes that the 30-day threshold specified in this measure was determined based on input from the expert panel who helped developer the measure.
- In its analyses on disparities, the developer examines 7, 21, and 60 days (but not 30 days, as specified) but does not link the timeframes to specific improvement.

Question for the Committee:

- *In the absence of empirical evidence from testing or evidence from the literature to support the 30-day threshold, do you agree that a 30-day threshold is a reasonable reflection of delayed follow-up?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

It is does not appear that empirical validity testing was conducted.

- NQF guidance indicates that validity testing can be conducted at the performance score level or at the data element level (for critical data elements) or both. Also, an assessment of the face validity of the **computed measure score** as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
  - NQF does not consider provision of description statistics (e.g., measure scores) to be adequate validity testing at the measure score level.
- Validity testing of data elements typically analyzes agreement with another authoritative source of the same information (e.g., what is reported in claims compared to what is included in the medical record, which is viewed as the gold standard). In section 2a2.2, the developer refers to studies validating data housed in administrative claims databases; however, no details of the methods, the results, or whether the studies assess all critical data elements are presented.
- Although the developer worked closely with an expert panel as part of the development process, there is no indication that a formal assessment of the face validity **of the computed measure score**, as required by NQF, was conducted.
- The developer provides some limited information about a comparison of the results of this measure and that of another mental health follow-up measure (NQF #0576). However, it is unclear what the comparison was meant to demonstrate and how the results should be interpreted.
- Per the NQF algorithm for validity testing,: Only descriptive statistics computed (box 3) and no face validity of the computed measure score were provided.

Questions for the Committee

- *Are you aware of evidence that administrative claims data accurately reflect mental health diagnoses and self-injury, suicide attempt, or suicidal ideation? NQF guidance permits citation to literature if all critical data elements from the measure have been validated by other sources.*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Children not continuously enrolled in a program for at least 180 days post-discharge are excluded, as are those who are re-admitted on the day of discharge.
- The developer did not provide information on the number of mental health discharges that were excluded from the measure.

Questions for the Committee

- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment:

- This measure is not risk-adjusted. As noted earlier, stratification by to age group, race/ethnicity, urban/rural status, county poverty level, insurance type, and benefit type is meant to illuminate possible disparities in care.
- In providing information about performance gap, the developer states its data indicate "convincing evidence that performance in this population differs by race and ethnicity." The developer also reports it has analyzed socioeconomic status based on poverty in the home county of each child and that better performance was found in more wealthy counties, although no specific data were provided. Lastly the developer specified an approach to rurality/urbanicity in the home county of each child and reports better performance in large urban vs. small urbans vs. rural counties, although no specific data were provided.

Questions for the Committee

- *Are the stratification variables appropriate?*
- *Given the data provided by the developer on differences in urbanicity, poverty, race and ethnicity, and SES (proxy of commercial vs. public insurance), does the Committee concur with the developer that risk adjustment for SDS is not appropriate?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer states it used chi square to assess statistical differences across different strata, but does not provide details as requested by NQF..
- The developer notes that meaningful differences were found between subpopulations in the testing data (CY2013 Medicaid claims data from NY State, for children with a mental health hospitalization), but no details are presented. The developer infers from this analysis that meaningful differences between measured entities (e.g., hospitals, states) would also be possible to identify.

Question for the Committee:

- *Can this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Because this measure has only one set of specifications (i.e., for claims data), this section is not applicable.

2b7. Missing Data

- The developer did not provide any information on missing data.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- This seems like a very "fruit cocktail" approach, mixing a wide group of patients with varying needs and risks. The exclusion of day of discharge follow-up to prevent gaming might punish a best practice.

2a2. Reliability testing

- I don't see evidence of reliability testing.
- The data elements are defined and ICD-9 and ICD-10 codes are provided, however empirical reliability testing was not conducted. It is possible that the measure be used at the facility level, but this may be challenging if the patient follow-up occurs outside the facility network (how will follow-up be monitored?).

2b1. Validity Specifications

- I don't see what is "magic" about a thirty day follow-up and this issue is non-trivial.

2b2. Validity Testing

- I don't see any evidence for validity testing. Certainly claims data will underestimate such concepts as suicidality/ideation and perhaps even suicide.
- Validity testing was not conducted. No face validity of the computed measure score were provided. Not aware of evidence that administrative claims data accurately reflect mental health diagnoses and self-injury, suicide attempt or suicidal ideation.

2b3-2b7. Threats to Validity

- Yes, the missing data constitute a threat to the validity of the measure.
- For above (no boxes for comments): Again, the exclusion of discharge day follow-up could penalize those systems most highly performing. There does not seem to be missing data considered in the attached documentation.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Data for the measure are obtained from administrative claims.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Probably can be done
- The data elements would be relatively available using administrative claims, but may be limited at the facility level.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- This is a new measure and is not currently in use.
- The developer states it is having initial conversations with partners about potential use of the measure for accountability applications, but no timeframes have been set. No details are provided about the partners.
- Information about improvement is not available, as this is a new measure.
- The developer did not identify any unintended negative consequences from the measure.

Questions for the Committee:

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Not clear--could enhance follow-up, but is that the important aspect of continuity? For example, if follow-up occurs in primary care but there is no timely communication of discharge meds and instructions, so what...
- As this is a new measure usability is yet to be determined. There does not seem to be any unintended consequences from the measure.

Criterion 5: Related and Competing Measures

- NQF staff identified that this measure is similar to NQF-endorsed 0576: Follow-up After Hospitalization for Mental Illness (NCQA). This new measure includes ages 0-21 years and the follow-up criteria appear to differ slightly; a code-by-code analysis would need to be performed by the developer. NQF 0576 reports two rates: percentage of discharges for which the patient received follow-up within 7 days and within 30 days of discharge.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: CAPQuaM PQMP Mental Health Follow Up Measure Timeliness 1: Delayed coordination of care following mental health discharge

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however,

serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Delay in mental health and primary care follow up visit following mental health discharge

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

This is a coordination of care measure that assesses which discharges from mental health admissions have delays in important follow up after discharge. The general conceptual model is that both continuity within the mental health care system and coordination with the primary care child health system are important to quality of

care. Our expert panel indicated that a timely MH follow up visit occurs on the 1-7th day after discharge and a timely PC visit occurs on days 1-21. For both PC and MH the panel established 30 days as the threshold for delayed care. (Figure 1)

Figure 1-Evidence: Simplified Conceptual Model Illustrating Continuity and Coordination

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ **Other – *complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

The evidence and measure resulted from CAPQuaM's peer-reviewed 360 degree method which in this case integrates a scoping literature review undertaken by a research librarian at Columbia University in collaboration with the CAPQuaM team, with the results of parent focus groups (conducted in Chicago) and conversations with our expert panelists and the expert panel findings themselves.

We conducted a two stage literature review, which began with an ad hoc review by CAPQuaM staff to orient ourselves to the literature and the topic. Round one was a targeted purposeful review performed by a key CAPQuaM investigator that was designed to be at the level of a graduate school term paper and intended to orient and inform the literature review and to provide an evidence base to guide our work pending the full scoping literature review. We called the product from round 1 the term paper.

For the Round 2 measure development, the original literature search conducted by a librarian at Columbia University resulted in 8,835 references that were not separated into mental health follow up and medication reconciliation, our two topics of interest at the time. The articles were first divided among pairs of reviewers (8 reviewers in total). Each pair of reviewers decided if each article was to be included or excluded and appropriate for mental health, appropriate for medication reconciliation, or for both. Results were merged into Excel and disagreements were discussed and resolved. The Mental Health library then had 920 articles. Two CAPQuaM staff then sorted all articles into topic areas based upon a hierarchical categorization list and excluded those reporting duplicate evidence or that were not informative. The Mental Health library then had 778 articles. Because of resource limitations, articles were prioritized using a 3 point rating scale. Articles rated in the lowest category by both reviewers based upon abstracts were excluded, leaving 653 articles that were reviewed and abstracted by the literature review team. We included evidence on children where possible and on adults and children where necessary

1a.8.2. Provide the citation and summary for each piece of evidence.

There is evidence that suggests that follow up rates are modifiable and that clinical characteristics may be associated with but are not dispositive of follow up rates (1-6). Various "bridging strategies" (7) that can be

effective range from telephone and letter prompting to various inpatient programmatic interventions aimed at discharge planning and linkage (2, 8-10) to involvement of the patient and treatment staff. (4, 11-13)

Continuity has been shown to impact a variety of health services factors including utilization, and patient attitudes about their illness, their hospitalization, and their subsequent ambulatory treatment. (14-19)

In addition to the constructs above, original studies from our group found the following performance gap in the late 1990's:

> In a Medicaid population in Massachusetts, there was poor coordination of care and incomplete communication from the mental health to the primary care system. A follow up with the primary care clinician could be documented 26% of the time within 30 days and 32.2% within 60 days. Among all mental health discharges, there was evidence in the chart that the PCP was aware of the mental health discharge only 46% of the time (n=242). Of those, 32% of the communication came directly from the patient and not another medical provider. Even among those who were seen in follow up, nearly one quarter did not show evidence of awareness of the MH discharge even after the "follow up" visit. The MassHealth medical director (herself a psychiatrist) estimated that in nearly every one of those admissions some form of medication change was made and lack of notation of awareness by the PCP was even more disturbing in that context.

Other interventions or approaches that have been examined in the literature includes attention to the physical proximity of services from different providers: for example a "medical home" paradigm applied specifically to mental health constructs, was shown to foster continuity of care in a controlled study. (20)

The importance of primary care coordination and of continuity are supported in the literature. A transition care-coordinator paradigm improves medical health constructs and can in fact save lives in medical settings. (21) Bates and Bitton April 2010 Health Affairs remind us that transitions are a vulnerable time for patients, concluding that "Hospitals need to let medical homes know when their patients leave, and medical homes need processes to contact these patients for follow-up… practices need electronic tools to assist with medication reconciliation, the process of identifying and updating the complete list of medications the patient is taking. One group is evaluating a tool that enables primary care providers to call up a patient's medication list at discharge and rapidly compare it to the electronic medication list that existed before admission."(22) Med Rec is a key aspect of follow up and this was supported by our expert panel.

The medical home paradigm is less available to youth with mental health problems, another performance gap. (23)

Despite workforce limitations, there is ample evidence that follow up is a manageable and consequential process of care and some institutions and systems do it better than others.

Gender, age, race, type of admission diagnosis, urban vs. other settings all seem to be predictors of continuity of care. Fragmented care for inner-city minority children with ADHD, system and human level factors that were perceived to impede coordination of care, need for better organizational policies that define provider responsibilities and accountability are all major issues. There is a need to support the coordination of care and

provide additional education and resources to improve collaboration. (14) This justifies our approach to stratification.

Follow-up is a key component of the optimal management of any number of medical conditions, but is especially critical for children with mental health diagnoses. Timely follow up with both primary care clinicians and mental health practitioners after a hospital discharge are imperative to deliver the best outcomes. According to a guideline developed by the American Academy of Child and Adolescent Psychiatry (AACAP) and the American Psychiatric Association (APA) (1997), there is a need for regular and timely assessments and documentation of the patient's response to all treatments. When considering follow up care, the literature makes the distinction between coordination across systems of care (in this case the primary care system and the mental health system), and continuity (in this case within the mental health system)." (24-25)

REFERENCES

1. Kirk SA. Who gets aftercare? A study of patients discharged from state hospitals in Kentucky. Hosp Community Psychiatry. Feb 1977;28(2):109-114.

2. Axelrod S, Wetzler S. Factors associated with better compliance with psychiatric aftercare. Hosp Community Psychiatry. Apr 1989;40(4):397-401.

3. Tessler R, Mason JH. Continuity of care in the delivery of mental health services. Am J Psychiatry. Oct 1979;136(10):1297-1301.

4. Fink EB, Heckerman CL. Treatment adherence after brief hospitalization. Compr Psychiatry. Jul-Aug 1981;22(4):379-386.

5. Leaf PJ, Livingston MM, Tischler GL, Weissman MM, Holzer CE, 3rd, Myers JK. Contact with health professionals for the treatment of psychiatric and emotional problems. Med Care. Dec 1985;23(12):1322-1337.

6. Mechanic D, Angel R, Davies L. Risk and selection processes between the general and the specialty mental health sectors. J Health Soc Behav. Mar 1991;32(1):49-64.

7. Meyerson AT, Herman GS. What's new in aftercare? A review of recent literature. Hosp Community Psychiatry. Apr 1983;34(4):333-342.

8. Bogin DL, Anish SS, Taub HA, Kline GE. The effects of a referral coordinator on compliance with psychiatric discharge plans. Hosp Community Psychiatry. Jul 1984;35(7):702-706.

9. Stickney SK, Hall RC, Garnder ER. The effect of referral procedures on aftercare compliance. Hosp Community Psychiatry. Aug 1980;31(8):567-569.

10. Wolkon GH, Peterson CL, Rogawski AS. A program for continuing care: implementation and outcome. Hosp Community Psychiatry. Apr 1978;29(4):254-256.

11. Rosenfield S, Caton C, Nachumi G, Robbins E. Closing the gaps: the effectiveness of linking programs connecting chronic mental patients from the hospital to the community. J Appl Behav Sci. 1986;22(4):411-423.

12. Olfson M, Mechanic D, Boyer CA, Hansell S. Linking inpatients with schizophrenia to outpatient care. Psychiatr Serv. Jul 1998;49(7):911-917.

13. Sullivan K, Bonovitz JS. Using Predischarge Appointments to Improve Continuity of Care for High-Risk Patients. Hosp Community Psych. 1981;32(9):638-639.

14. Kaizar E, Chisolm D, Seltman H, Greenhouse J, Kelleher KJ. The role of care location in diagnosis and treatment of pediatric psychosocial conditions. J Dev Behav Pediatr. Jun 2006;27(3):219-225.

15. Kazak AE, Hoagwood K, Weisz JR, et al. A Meta-Systems Approach to Evidence-Based Practice for Children and Adolescents. Am Psychol. Feb-Mar 2010;65(2):85-97.

16. Wissow LS, Brown JD, Krupnick J. Therapeutic alliance in pediatric primary care: preliminary evidence for a relationship with physician communication style and mothers' satisfaction. J Dev Behav Pediatr. Feb-Mar 2010;31(2):83-91.

17. Mahajan P, Thomas R, Rosenberg DR, et al. Evaluation of a child guidance model for visits for mental disorders to an inner-city pediatric emergency department. Pediatr Emerg Care. Apr 2007;23(4):212-217.

18. Leaf PJ, Livingston MM, Tischler GL, Weissman MM, Holzer CE, 3rd, Myers JK. Contact with health professionals for the treatment of psychiatric and emotional problems. Med Care. Dec 1985;23(12):1322-1337.

19. Simms MD, Dubowitz H, Szilagyi MA. Health care needs of children in the foster care system. Pediatrics. Oct 2000;106(4 Suppl):909-918.

20. Annunziato RA, Rubinstein D, Sheikh S, et al. Site matters: winning the hearts and minds of patients in a cardiology clinic. Psychosomatics. Sep-Oct 2008;49(5):386-391.

21. Annunziato RA, Shemesh E. Tackling the spectrum of transition: what can be done in pediatric settings? Pediatr Transplant. Nov 2010;14(7):820-822.

22. Bates DW, Bitton A. The future of health information technology in the patient-centered medical home. Health Aff (Millwood). Apr 2010;29(4):614-621.

23. Adams SH, Newacheck PW, Park MJ, Brindis CD, Irwin CE, Jr. Medical home for adolescents: low attainment rates for those with mental health problems and other vulnerable groups. Acad Pediatr. Mar 2013;13(2):113-121.

24. American Academy of Child and Adolescent Psychiatry, American Psychiatric Association. Criteria for short-term treatment of acute psychiatric illness. 1997.

25. National Committee for Quality Assurance (NCQA). HEDIS 2015: Healthcare Effectiveness Data and Information Set. Vol. 1, narrative. Washington (DC): National Committee for Quality Assurance (NCQA); 2014. Various p.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

evidence_attachment_-_MHFU_v3-635793142552381649.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

Pediatric mental health hospitalizations have increased 24% during 2007 – 2010. In total, US children spent 1,721,765 days in hospitals for mental health care in 2012. Recent estimates put the cost of mental health hospitalization of children at $11.6 billion between 2006 and 2011. This analysis also found mental health admissions were higher among black and white children compared with Hispanic children, and were more common for children with public insurance than private or no insurance. Additionally, children who are admitted to the hospital for a mental health condition are very likely to meet criteria for Children with a special health care need.

Follow-up is a key component of the optimal management of any number of medical conditions, but is especially critical for children with mental health diagnoses. Timely follow up with both primary care clinicians and mental health practitioners after a hospital discharge are imperative to deliver the best outcomes. According to a guideline developed by the American Academy of Child and Adolescent Psychiatry (AACAP) and the American Psychiatric Association (APA) (1997), there is a need for regular and timely assessments and documentation of the patient´s response to all treatments. When considering follow up care, the literature makes the distinction between coordination across systems of care (in this case the primary care system and the mental health system), and continuity (in this case within the mental health system)."

Therefore, this measure looks at continuity of care as a component of coordination of care and assess follow up appointments both within and outside of the mental healthcare system. This measure reflects one aspect of coordination of care following the discharge of a child who has been hospitalized for a primary diagnosis that is specified as being a mental health diagnosis. This measure describes key attributes regarding the timeliness of follow up after a mental health discharge for children. Specifically, this measure set looks at the failure to establish timely follow up care subsequent to the day of discharge in both the primary care and mental healthcare systems. Stratification by type of failure enhances the granularity of the measure and provides clear data to support improvement initiatives. The measure is specified to be reported as an aggregate for the included age groups (Birth-21 years, with 19-21 optional) and also stratified by age group (Birth–5 years, 6-11 years, 12-18 years, 19-21 years (optional).

With a better understanding of follow up patterns after hospitalization for a mental health condition, health care organizations and policy makers can develop better informed services, health policy and planning for children with mental health conditions. Coordination of care is an emerging interest.

References:

American Academy of Child and Adolescent Psychiatry, American Psychiatric Association. Criteria for short-term treatment of acute psychiatric illness. 1997.

National Committee for Quality Assurance (NCQA). HEDIS 2015: Healthcare Effectiveness Data and Information Set. Vol. 1, narrative. Washington (DC): National Committee for Quality Assurance (NCQA); 2014. various p.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

We have analyzed performance in New York State Medicaid Managed Care and found in a recent year that 74.4% of more than

13,000 mental health discharges had delayed follow up in either primary care or mental health appointments. Stratified, there were 7.52% with delayed Primary Care Follow up only, 22.37% with delayed Mental Health follow up only and 44.47% with delays in both. There were 13,692 discharges. The SEs of the estimates are 0.37%, 0.23%, 0.42%, and 0.36% respectively.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Race/Ethnicity
We use existing data to describe race/ethnicity. We have specified these data to be stratified by race/ethnicity. In New York State we were capable of identifying differences across a variety of measures assessed during development.

For example, among discharges assessed among Blacks (N=3210), Whites (N=4290) and Hispanics (N=3633) 0-21 years of age, using slightly different specifications we found:
Timely follow up varied as follows:

MH follow up within 7 days:
Black 15.9%, White 13.6%, and Hispanic 21.5%
PC follow up within 21 days:
Black 6.8%, White 8.8%, and Hispanic 10.5%

Delays in follow up beyond 60 days varied as follows:
MH no follow up within 60 days:
Black 66.3%, White 67.2%, and Hispanic 59.5%
PC no follow up within 21 days:
Black 85.3%, White 82.5%, and Hispanic 79.2%

These data offer convincing evidence that performance in this population differs by race and ethnicity.

We will have updated data for the current specifications at time of our revision to submission:
Any delay: Blacks xx (xx), Whites xx(xx), Hispanics xx (xx), all others
PCFU only: Blacks xx (xx), Whites xx(xx), Hispanics xx (xx), all others
MH FU only: Blacks xx (xx), Whites xx(xx), Hispanics xx (xx), all others
Both: Blacks xx (xx), Whites xx(xx), Hispanics xx (xx), all others

Socioeconomic Status
We have specified an approach to looking at poverty in the home county of each child. In NY State data analyses of our measures were sensitive to differences in the three categories that are present in NY State. Values were more favorable in more wealthy counties.

Rurality/Urbanicity
We have specified an approach to looking at the rurality/urbanicity in the home county of each child. In NY State data analyses of our measures were sensitive to differences in the three categories that are present in NY State. Performance was more favorable in large urban as compared to small urban compared to rural counties.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
Data provided in 1b.4.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
Our analysis identified 13,692 hospital discharges in one year in New York State Medicaid Managed Care (MMC). 183,000 Medicaid children saw a clinician for a medical visit or admission that included at least one mental health diagnosis and 132,000 of them were in MMC. Mental health discharges are common and occur at a rate of about 1 per 10 children with mental health diagnosis in this population.

Mental health is a critical component in the development of a child's emotional and physical well-being. In 2009, nearly 10% of pediatric hospitalizations were for a primary mental health diagnosis with depression, bipolar disorder, and psychosis as the most frequent reasons [1]. Pediatric mental health hospitalizations have increased 24% during 2007 – 2010 [1]. Our analysis of discharges from general and children's hospitals in the US with a primary mental health diagnosis using the 2012 Kids' Inpatient Database (KID) found that mood disorders accounted for 55% of primary diagnoses, followed by psychotic (9%) and substance abuse disorders (8%). In total, US children spent 1,721,765 days in hospitals for mental health care in 2012. This analysis also found mental health admissions were higher among black and white children compared with Hispanic children, and were more common for children with public insurance than private or no insurance. Recent estimates put the cost of mental health hospitalization of children at $11.6 billion between 2006 and 2011.[2] Our analysis suggested a particular burden for Medicaid. We note additionally that children who are admitted to the hospital for a mental health condition are very likely to meet criteria for CSHCN and hence this measure set is of importance for this population of interest.

Follow-up is a key component of the optimal management of any number of medical conditions, but is especially critical for children with mental health diagnoses. Timely follow-up with both primary care providers and mental health practitioners after a hospital discharge are imperative to deliver the best outcomes. There is broad acceptance that follow up may also reduce re-hospitalizations and associated costs. Still, the capacity (facilities and clinicians) needed to provide follow-up for children with mental health diagnoses remain insufficient. In Massachusetts, one study found that 80% of pediatricians reported that their patients struggled to find mental health services. [3] Our project's focus groups with parents of children with mental illness indicated the burden on parents to identify and secure outpatient mental health services is substantial, including for children with private insurance, and that clinical resources are scarce. Children with mental health issues are more vulnerable to incomplete follow-up because of a lack of available services. The challenges are increased because care coordination for pediatric mental health patients is made more complex by a variety of issues such as the potential for stigma, frequent involvement of one or both of the school and juvenile justice systems, the frequent involvement of child protective services, and the potential for concomitant substance abuse. [4] Clinically, complexity is added by the particular reluctance of some mental health professionals to share information even within the clinical team. [5, 6]

Follow-up after discharge of a hospitalized mental health patient is a current National Committee for Quality Assurance (NCQA) quality measure in the Health Effectiveness Data and Information Set (HEDIS) that tracks the percentage of patient appointments with a mental health practitioner. The current HEDIS calculates the percentage of members who received a follow-up at 7 and 30 days after discharge for patients over 6 years of age. Follow-up under HEDIS guidelines can occur as an outpatient visit, an intensive outpatient encounter or partial hospitalization with a mental health practitioner.

This measure has several important merits including its comprehensive definition of a mental health condition. However we were assigned by AHRQ and CMS to update the measure and to optimize it for the child health setting. As there are no provisions to assess follow up outside of the mental health system, such as with primary care providers, this represented one clear opportunity for enhancement.

CAPQuaM builds from the current HEDIS measure by looking at continuity of care as a component of coordination of care by

comparing follow-up appointments both within and outside of the mental healthcare system. The proposed measure is for children and adolescents only, extends the pediatric age range included, and is specified to be stratified by age.

Barbara Starfield defined primary care as "that level of a health service system that provides entry into the system for all new needs and problems, provides person-focused (not disease-oriented) care over time, provides care for all but very uncommon or unusual conditions, and coordinates or integrates care provided elsewhere by others." [20] Coordination of care sees the primary care practice as integrating all aspects of its patients' care, even when being seen elsewhere. [21] This coordination is especially important for those children with special healthcare needs (mental health conditions included) and has become a key aspect in the medical home model, which strives to provide a single point of care from which all other health care services can be integrated. [22] Coordination of care implies continuity, but continuity can happen with only minimal or inadequate coordination and is not sufficient to qualify as meaningful, high-quality care. The submitted measure set uses follow-up visits to a mental health provider to signify continuity of care and follow-up with a primary care provider to signify coordination of care.

Children with mental health diagnoses comprise a critically important population of high interest to Medicaid. According to one report conducted by the Center for Health Care Strategies, less than 10% of children in Medicaid utilize behavioral health care, but behavioral health care accounts for 38% of Medicaid expenditures for children. [23] Furthermore, one third of the Medicaid child population utilizing behavioral healthcare is in the foster care system. These children represent 56% of the total behavioral health expenses for all children enrolled in Medicaid. Our analysis of both the National Survey of Children's Health data (NSCH, 2011/12), and of the 2012 Kids' Inpatient Database (KID) confirmed the importance of mental healthcare in the Medicaid population. The analysis of the NSCH data estimates that approximately 5.2 million children between the ages of 0-17 years old in the U.S. have been told that they have an emotional, behavioral, or developmental issue. Fifty six percent of these children are of low income and have public insurance. Three out of every 1,000 mental health hospital admissions are children with public insurance.

We have done systematic and iterative analyses to assess various approaches to identifying and counting hospital admissions for children with a mental health diagnosis using New York State Medicaid data, resulting in the current specification as most sensitive while retaining appropriate selectivity. In 2013, we identified 14,488 inpatient discharges for children 0-21 in New York State Medicaid, of which more than 11,000 were in children 0-18.

1c.4. Citations for data demonstrating high priority provided in 1a.3

1.Bardach, N.S., et al., Common and costly hospitalizations for pediatric mental health disorders. Pediatrics, 2014. 133(4): p. 602-9.
2.CM Torio, W.E., T Berdahl, MC MCCOrmich, LA Simpson, Annual Report on Health Care for Children and Youth in the United States: National Estimates of Cost, Utilization and Expenditures for Children With Mental Health Conditions, in Pediatrics. 2014. p. 19-35.
3.Perrin, E.C. and R.C. Sheldrick, The challenge of mental health care in pediatrics. Arch Pediatr Adolesc Med, 2012. 166(3): p. 287-8.
4.Kazak, A.E., et al., A Meta-Systems Approach to Evidence-Based Practice for Children and Adolescents. American Psychologist, 2010. 65(2): p. 85-97.
5.Weiss, A.P., Special protections for mental health treatment notes. Virtual Mentor, 2012. 14(6): p. 445-8.
6.Coffey, R.M., et al., Transforming mental health and substance abuse data systems in the United States. Psychiatr Serv, 2008. 59(11): p. 1257-63.
7.Kirk, S.A., Who gets aftercare? A study of patients discharged from state hospitals in Kentucky. Hosp Community Psychiatry, 1977. 28(2): p. 109-14.
8.Axelrod, S. and S. Wetzler, Factors associated with better compliance with psychiatric aftercare. Hosp Community Psychiatry, 1989. 40(4): p. 397-401.
9.Wolkon, G.H., Characteristics of clients and continuity of care into the community. Community Ment Health J, 1970. 6(3): p. 215-21.
10.Tessler, R. and J.H. Mason, Continuity of care in the delivery of mental health services. Am J Psychiatry, 1979. 136(10): p. 1297-1301.
11.Meyerson, A.T. and G.S. Herman, What´s new in aftercare? A review of recent literature. Hosp Community Psychiatry, 1983. 34(4): p. 333-42.
12.Bogin, D.L., et al., The effects of a referral coordinator on compliance with psychiatric discharge plans. Hosp Community Psychiatry, 1984. 35(7): p. 702-6.
13.Stickney, S.K., R.C. Hall, and E.R. Garnder, The effect of referral procedures on aftercare compliance. Hosp Community Psychiatry, 1980. 31(8): p. 567-9.
14.Wolkon, G.H., C.L. Peterson, and A.S. Rogawski, A program for continuing care: implementation and outcome. Hosp Community Psychiatry, 1978. 29(4): p. 254-6.
15.Fink, E.B. and C.L. Heckerman, Treatment adherence after brief hospitalization. Compr Psychiatry, 1981. 22(4): p. 379-86.
16.Sullivan, K. and J.S. Bonovitz, Using predischarge appointments to improve continuity of care for high-risk patients. Hosp Community Psychiatry, 1981. 32(9): p. 638-9.
17.Rosenfield, S., et al., Closing the gaps: the effectiveness of linking programs connecting chronic mental patients from the hospital

to the community. J Appl Behav Sci, 1986. 22(4): p. 411-23.
18.Olfson, M., et al., Linking inpatients with schizophrenia to outpatient care. Psychiatr Serv, 1998. 49(7): p. 911-7.
19.Kazak, A.E., et al., A meta-systems approach to evidence-based practice for children and adolescents. Am Psychol, 2010. 65(2): p. 85-97.
20.Starfield, B., Primary Care: Balancing Health Needs, Services, and Technology 1998: Oxford University Press.
21.Starfield, B., L. Shi, and J. Macinko, Contribution of primary care to health systems and health. Milbank Q, 2005. 83(3): p. 457-502.
22.Stille, C.J. and R.C. Antonelli, Coordination of care for children with special health care needs. Curr Opin Pediatr, 2004. 16(6): p. 700-5.
23. Pires, S.G., KE; Allen, KD; Gilmer, T; Mahadevan, RM, Examining Children´s Behavioral Health Service Utilization and Expenditures in Faces of Medicaid, C.o.H.C.S. Inc., Editor. 2013.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Mental Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access, Care Coordination, Disparities, Safety, Safety : Medication Safety, Safety : Readmissions

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
We currently do not have a web page. We will ensure that this measure will be publicly available.

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** FINAL_CAPQuaM_MHFU_ICD_Conversion.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Whether or not follow up visits to a primary care clinician or a behavioral health clinician were delayed past 30 days after discharge from a qualifying hospitalization.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back*

to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)
The data are reported for one year (the year of discharge) and also require the 6 months following the reporting year for assessment.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The numerator is the number of discharges for which the first follow up visit to a mental health clinician does not occur in days 1 to 30 following discharge OR for which the first follow up to a primary care clinician does not occur in same time period. The day of discharge is considered Day 0.

The measure is further stratified as:

a. Delayed receipt of initial mental health follow up visit (percent first follow up visit with MH clinician after day of discharge is > 30 days, ONLY;

b. Delayed receipt of initial primary care follow up visit (percent first follow up visit with PC clinician after day of discharge is > 30 days, ONLY;

c. Delayed receipt of both primary care and mental health follow up visits (no visits to MH AND no visits to PC Clinicians in days 1-30).

Definitions of how to identify follow up and clinician types can be found in the appendix, particularly Tables 5-7.

Our online specifications incorporate ICD-9 codes only. For the specified ICD-10 codes and a detailed listing of ICD 9 codes see attached spreadsheet in S2.b.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Hospital discharges of children from birth through their 21st birthday (0-21) discharged from an inpatient visit in a mental health facility or from any facility with a primary mental health diagnosis.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Denominator Elements:

1) Age of the child
2) Evidence of qualifying discharges using the specified mix of ICD9, CPT, HCSPCS, Revenue, and POS codes.
3) Discharge status (alive, not transferred to inpatient facility)
4) Date of discharge

For stratifications and at the option of the accountability entity:

5) County of residence of the caregiver
6) Race/ethnicity
7) Insurance type
8) Benefit type

All children from birth through their 21st birthday (optionally 18 at the preference of the accountability entity) who are:

- Discharged from an inpatient hospitalization with either a primary mental health ICD9 diagnosis (Primary Diagnosis 290xx through 314xx and 316xx) or any primary diagnosis with a V (62.84) or an E (950xx-959xx) code indicating self-injury, suicide

attempt, or suicidal ideation

OR

- Discharged for any diagnosis from place of service 51, 55, or 56.

Detailed specifications and algorithm for identifying discharges is shown in the appendix in Tables 1-4.

These details incorporate ICD-9 codes only. For the specified ICD-10 codes and a detailed listing of ICD 9 codes see attached spreadsheet in S2.b.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Children who are not continuously enrolled in any a program reporting data available to the reporting or accountability entity for at least 180 days following the date of discharge.

Children who are re-admitted to any hospital on the day of discharge.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Exclude if three are not six months (180 days) of continuous enrollment from the date of index discharge. For adolescents and children in Medicaid, the specific plan may be changed so long as Medicaid eligibility is continuous. For private insurers, the continuity should be within the health plan or across health plans where an all payer data base is available. The exclusion relates solely to the availability of data rather than attribution. As this is a measure of coordination, "If a plan touches a patient in the time frame it owns them" for this measure.

Exclude from the measure any otherwise qualified discharge for which there is a readmission that meets inclusion criteria on the identical date as the date of discharge.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
The top line for the measure is the number that have delay in follow up with primary care clinicians or mental health clinicians, inclusive. The measure also should be reported stratified by those who have delay:
a. Delay only in seeing mental health clinician
b. Delay only in seeing primary care clinician
c. Delay in seeing primary care clinician and mental health clinician.

All should be reported as percent of eligible discharges (to 2 digits)

The measure and stratifications are intended to be reported overall and stratified by age group: children before their sixth birthday (0-5), children from their sixth birthday and prior to their 12th birthday, children from their 12th birthday and prior to their 19th birthday, and from their 19th birthday until their 21st birthday. For this measure set the age of record is the child or adolescent's age at the date of discharge.

Categorize by age group: children before their sixth birthday (0-5), children from their sixth birthday and prior to their 12th birthday, children from their 12th birthday and prior to their 19th birthday, and (if included) from their 19th birthday until their 21st birthday. For this measure the age of record is the child or adolescent's age at the date of discharge.

Additional stratifications should be reported by the following variables:
Race/ethnicity, urban influence, level of poverty in the caregiver's county of residence, insurance type, and benefit type. ZIP code data (or county FIPS code if zip not available) are used to derive the urban influence and level of poverty variables.

To create other stratification variables:

i. Identify County equivalent of child's residence (based upon primary caregiver). If County and State or FIPS code are not in the administrative data, the zip codes can be linked to County indirectly, using the Missouri Census Data Center

(http://mcdc.missouri.edu/). These data will link to County or County equivalents as used in various states.

ii. Identify the Urban Influence Code (1) or UIC for the county of child's residence. (2013 urban influence codes available at: http://www.ers.usda.gov/data-products/urban-influence- codes.aspx#.UZUvG2cVoj8). Use one of two schema to identify rurality/urbanicity if desired. The former differentiates better various rural communities, while the latter better differentiates different urban settings. One may incorporate aspects of both as shown in C. Depending on the setting and interests of the accountability entity, all rural areas may be aggregated, although this should not be done to obscure findings in frontier areas:

a.After Bennett et al (SC Rural research Center):
i.UIC 1 & 2 are classified as Urban
ii.UIC 3,5,& 8 as micropolitan Rural
iii.UIC 4,6,& 7 Rural Adjacent to a metro area
iv.UIC 9-12 remote rural

b.Modified after Hart (UND Center for Rural Health)
i.UIC 1 Large Urban
ii.UIC 2 Small Urban
iii.UIC 3-8 Rural
iv.UIC 9-12 remote rural (may be used to approximate frontier)

c.Modified integrated approach:
i.UIC 1 Large Urban
ii.UIC 2 Small Urban
iii.UIC 3,5,& 8 as micropolitan Rural
iv.UIC 4,6,& 7 Rural Adjacent to a metro area
v.UIC 9-12 remote rural

iii.Identify the Level of Poverty in the caregiver's county of residence. The percent of all residents in poverty by county or county equivalent are available from the US Department of Agriculture at http://www.ers.usda.gov/data-products/county-level-data-sets/download-data.aspx . Our stratification standards are based on 2011 US population data that we have analyzed with SAS 9.3. Using Mother's state and county of residence (or equivalent) or FIPS code, use the variable PCTPOVALL_2011 to categorize into one of 5 Strata:
a.Lowest Quartile of Poverty if percent in poverty is <=12.5%
b.Second Quartile of Poverty if percent in poverty is >12.5% and <=16.5%
c.Third Quartile of poverty if percent in poverty is >16.5% and <=20.7%
d.First Upper Quartile (75th-90th) if percent in poverty is >20.7% and <=25.7%
e.Second Upper Quartile (>90th percentile)

iv.Categorize Race/Ethnicity as Hispanic, Non-Hispanic White, Non- Hispanic Black, Non-Hispanic Asian/Pacific Islander, and Non-Hispanic Other

v.Categorize Insurance Type as Private (Commercial), Public, None or Other

vi.Categorize benefit type as HMO, PPO, FFS, PCCM, or Other

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate

worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
N/A

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Identify the admission and discharge dates of all hospital discharges that occurred in the reporting year for children in the eligible age groups, using the Codes indicated in Appendix 1: Tables Admit 1 – Admit 4. Qualification using any one of the tables is sufficient. That is eligibility may be qualified via Table Admit 1 OR Table Admit 2 OR Table Admit 3 OR Table Admit 4. Admissions should be reviewed and de-duplicated. The basic unit of analysis is the hospital discharge, so children with multiple admissions should be included in the measure distinctly for each admission.

Tables 1-4 are described here:
Table 1 defining admission on the basis of CPT codes, place of service, diagnosis, and dates of admission and discharge

Table 2 defining admission on the basis of HCPCS codes, place of service, diagnosis, and dates of admission and discharge

Table 3 defining admission on the basis of Revenue codes, place of service, diagnosis, and dates of admission and discharge

Table 4 defining admission on the basis of POS codes and dates of admission and discharge

Step 2: Group all discharges by patient in chronological order. Remove from the measure any otherwise qualified discharge for which there is an admission for any inpatient hospitalization for which admission is on the date of discharge and discharge of the subsequent hospitalization is on a later date.

Step 3: Confirm that all remaining discharges were for children and adolescents who have not had at least one eligible mental health discharge (MHD) in the reporting year.

Step 4: Create Denominator 1: Consistent with the above, eliminate all MHD for children who are not enrolled continuously (continuous enrollement criterion represents the need for the capacity for data capture and is not specific to any health plan) for 180 or more days after the MHD. Consider the day of discharge to be Day 0. Eliminate all MHD for which the child/adolescent is readmitted for an MHD on Day 0. The Denominator should be created for all age groups (0 through 18, or 0 to 21) and for each age stratum (0-5, 6-11, 12-18, 19-21). Other such stratifications as specified herein and requested by the accountability agency.

Step 5: Create MHFU numerators. Qualifying events include specified outpatient or inpatient mental health visits, as shown in the POS Table (Appendix - Table 5) and algorithm. For each numerator, create an appropriate flag in the record regarding qualification status and another variable reporting day post admission of the qualifying event:

Table 5: POS Class defines inpatient and outpatient codes

a.Identify qualifying outpatient visits (per POS table) and the days after discharge of the event. Search from post-hospital Day 1 (not including Day 0) forward to 180 days. Use provider types as used in the data set to identify those visits that were to specified MH clinicians or to PC clinicians. For all that qualify, please identify the first MH qualified and the first PH qualified that follow the discharge and record for each the day following discharge on which it occurred. The first of each these types of visits is considered the initial. Record the day post discharge on which each visit occurred. The numerator is satisfied if either the initial primary care

visit or initial mental health visit occurred after day 30 or if no such visit was identified. For each discharge in the numerator identify whether it qualified because of mental health alone or primary care alone or both.

Please note that the Table of MH providers (Appendix 1: Table 6) is more specific than for the HEDIS measures and is optimized for child health.

PC Clinicians are specified to include pediatricians (including medicine-pediatrics physicians), adolescent medicine physicians, family physicians, internists, and advance practice nurses working with any of these.(Appendix 1: Table 7)

Table 6: List of MH Clinicians
Table 7: List of PC Clinicians

Step 6: Calculate and report the measure as described below:

Report Denominator 1's value as "N" for the measure and each Stratum reported.

II.Delayed coordination of care following mental health discharge. Report percent to 2 digits.

a.Delayed receipt of initial mental health follow up visit (percent first follow up visit with MH clinician > 30 days); OR
b.Delayed receipt of initial primary care follow up visit (percent first follow up visit with PC clinician > 30 days);
c.Stratify all numerator events as follows (percent of all discharges):
i.Meets criterion a only: Delayed coordination of care with mental health clinician
ii.Meets criterion b only: Delayed coordination of care with primary care clinician
iii.Meets criterion a and b: Delayed coordination and continuation of care

Step 7: Create stratification variables, as specified above and as requested by the accountability entity.

Step 8: As requested by accountability entity, describe variability as 95% confidence intervals. Recall that proportions are percents divided by 100. The CI is found as the mean percent plus or minus the product 196*[Square root of the [quotient of the (proportion meeting criterion) multiplied by (the proportion not meeting the criteria) divided by Denominator 1].

Step 9: Repeat Steps as needed to describe findings by strata—Age category, Race/Ethnicity, UIC or urbanicity, County Poverty Level, Insurance Type, and Benefit Type. Report by Race/Ethnicity within Age strata and repeat that analysis by UIC, and also by County Poverty Level. Report by Insurance Type and Benefit Type within Race/Ethnicity. Additional Cross tabulations are supported by these specifications and may be requested by an accountability entity.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Inclusion criteria require the availability of data to identify qualifying discharges and including the 180 day period following discharge.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.

Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The preferred data source is a comprehensive encounter and/or billing administrative database, along with enrollment data. Should a comprehensive database not be available, a combination of a discharge abstract and an ambulatory data abstract can be used as an alternate.

General data elements include:
-Age
-Race and ethnicity
-Insurance type (Medicaid, Private, Other)
-ICD9, CPT, Revenue, and Place of Service codes
-Benefit type among insured (HMO, PPO, FFS, Medicaid Primary Care Case Management Plan [PCCM], Other)
-ZIP code or State and County of residence and FIPS where available
-Enrollment status
-Provider type

Administrative data with billing (procedure) codes, diagnosis codes, place of service codes, revenue codes, and provider type codes are used to identify:
-Eligibility, which requires a hospital discharge for a mental health condition as specified;
-Qualifying numerator events such as:
  oOutpatient visits to a mental health clinician;
  oOutpatient visits to a primary care clinician;
  oSpecified mental health readmissions.
-Potentially disqualifying events, such as:
  oSpecified mental health readmissions;
  oSpecified non-mental health hospital admissions.
-Date of service should be recorded for all relevant services.
-Insurance benefit type.
-ZIP code or State and County of residence and FIPS where available.
-Race and ethnicity (from hospital administrative data or charts if not in administrative data from plan).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : National, Population : Regional, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic, Behavioral Health/Psychiatric : Inpatient, Behavioral Health/Psychiatric : Outpatient, Hospital/Acute Care Facility, Other, Post Acute/Long Term Care Facility : Inpatient Rehabilitation Facility, Post Acute/Long Term Care Facility : Long Term Acute Care Hospital
If other: Coordination of inpatient and ambulatory care relevant for behavioral health and primary care. Health plan; Integrated delivery system

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
CAPQuaM_NQF_Testing_submission_form_-mhfu_final.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: CAPQuaM PQMP Mental Health Follow Up Measure Timeliness 1: Delayed coordination of care following mental health discharge
Date of Submission: 9/30/2015
Type of Measure: Coordination of Care

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

We used Medicaid claims predominantly in our testing.

Our cross sectional study of the 2012 KID database found the rate of pediatric hospitalizations for mental disorders in 2012 was 2.96 per 1000 children, representing 4% (257,882) of total pediatric hospitalizations. Expected variation in admission rate by age group was seen: 0.13, 1.04, 5.36 and 7.49 (P<.001) per 1000 in children less than 6, 6-11, 12-18 and 19-20 years old, respectively. Admissions were most common in children with public insurance (3.0 per 1000), compared to private insurance (2.0 per 1000) and those without insurance (1.0 per 1000), P<.001. Median length of stay was 4.2 (IQR 2.3-6.8) days. Children in US spent 1,721,765 days in hospitals for mental health care in 2012. An approximately equal number of children were diagnosed primarily for physical health disorders who also had a mental health diagnosis noted, highlighting the critical importance of coordinating care across the MH and primary care systems to optimize integrated care for children. This justifies choice of measure and age stratification.

Our study of 2013 data in NY Medicaid found more than 11,000 primary mental health discharges in children 0-18 and another 3,000 or so in 19 and 20 year olds (overall N=13,692). We present combination measures of MH (mental health) and PCP (primary care clinician) follow up, followed by these broken out by type of visit. We found 66.8% of initial mental health visits were delayed, 52.0% of initial primary care visits were delayed, with 7.5% having delay in primary care visit only, 22.4% having delay in mental health visit only and 44.5% experiencing delay in both. We have also demonstrated important variations by race/ethnicity, age, percent poverty in the county, and urbanicity. The low rates of timely follow up and the high rates of MH readmission strongly suggest the clinical importance of this measure.

We use administrative claims data as used by Medicaid and CMS for billing. We use broad classes of mental health disorders codes, so that coding errors among or shifts between similar diagnoses will not be a problem. We include all the codes that indicate suicidality or self-harm, again reducing opportunity for errors. And these data sources are in use by exiting NQF measures (0576) and currently by researchers in the peer review literature. While claims data are imperfect, Dr. Kleinman and Dr. Shemesh (a member of the CAPQuaM team) reviewed the validity of administrative claims (and its history) in their commentary in the 2014 Yearbook of Pediatrics.

1.3. What are the dates of the data used in testing? 2013 data in NY Medicaid

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.26*)**	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: State Mediciad (population), County	☒ other: State Mediciad (population), County

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

NY Medicaid covers 183,000 children who had a mental health diagnosis in 2013, 132,000 of them were in Medicaid managed care, which was our primary data source. These children had 13,692 admissions.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

13,692 children 0-21 with mental health diagnoses.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Race/ethnicity. Rurality/urbanicity and level of poverty of the county of the caregiver's residence.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The basis for the scientific soundness, validity, and reproducibility of this measure lies CAPQuaM's peer reviewed 360 degree process which include a structured literature review, focus groups with parents/caregivers, an expert panel rating specifications and constructs using a RAND type modified Delphi process (2 round) and in the participation of a national steering committee and a large and diverse senior advisory board. (Table 1)

Table 1: Overview and summary of CAPQuaM's peer reviewed 360 degree method

Stage	Phase	Innovation	Product(s)
1. Clinical Criteria Development	a. Input Development	1. Focus groups of caregivers of children with asthma who have used the ED 2. Interviews with front line clinicians: primary care, asthma docs, and ED docs	1. Literature review 2. Summary of consumer perspectives, values and understanding relevant to clinical issue of interest 3. Summary of findings form clinician interviews
	b. RAND/UCLA 2 Round Modified Delphi Process	1. Inclusion of consumer perspectives as a key input; 2. Use of this method to identify appropriateness criteria in national performance measure development;	1. Explicit criteria that rank a comprehensive and mutually exclusive set of clinically detailed scenarios;
2. Boundary Guideline Development	Criteria Enhancement	1. Iterative process to enhance reliability and internal consistency of the explicit criteria set with a goal of outlining three boundary spaces	1. Internally consistent set of explicit criteria that are stable in their representation of the expert panel perspective. "Enhanced criteria"
	Guideline Articulation	1. Stakeholder (including experts, users, clinicians, consumers and others) informed review of the enhanced criteria. 2. Definition of zones of potential overuse, potential underuse, and professional interaction and decision-making based upon the explicit criteria 3. Stakeholder valuations of potential deviations from guideline 4. Boundary Guideline	1. Boundary Guideline 2. Prioritization list

3. Creation of Measure	Specification	1. Translation of guideline into specification of necessary data 2. Iterative process to define optimally efficient sources of data to allow for measurement and stratification	1. Initial specification of measure
	Review	1. Constructive peer review of specifications by stakeholders in Steering Committee and SAB	1. Final specifications of measure including variables for stratification as needed
	Fielding and testing of measure	1. Measure testing	1. Functional experience and practical understanding of measure, its scoring, variability, and interpretation

It further benefits from the validity of administrative claims data, particularly when used as we have used them, without asking them to make fine diagnostic distinction. Our findings and our standard errors suggest high signal to noise ratio of the measure as well as sensitivity to small differences in the specifications (ie 7 days, 14 days, 21 days, 30 days, 60 days etc).

Such data are used for billing by private and public insurers, quality improvement initiatives and analogous measures of which we are aware, including 0576. AHRQ just funded us to conduct an R01 using this data source to assess policy changes in the child mental health population of NY State Medicaid.

Most databases contain consistent elements, are available in a timely manner, provide information about large numbers of individuals, and are relatively inexpensive to obtain and use. Validity of many databases has been established, and their strengths and weaknesses relative to data abstracted from medical records and obtained via survey have been documented. [40] Administrative data are supported, if not encouraged by federal agencies such as NIH, AHRQ, HCFA, and the VA. The Centers for Medicare & Medicaid Services made clear to the participating AHRQ-CMS CHIPRA Centers of Excellence funded to develop measures in the Pediatric Quality Measures Program that it places a premium on feasibility when assessing those measures that it will most highly recommend to states to complete. The sources of data for the existing measure and other similar measures are typically based upon administrative data providing consensual validation for the appropriate primary data source.
Constructs underlying these measures:

- Identifying children with a mental health diagnosis through the use of diagnostic and billing codes
- Identifying specific services children received in the specified times frames following their mental health admission: primary care visits and visits with a mental health care provider
- Incorporating widely used coding schema, including HCPCS, CPT, and CMS's revenue codes and place of service in ways consistent with previous usage
- Identifying the type of facility providing the service using CMS's place of service codes.

We were guided in our inclusion criteria for a mental health hospitalization by the results of a formal RAND/UCLA modified Delphi process conducted with a multidisciplinary panel of national experts, which included a pediatrician, pediatric hospitalist, family physician, child psychiatrist, adult psychiatrist, adolescent physician, family advocate, discharge planner, and a licensed psychologist. The definitions were specified to allow their use with data elements that are typically available in electronic form to a responsible entity, such as a health plan or state Medicaid program. Part of our validation process using New York State Medicaid data for iterative testing to refine our specifications. We conducted at least 8 distinct rounds of testing using these data. We assessed the number of discharges identified using for example CPT codes with and without revenue codes; we looked at findings using various thresholds that were consistent with expert panel recommendations; and we modified our specifications to not include day of discharge when we found that a disproportionately large

proportion of follow up visits that were captured using our initial specifications occurred on the day of discharge and therefore failed to provide longitudinal assessment as is necessary for good follow up care.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Standard errors were consistently small, indicating a high level of precision and the capacity to distinguish signal from noise both within and across populations.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

High reliability with excellent precision.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☐ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The reliability section also contains information related to validity and is incorporated here by reference. Administrative data using ICD9 and CPT4 codes have been shown to be reliable and effective means to identify clinical encounters. We have previously shown (in our asthma measure development work) that the validity (particularly the sensitivity without cost to the specificity) of administrative data can be enhanced by using Revenue codes. Like CPT codes, place of service codes are sufficiently valid to be used by CMS for payment decisions. We have been reassured by our NY State Medicaid partners regarding the validity of provider type coding within state Medicaid data sets and from their national experience in managed care data sets as well.

We tested the measure in 2013 Medicaid claims data from NY State. Analyses were conducted by NY State Medicaid employees using the actual data set and there were no issues with feasibility and programming in SAS was straight forward. Our results were logical and as expected and were sensitive to differences across the various substrata analyses that we considered. We looked at various cutoff periods and variation was as one would expect by the design, ie there were more follow up visits at 14 days than 7, at 21 than 14, etc. During our initial testing we found that more than half of timely MH follow up visits occurred on the day of discharge, leading us to specify the follow up period as beginning on the day after discharge. The pattern did not fit anything we could justify based in the literature and we were concerned this might have represented gaming of the system. Visits on day 0 also failed to provide longitudinal care as is critical for follow up. The psychiatric guidelines previously referenced support the importance of ongoing longitudinal care.

Our results were all plausible and no other results demanded challenging explanations.

Measure results were:
Delayed coordination of care: 74.46% (10,181/13,692)
Delayed MH only: 22.37% (3,063/13,692)
Delayed PC only: 7.52% (1,029/13,692)
Delayed Both: 44.47% (6,089/13,692)

Use of expert panels has been demonstrated to be useful in measure development and health care evaluation, including for children. [41] Practitioners have been identified as a resource for researchers in developing and revising measures, since they are on the frontlines working with the populations who often become research participants. Involving practitioners can assist researchers in the creation of measures that are appropriate and easily administered. [42] Our expert panel supported measures that assessed the presence of prompt follow up with a mental health professional following hospitalization for mental health and also with a primary care clinician. Our expert panel further defined the age ranges and range of diagnoses to be considered as mental health discharges, and who could be considered a primary care clinician and mental health clinician for the purposes of follow up. We worked closely with our partners in the New York State Medicaid program to map the intended constructs to administrative data fields that were both available in New York and that would typically be available. Finally, our expert panel defined what constitutes delay (> 30 days for MH and for PC follow up visits).

Key reference materials for our work included our partner NCQA and HEDIS's specifications for their measure on follow up after mental health discharge, and articles in the literature including one co-authored by Senior Advisory Board Member Harold Pincus [44], AHRQ's specifications for its clinical classification software, the standard reference manuals for ICD-9CM and CPT-4 published by Ingenix, and CMS' own Revenue Codes and Place of Service codes. [45] We were also informed by a recently published annual report on mental health admissions for children, [46] and have conducted analysis of the KIDS database to enhance our understanding of this area.

Our final definitions operationalize the recommendations of our expert panel. As needed we guided decisions with reference to the sources noted to the previous paragraph and also our own analyses of HCUP and NY State Medicaid data. Specific pretesting included iterative analyses in NY State Medicaid data, which demonstrated that our parameters (definitions of admissions and follow-up) were selective but not overly restrictive, especially in regards to the current HEDIS measure. This helped us achieve our goals of more accurately reporting follow-up rates among pediatric populations.

ICD10 Code Development

The development team's goal was to develop an ICD10 code set that was fully consistent with the intent of the original measure.

Our process began by performing general equivalency mapping using the forward mapping from www.icd9data.com. We then did a de novo review of the CMS ICD 10 CM set to seek to identify codes that might be appropriate for this measure. We reviewed potential codes identified by both sources and developed a new list of codes appropriate for inclusion criteria. To assist with transcription we used a code list from an excel spread sheet available from CMS. We developed two lists, one to include all codes that determine eligibility only when they are primary diagnosis and a second set of codes that allow for eligibility regardless of where they are in the diagnosis list. These latter codes represent codes for self harm, suicidal ideation, homicidal ideation and the like. Key team members for this work were Suzanne Lo, MPH who staffed and coordinated this work, Eyal Shemesh, MD and Lawrence Kleinman, MD, MPH. Dr. Shemesh is a pediatrician, psychiatrist, child psychiatrist and Director of Behavioral Developmental Pediatrics at Icahn School of Medicine and was a lead developer for this measure. Dr. Kleinman is both CAPQuaM PI and was a lead developer for this measure. Dr. Kleinman and Dr. Shemesh reviewed the lists independently and achieved

consensus in a conference call review and discussion. The guidance for the intended constructs for both ICD9 and ICD10 coding were the findings from a RAND style modified Delphi panel that incorporated 9 national experts over the course of the measure development process.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

We concluded that our specifications included a substantially different population than the current measure 0576, that we were able to identify with our specification a distinct population with mental health discharges and that we were able to identify the timing of follow up and the specialty of the follow up provider. We further found that while there is substantial overlap between the two findings of primary care and mental health follow up, they do not duplicate information: Kappa = 0.39. McNemar's Chi is >1000, showing the marginal proportions are different as well.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☒ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

There are no biological hypotheses supporting greater risk of delayed follow up across strata. Entities are responsible for managing the populations that they manage and risk adjusting for these factors would obscure

real differences in performance. The use of stratifications allows a like to like comparison to be made when the accountability entity chooses to see that in addition to the top line results.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

In working with AHRQ, CAPQuaM has standardized its strata across various CAPQuaM PQMP measures to include race/ethnicity, age strata (individualized for each measure), rurality/urbanicity, poverty, health plan type, as described above.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the*

information provided related to performance gap in 1b)

We used chi square to assess statistical differences across different strata.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

We found statistical differences across strata, such as race, rurality, age.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The measures are sensitive enough to detect meaningful differences as observed within a population. Since the sum of squares across populations is expected to be greater in distinct populations, we expect the measure to perform very well when comparing across populations as well. Since the effective sample size of within population comparisons (such as we have conducted) is diminished by a variable intraclass correlation coefficient, we would expect greater power for equal sample size to detect differences between entities than we had in our testing of various subpopulations within a single state. This supports the same conclusion. The signal to noise ratio is very strong for these measures.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Our specification and definitions do not include hospital discharges for which the reporting entity is not expected to have the relevant data.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry), Other

If other: Race and zip code may be from medical record or encounter data

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

In testing, these measures were able to be completed using administrative data from the NY State Medicaid Program.

Limitations of the measure

These measures suffer from the usual limitations of administrative data analysis. Our careful and iterative processes have mitigated these limitations to the extent possible.

We do not consider in this measure specific processes that may enhance follow up, limiting the opportunity for this measure to inform regarding mechanisms for achieving improvement.

The current measure does not included a patient reported component that may be informative regarding follow up.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

None at present.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Not in use	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

The measure is not currently in use because it is newly developed and awaiting NQF endorsement.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

As a part of our work with PQMP, we are working on specific plans for dissemination and use. Our plan for implementation includes submitting our application for measurement endorsement from the National Quality Forum. We are having conversations with partners regarding the application and use of this measure. No time frames have been established, as the measure requires endorsement before it is implemented. Meeting the expected timeframes of NQF, the plan will include an accountability application within 3 years of initial endorsement and will be publicly reported within six years of initial endorsement.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

The measure is clearly defined and readily understandable to consumers, patients, clinicians, providers, purchasers, health plans, policy makers and others.

The timeliness of follow up visits following hospital discharge of children with a primary mental health diagnosis, specifically examining delayed coordination of care following mental health discharge, is an important measure that helps ensure high-quality, efficient healthcare for individuals and populations. Timely follow-up is a key component of the optimal management of any number of medical conditions, but is especially critical for children with mental health diagnoses. Timely follow-up with both primary care providers and mental health practitioners after a hospital discharge is imperative to deliver the best mental and physical health outcomes, reduce hospitalization and associated health costs.

Readmission rates for a mental health diagnosis following mental health discharge is common: in our NY State data 19.4% were readmitted for a mental health diagnosis within 90 days and 28.2% within 180. These are children with substantial health care needs that tax the capacity of the ambulatory setting. Successful ambulatory care on a population level requires both primary are and mental health professionals providing a substantial amount of well-coordinated care.

A variety of stakeholders would benefit from measuring delayed coordination of care following mental health discharge. Purchasers, health plans, consumers would all benefit. Policy makers could use this information, of example, to look at the impact of work force interventions. The measure has meaning at the level of populations, systems, health plans and within geographic areas at the level of the hospital. The granularity of the various stratifications allow targeting in terms of types of clinicians and in terms of populations served.

Furthermore, it would allow stakeholders, including researchers, to explore the association of timeliness of follow up to mental and physical health providers and the child's future health, specifically looking at rates of readmission or future mental and physical diagnosis or health status.

More importantly, this measure would allow the categorization of types of delay. In other words, providers, health systems, states and researchers can categorize patient's delay in mental health or primary care follow up visit by greater than 30 days, which helps to understand patients, their care trajectory, barriers and opportunities to timely follow up care. The information derived from this measure can help stakeholders to design innovative targeted healthcare models. With a better understanding of follow-up patterns after hospitalization for a mental health condition, health care organizations and policy makers can develop better informed services for children with mental health conditions.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

There has not been any evidence of unintended negative consequences to individual or populations. There are no anticipated unintended consequences if measuring at the level of comparing states, geographic regions, payment models, or health plans. In general, may need to stratify by age to avoid confounding. When comparing hospitals it may be important to incorporate strata that define context, such as rurality, poverty, and types of insurance of the discharged patients.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

0576 : Follow-Up After Hospitalization for Mental Illness (FUH)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
This measure is a purposeful enhancement of an existing measure. Unlike 0576, this measure is designed, optimized and specified for children. It incorporates primary care coordination as well as continuity in the mental health system. Visits on the day of discharge do not satisfy the criteria for this measure. This is because testing showed that a disproportionate number of MH FU visits were on Day 0, wondering if organizations had found a way to game the system, perhaps with an outpatient stop on the way home. These are administrative data requiring one time programming so the administrative burden should be trivial while the potential for enhanced measurement is significant. We view 0576 as a critical building block. CAPQuaM makes explicit the various components of follow-up, specifically, continuity within the mental health specialty and the coordination that occurs across specialties (i.e. PCP and mental health provider). Furthermore, expanding on NQF 0576, the measure can be stratified by the type of delay, MH only, PC only, or both. This measure is specified as a delay measure, consistent with guidance from our 360 method which includes a structured literature review, focus groups with caregivers/parents, discussions with professionals, a national expert panel using a RAND modified Delphi method to define the specifications included in the measure, a national steering committee and a national senior advisory board all contributing to the development of this measure.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
This measure expands the population covered by 0576. It is part of a measure set developed by CAPQuaM to measure and report rates of follow up after a mental health hospitalization in a pediatric population. CAPQuaM was asked by CMS and AHRQ to enhance an existing measure in the Health Effectiveness Data and Information Set (HEDIS) that was developed by the National Committee for Quality Assurance (NCQA) by optimizing it for child health concerns.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** FINAL_CAPQuaM_MHFU_Appendix.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** University Hospitals Cleveland Medical Center
Co.2 **Point of Contact:** Lawrence, Kleinman, drlarrykleinman@gmail.com, 617-699-3357-
Co.3 **Measure Developer if different from Measure Steward:** Collaboration for Pediatric Quality Measures (CAPQuaM)

Co.4 **Point of Contact:** Lawrence, Kleinman, drlarrykleinman@gmail.com, 617-699-3357-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Role: Expert Panelists

Moira Szilagy University California Los Angeles (previous University of Rochester)
Brian Pate University of Kansas School of Medicine-Wichita
Chris Lupold Lancaster General Hospital
Bradley Stein RAND Corporation Pittsburgh
Charles Saldanha Contra Costa Regional Medical Center
Jonathan Pletcher Children's Hospital of Pittsburgh of UPMC
Teresa King National Federation of Families for Children´s Mental Health and National Technical Assistance Center for Children's Mental Health, Georgetown University
Oona Caplan Mount Sinai Medical Center
Julie Carbray University of Illinois at Chicago

ROLE: Steering Committee (and Investigators)

Wilson Pace, MD American Academy of Family Physicians – DARTNET Institute - University of Colorado
Lynn Olson, PhD American Academy of Pediatrics
Christina Bethell, PhD, MBA, MPH Child and Adolescent Health Measurement Initiative, Johns Hopkins University (Previous OHSU)
Elizabeth Howell, MD Icahn School of Medicine at Mount Sinai
Harold Kaplan, MD Icahn School of Medicine at Mount Sinai
Lawrence Kleinman, MD, MPH Icahn School of Medicine at Mount Sinai
Rebecca Anderson Mount Sinai Medical Center
Eyal Shemesh, MD Icahn School of Medicine at Mount Sinai
Mary Barton, MD National Committee on Quality Assurance
Charles Homer, MD, MPH US Department of HHS (previous National Institute for Child Health Quality)
Marla Clayman, PhD American Institutes for Research (previous Northwestern University)
Foster Gesten, MD New York State Dept. of Health, Office of Health Insurance Programs
Jerod M. Loeb, PhD The Joint Commission
Robert Rehm National Committee for Quality Assurance
Steve Kairys, MD American Academy of Pediatrics/QuIIN
Erin DuPree, MD The Joint Commission (previous: Icahn School of Medicine at Mount Sinai)
Beverley Johnson, BSN* Institute for Patient- and Family-Centered Care
Doris Peter Consumers Union

ROLE: Senior Advisory Board Member and Investigator

Shoshanna Sofaer, DrPH American Institutes for Research (previous CUNY Baruch)
Harold Pincus, MD Columbia University, NYSPI
Lynne Richardson, MD Icahn School of Medicine at Mount Sinai
Ian Holzman, MD Icahn School of Medicine at Mount Sinai
Marilyn Kacica, MD New York State Dept. of Health, Division of Family Health

ROLE: Senior Advisory Board Member

Marc Lashley, MD Allied Pediatrics
Gary Mirkin, MD Allied Pediatrics
John Santa, MD* (previous Consumers Union)
John Clarke, MD ECRI, PA Patient Safety Authority
Scott Breidbart, MD Empire Blue Cross Blue Shield/ Anthem
Robert St. Peter, MD, MPH Kansas Health Institute
Ruth Stein, MD Montefiore Children's Hospital
Arthur Aufses, MD Icahn School of Medicine at Mount Sinai
Eric Rose, MD Icahn School of Medicine at Mount Sinai
Wendy Brennan, MS National Alliance on Mental Illness of New York City (NAMI - NYC Metro)
Laurel Pickering, MPH Northeast Business Group on Health

Martin Hatlie, JD Partnership 4 Pt. Safety, Consumers Advancing Patient Safety
Paul Wise, MD Stanford University
Lisa Simpson, BCh Academy Health

ROLE: Investigators/Key Staff
Rusty McLouth, MS AAFP
Elise Barrow, MPH Icahn School of Medicine at Mount Sinai
Natalia Egorova, PhD Icahn School of Medicine at Mount Sinai
Elizabeth Howell, MD Icahn School of Medicine at Mount Sinai
Harold Kaplan, MD Icahn School of Medicine at Mount Sinai
Barbara Rabin, MHA Icahn School of Medicine at Mount Sinai
Carolyn Rosen, MD Icahn School of Medicine at Mount Sinai
Melissa Saperstein, MSW Icahn School of Medicine at Mount Sinai
Eyal Shemesh, MD Icahn School of Medicine at Mount Sinai
Virginia Walther, MSW Icahn School of Medicine at Mount Sinai
Marianne McPherson, PhD, MS NICHQ
Joseph Anarella, MPH NYS DOH, OHIP
Lee Sanders, MD, MPH Stanford University
Kasey Coyne Northwestern
Victoria Wagner New York State Health Department Office Quality Patient Safety
Wei Jing New York State Health Department Office Quality Patient Safety
Judy Stribling Icahn School of Medicine at Mount Sinai
Louise Falzon Columbia University
Keri Thiessen AAP – QuIIN
Kasey McCracken OHSU
Sandeep Sharma Icahn School of Medicine at Mount Sinai
Ann Nevar UH Rainbow Babies and Children's Hospital/ CAPQuaM
Suzanne Lo UH Rainbow Babies and Children's Hospital/ CAPQuaM
Amy Balbierz Icahn School of Medicine at Mount Sinai
Samantha Raymond Icahn School of Medicine at Mount Sinai
Allisyn Vachon Icahn School of Medicine at Mount Sinai
Shannon Weber Icahn School of Medicine at Mount Sinai

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2817
Measure Title: Accurate ADHD Diagnosis
Measure Steward: American Academy of Pediatrics
Brief Description of Measure: Percentage of patients aged 4 through 18 years whose diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) was based on a clinical exam with a physician or other healthcare professional, as appropriate which includes: confirmation of functional impairment in two or more settings AND assessment of core symptoms of ADHD including inattention, hyperactivity, and impulsivity, either through use of a validated diagnostic tool based on DMS-IV-TR criteria for ADHD or through direct assessment of the patient.
Developer Rationale: According to statistics provided by the Centers for Disease Control and Prevention, 5 million children (9%) aged 4-17 years have ADHD, the percentage of children with parent-reported ADHD increased by 22% between 2003 and 2007, and rates of ADHD diagnosis increased an average of 3% per year from 1997 to 2006 and an average of 5.5% per year from 2003-2007. (1) In November of 2011, the American Academy of Pediatrics (AAP) published a new evidence based guideline for ADHD diagnosis, follow-up, and treatment, which was based on extensive review of the existing evidence. One recommendation with a high level of evidence indicated that when diagnosing ADHD in children 4-18 years of age, primary care clinicians should determine that "Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition" (DSM-IV) criteria have been met (including documentation of impairment in more than one major setting), with information obtained from reports of parents or guardians, teachers, and other school and mental health clinicians involved in the child´s care. Furthermore, the primary care clinician should rule out any alternative cause and include assessment for other conditions that might coexist or be comorbid or consequent to ADHD, including emotional or behavioral, developmental, and physical conditions (2). Validated tools based on DSM-IV criteria have demonstrated effectiveness for diagnosing ADHD and for distinguishing ADHD from the diagnosis of other conditions which may have the same symptomology and/or impairment. When less rigorous methods are applied to the diagnosis of ADHD, the positive existence of the condition ADHD may be missed, leading to potential social and academic struggle. A diagnosis of ADHD may also be made erroneously when another condition is present that may need immediate attention to prevent increased severity. Either false negative or false positive diagnostic errors can lead to poor quality of care and potential harm.

1. Centers for Disease Control and Prevention. Summary health statistics for U.S. children: National health interview survey, 2009. Vital and Health Statistics Series. 2010;10(247).

2. Subcommittee on Attention-Deficit Disorder/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Pediatrics. 2011; 128(5):1-16.

Numerator Statement: Patients whose diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) was based on a clinical exam with a physician or other healthcare professional, as appropriate which includes: confirmation of functional impairment in two or more settings (1) AND assessment of core symptoms of ADHD including inattention, hyperactivity, and impulsivity, either through use of a validated diagnostic tool (2) based on DMS-IV-TR criteria for ADHD or through direct assessment of the patient.

(1) Settings: Includes home, school, and community
(2) Validated diagnostic tool used may include any of the following examples, all of which are based on the DSM-IV criteria for ADHD:
Conners Rating Scales
Barkley ADHD Rating Scale
Vanderbilt Parent and Teacher Assessment Scales
ADHD Rating Scale-IV (DuPaul)
Swanson, Nolan, and Pelham-IV (SNAP IV) Questionnaire

Other ADHD diagnostic tools may be determined valid based on DSM-IV criteria and therefore would be acceptable for this measure and will be added to the list at periodic updates.
Denominator Statement: All patients aged 4 through 18 years with a diagnosis of ADHD.
Denominator Exclusions: n/a

Measure Type: Process
Data Source: Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: n/a

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following information for this process measure (Level of Analysis = Clinician: Group/Practice and Facility):

- The developer linked accurate diagnosis of ADHD to increases in appropriate treatment and decreases in inappropriate treatment, which leads to patients resulting in improved quality of life, grades, and functionality. Evidence for this process measure should demonstrate that when an ADHD diagnosis is accurate, appropriate treatment is more likely to occur, which will lead to the desired outcomes of improved quality of life, better grades, and increased functionality.
- The measure is based on a recommendation from the 2011 American Academy of Pediatrics' *Clinical Practice Guideline for the Diagnosis, Evaluation, and Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescent* that is based on grade B evidence. This grade evidence indicates that it includes RCTs or diagnostic studies with minor limitations and overwhelmingly consistent evidence from observational studies. The recommendation is graded as "strong", meaning that it is based on high- to moderate-quality scientific evidence and a preponderance of benefit over harm.
- The developer reported that the body of evidence underlying the clinical practice guideline included 14 studies that ranged from 1996 to 2009. The developer notes that the DSM-IV criteria can be applied to pre-school children, school-age children, and adolescents, although they note a few caveats. They also state that the "*DSM-IV system does not specifically provide for development-level differences and might lead to some over diagnosis*".
- Per the **NQF Algorithm for Evidence**, the eligible ratings are HIGH, MODERATE, or LOW because the developer identifies a systematic review that is graded and that assesses the quantity, quality, and consistency of the evidence (box 3-->4).

Questions for the Committee

- *The numerator construction is "including inattention, hyperactivity, and impulsivity." Is there evidence that other symptoms should be included? Does the Committee wish to discuss with the developer whether ALL three must be present?*
- *Do you know of validated instruments that allow ADHD diagnosis according to the DSM-IV-TR? The*

specifications permit use of such an instrument OR direct assessment of the patient. Does the Committee wish to discuss with the developer the strength of evidence for one or the other options?

- *Is the relationship of this measure to patient outcomes reasonable, and how strong is the evidence for the relationship?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer provided performance results for this measure using data abstracted from 118 charts across four outpatient clinician office networks in the Chicago area.
  - Performance rates varied from 63.41% to 92.86% across the four sites.
  - The developer notes that racial/ethnic disparities were found among the patient population included in the four outpatient clinician office networks.
    - Among Asian patients (N=3), the measure performance was 66.67%
    - Among Black patients (N=31), the measure performance was 54.84%
    - Among Hispanic patients (N=27), the measure performance was 55.56%
    - Among White patients (N=32), the measure performance was 81.25%; and
    - Among patients whose race/ethnicity is Unknown (N=19), the measure performance was 73.68%

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- Process measure leads to diagnosis although ruling out other diagnoses is vague. 3 types of ADHD: primarily inattentive, impulsive-hyperactive and combined which are not addressed here. As well there is a evidence that ADHD presents differently in girls than in boys especially regarding the presenting symptoms. Checklists are limited, have not been developed using DSM IV or IV TR or V criteria as many predate these criteria and may not include those used by psychology or developmental pediatrics. DSM version may be an issue as DSM 5 is now in use. Also the measure states DSM IV TR and DSM IV. Diagnosis typically must be based on evaluating all information (checklists, observation, and history in multiple settings) as inconsistencies may point to another diagnosis. Checklists should be from multiple settings if possible (e.g. teacher and parent). I know of no checklists based on the DSM IV TR criteria specifically. Appropriate diagnosis even if there is a comorbidity does lead to appropriate treatment of the condition and improved function. This strategy does not rule out intellectual disabilities as a diagnosis which is a diagnosis that might mimic ADHD, but would be treated very differently.
- Measure applies directly to process, linkage to outcome is less clear.
- There is good evidence that treatment improves outcomes, and reasonable inference that accurate diagnosis is needed to appropriately target treatment. The evidence that the accuracy of diagnosis is significantly improved (over global assessment by a provider) by the use of either the standardized instruments or documentation of the clinician's observations is limited.
- Strong evidence High
- Links evidence supporting accurate dx of ADHD to increase in appropriate tx and decrease in inappropriate tx leading to improved QOL and functioning. Based on the recommendation from the AAP in 2011 (14 studies).

1b. Performance Gap.

- Disparities in diagnosis and access to diagnosis and treatment exist in racial and ethnic minorities. There also

may be confounders that make ADHD symptoms more common in some populations including living in poverty. African American males living in poverty are more likely to have the symptoms, but not clear if the confounding factors may be contributing to the symptoms. This measure should measure practitioner performance and not access to care or other potential disparities, but could practitioners be treating kids of racial and ethnic minorities differently (e.g. not obtaining the appropriate information before making the diagnosis) or is it more difficult to get the information (checklists etc.) in some areas which affects children of racial and ethnic minorities more? There is a gap in performance that would be amenable to improvement even in the highest performing group. This is a common condition and one that is diagnosed by various practitioners, but the diagnosis should be made in the same manner regardless of the practitioner.

- Although it would be nice to have more evidence, there is plenty of evidence in other areas to suggest that the variation detected is likely to occur throughout the country.
- using the measure, the developers show that many diagnoses of ADHD are not sufficiently supported by documentation, and that there are racial disparities in this measure. They do not show that large numbers of children are misdiagnosed or go undiagnosed.
- Uncertain how subpopulations represent disparities. What is the data on subpopulations and rates of diagnosis? Unclear how many assessed at the different sites to determine the importance of the gaps identified.
- In some settings there is a significant performance gap and there is a performance gap that is greater in minorities.
- Performance data provided for 118 charts across 4 outpatient clinical offices. Data given by ethnic groups.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

The developer provides the following information:

- This measure is specified at the group practice and facility levels of analysis for the (clinician office and inpatient psychiatric settings, respectively). It uses data from either electronic health records or paper medical records.
- This measure captures children ages 4-18 years with a diagnosis of ADHD. There are no exclusions specified for the measure.
- ICD-9 and ICD-10 codes to identify patients diagnosed with ADHD (the denominator) are provided. The developer notes in the supplement excel file that the ICD-10 codes were identified through use of www.ICD10data.com. Information needed for the numerator must be obtained from the medical record (no standardized codes required).
- Assessment of core symptoms can be done through use of a validated instrument or through direct assessment. The developer lists examples of validated instruments that can be used but note that others may be acceptable.
- The calculation algorithm is provided. In this algorithm, the developer suggests simple random sampling of charts, but does not provide any guidance on the number of charts needed for reliable measurement.
- The developer also encourages stratification of measure results according to gender, age group, race, language, and insurance type. However, this stratification is optional and is not meant to serve as a method of risk-adjustment.

Questions for the Committee

- *Are the instruments cited as examples in the submission appropriate for assessing the core symptoms of ADHD?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high

proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer provides the following information:

- Empirical reliability testing for this measure was conducted at the data element level.
- The developer tested the measure using data from 118 medical records obtained from four outpatient clinician office networks in the Chicago area. The timeframe for testing was December 2011-June 2012.
    - Testing was conducted using both paper and electronic records housed in outpatient clinics; random sampling of medical records was done for one clinic.
    - The developer provides demographic information for patients included in the testing. The clinics are not specifically described, although the consortium to which they belong is.
    - The developer does not provide information on testing for inpatient psychiatric facilities, although the developer indicated that the Level of Analysis included facilities.
    - Inter-rater reliability was assessed by computing percentage agreement and the Kappa value. Kappa is a statistic that represents the proportion of agreement between two abstractors that is not explained by chance alone. Values for kappa range between -1.0 and 1.0. A value of 1.0 reflects perfect agreement; a value of 0 reflects agreement that is no better than what would be expected by chance alone; a value less than zero reflects agreement that is worse than what would be expected by chance.
    - The developer reports Kappas ranging from 0.27 to 0.60 for the numerator. Specifically,
        - Evidence of clinical exam by physician in chart (yes/no) = Kappa 0.27
        - Evidence in the chart of assessment of core symptoms of ADHD, including inattention, hyperactivity and impulsivity through a validated diagnostic tool AND through direct assessment of the patient (yes/no) = Kappa 0.60
        - Evidence in the chart of assessment of impairment in two settings (yes/no) = Kappa 0.36
        - Overall ADHD measure (clinical exam by MD, evidence of impairment in two settings; and either assessment through validated tool or direct assessment) = Kappa 0.27
        - The literature generally considers a Kappa of 0.27 represents "fair" agreement, and one of 0.60 indicates "moderate" agreement.
    - No information is provided on reliability testing of the denominator, except the developer indicates the abstractors "received training on how to identify and select the charts for inclusion in testing." The developer further indicates the denominator ADHD diagnosis "can be identified by looking for an ADHD diagnostic code in the patient medical record."
- **Per the NQF Algorithm on Reliability,** testing at the data element is eligible for a MODERATE or LOW rating (box 8-->9).

Questions for the Committee

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is the reliability testing of only the numerator appropriate? Does the Committee wish to discuss with the developer the lack of empirical assessment for the denominator? Given the measure construct, what would be appropriate?*
- *The developer indicates the Level of Analysis also is facility, but no facility-level (e.g., in-patient psychiatric facility) testing data are presented for that care setting. Does the Committee wish to discuss with the developer the availability of such data, given NQF's requirement to test for the applicable Level of Analysis?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the

- The AAP ADHD guideline notes that to make a diagnosis of ADHD, the primary care clinician should determine that DSM-IV criteria have been met, including documentation of impairment in more than one major setting.

The measure specifications correspond and require two or more settings.

- The criteria in the guideline specify academic or behavioral problems and symptoms of inattention, hyperactivity, and impulsivity; the measure specifications correspond.
- The age group included in the measure corresponds to the guideline.

Question for the Committee

- *Are the specifications consistent with the evidence?*
- *Since the DSM-IV criteria for diagnosing ADHD are cited, do they include assessment of inattention, hyperactivity, and impulsivity?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- The developer did not conduct empirical validity testing for this measure.
- NQF guidance indicates that the assessment of face validity of the measure score as an indication of quality is an acceptable method for measure validation if systematically assessed by recognized experts.
- The developer states that the 25-member Expert Panel that helped develop the measure agreed that the measure can be used to distinguish good and poor quality care. However, they did not describe the process of obtaining this agreement and did not provide the data associated with the assessment.
- The developer also noted that face validity was assessed via a 21-day public commenting period and listed the organizations that provided comments. However, they do not describe if or how public commenters provided an assessment of the measure score as an indicator of quality and this is not included in the results provided.
- Per the **NQF Algorithm for Validity**, the eligible ratings for this measure are MODERATE and LOW since there is no empirical testing, but face validity was assessed at the computed measure score (box3-->4).

Questions for the Committee:

- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- There are no exclusions for this measure.

2b4. Risk adjustment:

- This measure is not risk-adjusted.

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer provides 95% confidence intervals for the performance rates for the four testing sites. The rates across sites varied substantially, and the confidence intervals for site 1 and site 4 did not overlap. This suggests that performance rates across sites may be statistically different.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Because this measure has only one set of specifications (i.e., for claims data), this section is not applicable.

2b7. Missing Data

- The developer states that denominator criteria were missing in approximately 5% of cases and numerator criteria were missing in 34% of cases.
- The developer notes that if data are missing for the numerator elements, the measure is calculated as not being met.
- It is unclear how the developer determined that data were missing for the denominator (i.e., having an ADHD diagnosis).

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- The reliability isn't very good and the limitations of using only 4 sites from the same geographic area makes this problematic. As well the use of this in an inpatient psychiatric facility where admission is complicated by the reason for admission (which is almost never ADHD) makes that a different population altogether. The diagnosis should be based on the same information, but may be much more complex to do. This seems like it doesn't belong. Clinical exam is not always reliable and shouldn't really be used alone as it may reflect other things that are happening at the time and not necessarily only how the child is all the time that they are not at the physician or psychologists. Facility should be discussed and consideration for dropping it as a potential location. The only measure with some reliability is the documentation of the presence of symptoms consistent with the diagnosis. The denominator should be assessed as well since using the diagnosis code for this doesn't pick up those kids that came in with symptoms and it was determined to be something else.
- The overall assessment and assessment of core symptoms had the lowest kappas, suggesting that these are the least well specified.
- The measure specifications are consistent with the evidence.

2a2. Reliability testing

- The instruments are appropriate even though they were not developed using the diagnostic criteria from the DSM (the core symptoms of ADHD have not changed dramatically over time). The diagnosis should be made using both standardized checklists and observation and history not "or" as there can be discrepancies that might lead to a different diagnosis and observation of the symptoms is not always appropriate or helpful especially in a situation where anxiety might be present as well. Not sure how measurement of multiple settings is going to be captured in the EHR. Likely would require reading the actual chart rather than a measure that could be pulled. Sampling of charts to get a statistically significant number is complex. Clearer guidance about how to do this could be included. Some elements are straightforward to assess (e.g. presence of a checklist, symptoms matching the checklist, etc.), however might be stronger if they required actually using a diagnostic checklist to identify symptoms present. The standard checklists all have scores, but they are not the same and not always "diagnostic". Complexity exists if one is looking at ADHD with one of the other subtypes (inattentive or impulsive-hyperactive). Unclear what stratifying the data by demographics would do although there are gender differences in symptoms and likely disparities based on race and ethnicity.
- Denominator is not well defined. Criteria is "an ADHD diagnosis within the record". Children may meet criteria for ADHD at one point in time and later find their diagnosis changed as more symptoms come to light.
- From the narrative, it's not clear whether all three of the core symptoms must be documented to be present. Most concerning issue is the low Kappas, using well trained research nurses for chart review.
- How far is the lookback for the tools used for diagnosis? Assessments may not be completed in a single visit or time frame.
- There are validated instruments. It is likely that this measure can be consistently implemented.
- Data based on 118 records. Kappas given for numerator data (dx by physician, core symptoms present, 2 settings, use of validated tools. No info given on testing denonminator.

2b1. Validity Specifications

- The specifications and evidence match except that DSM IV (and the current version 5) include three subtypes of ADHD: inattentive, impulsive-hyperactive, and combination. Only the combination has all three symptoms. Demonstration that the child meets the DSM criteria is not really measured here even though those are the

diagnostic criteria (i.e. it is not just are inattention, hyperactivity and impulsivity present, but are they manifest in the manners described by the criteria in the DSM?).

- Specifications are consistent with the AAP guideline, with caveat about denominator.
- The criteria for diagnosis are based on an accepted guideline; developer does not present the evidence underlying the diagnostic criteria (this is assumed). I do not know this literature well enough to assess validity of the diagnostic criteria.
- Important specifications present.

2b2. Validity Testing

- The presence of symptoms in multiple settings which is what this measure includes is the gold standard for diagnosis. There is no score and unclear how the "experts" agreed on this. This measure would be stronger if the "presence of symptoms" were measured by a score using the DSM criteria.
- Tested on a single small local sample. May not be generalizable across the country. Review by 25 member expert panel is likely a reasonable approach to assess face validity. Need more information about methodology to know if it was representative of anything.
- Not empirically tested. Expert panel liked the measure but process for evaluating the measure appears to have been a global assessment rather than assessment of elements or other more detailed evaluation.
- The validity testing for this measure was low. 25 member panel agreed, but did not provide specifics. Lack of detail on the 21 day comment period results
- No formal empirical testing was completed. Face validity performed through public comment but details missing.

2b3-2b7. Threats to Validity

- Statistically different performance indicates that it measures performance differences. Data missing from psychiatric inpatient would indicate that this measure really can't be used in that setting which is dramatically different from the clinic setting. Very unclear how the denominator was chosen if the criteria was an ADHD diagnosis and that was missing. The sample shouldn't have included any of those if the ICD codes were used to pull it.
- Unclear how missing data in denominator can be found. Without understanding that, it is possible that the measure will "wobble" as the sample changes.
- No exclusions for denominator seems appropriate and it is likely that most children who have the diagnosis will be captured using administrative data.
- Missing data would threaten validity and the structure of EHR would affect the ease of recording evidence that the process measure was met.
- Data given for missing data in their test sample for numerator and denominator. Details not given regarding how it is determined that data is missing.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The data source for this measure is paper medical records or electronic medical records.
- The developer reported that there are improvements that could facilitate collection and reporting of this measure, such as the documentation of the use of a specific validated tool and the setting in which the evaluation took place in queriable fields.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *How likely is it that the required data elements are available in structured fields in EHRs*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Elements are not part of the general pediatric exam typically and in order to be measured would either need modification of an EHR or text fields to support this. Paper charts would have the copies of the checklists etc. In

order to identify the presence of the symptoms the chart would have to have it in the history which means that would have to be read. Meeting the criteria for the diagnosis based on DSM criteria does not seem to be measured as the measure is just that the symptoms are present and the checklists are not specific to the diagnostic criteria in the DSM.
- Components of diagnosis will require chart audit- information on the elements of a diagnosis in not usually collected in a standardized way.
- Main concern is reliability of the numerator data when a standardized instrument is not used. Would have been useful to know kappa for those instances.
- Concerns about the timing over which a diagnosis may be made--how much of the chart and over what time is it needed to look for tools used for assessment.
- EHR may need to be revised to facilitate feasible extraction.
- It is unclear whether the information needed for the numerator is built into most medical records and/or recorded routinely. For example - information from multiple environments - is this a yes/no question that the MD documents?

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- The measure is currently used by the American Board of Pediatrics for both pediatrician performance improvement and as a requirement for professional recertification.
- The developer did not describe how it plans to use this measure for payment and public reporting.
- No information regarding improvement was provided.
- The developer stated that no unintended consequences were identified during testing of this measure.

***Questions for the Committee**:*

- *Can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Is the measure appropriate for accountability purposes (NQF's primary endorsement focus)?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- If tightened then this measure could indicate quality of evaluation for ADHD which is by its nature behavior and therefore requires history and information from multiple sources (as opposed to having a blood test or MRI finding that makes the diagnosis). It is an area that general pediatric practitioners can and do make the diagnosis and in general the quality of this is poor. As well the diagnosis leads to appropriate treatment and improvement in function which is often life-long. However, the measure while process may not measure what is actually happening. Also the use of the measure for inpatient psychiatric facilities without any data that it is valid or reliable in that setting is problematic.
- ADHD is best diagnosed through a careful clinical evaluation, supported by input from standardized tools like the Vanderbilt. Ease of measuring questionnaire use could privilege that avenue of diagnosis, reducing the nuance present in a careful clinical evaluation.
- This measure is intuitively appealing but the measure itself doesn't perform as well as one might hope.
- Usability is high for potential to prevent over-diagnosis and unnecessary treatment.
- If the information is collected routinely, could be an important measure. Unclear how easy it will be to obtain all necessary information.

Criterion 5: Related and Competing Measures

- 0108 : Follow-Up Care for Children Prescribed ADHD Medication (ADD) (NQF-endorsed) is related to this measure.
- Both measures focus on children and adolescents with ADHD diagnoses, however, this measure considers children and adolescents ages 4-18 and focuses on accurate diagnosis. # 0108 considers children ages 6-12 with

a new prescription for ADHD medication who had at least three follow-up care visits within a 10-month period, one of which is within 30 days of when the first ADHD medication was dispensed.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Accurate ADHD Diagnosis

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Accurate ADHD diagnosis

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Accurate diagnosis is a core element of safe, high quality care. When identifying and treating children with mental illness, pediatricians are least likely to agree on identifying and treating learning problems; however, approximately, 66% of physicians believe that pediatricians should treat or manage ADHD. [1] Furthermore, only about 50% of children with ADHD seen in practice settings obtain care that matches guidelines of the American Academy of Child and Adolescent Psychiatry. There is significant room for quality improvement in the area of accurate ADHD diagnosis. However, validated tools now exist that facilitate the complete and reliable evaluation of core symptoms and impairment for ADHD to ensure all elements of the DSM-IV criteria are assessed. When less vigorous methods are applied to the diagnosis of ADHD, the positive existence of the condition ADHD may be missed, leading to potential social and academic struggle OR a diagnosis of ADHD may be made erroneously when another condition is present that may need immediate attention to prevent increased severity. Either false negative or false positive diagnostic errors can lead to poor quality of care and potential harm. For example:

ADHD diagnosis = accurate → Increase appropriate treatment

- ➔ Decrease inappropriate treatment
- ➔ Improve quality of life
- ➔ Improve grades
- ➔ Improve functionality

This proposed measure represents an enhancement in the delivery of ADHD diagnostic care and will facilitate the accuracy of ADHD diagnosis and care. This proposed measure, Accurate ADHD Diagnosis, aims to increase appropriate diagnosis and treatment while decreasing inappropriate diagnosis and treatment ultimately improving quality of life, grades, and functionality in children with ADHD.

1. Stein RE, Horwitz SM, Storfer-Isser A, Heneghan A, Olson L, Hoagwood KE. Do pediatricians think they are responsible for identification and management of child mental health problems? Results of the AAP periodic survey. *Ambul Pediatr*. 2008;8(1):11-17.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management, Wolraich M, Brown L, Brown RT, DuPaul G, Earls M, Feldman HM, Ganiats TG, Kaplanek B, Meyer B, Perrin J, Pierce K, Reiff M, Stein MT, Visser S. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. *Pediatrics*. 2011;128(5):1007-1022.

http://pediatrics.aappublications.org/content/early/2011/10/14/peds.2011-2654

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

The guideline of interest is Action Statement 2 on page 6, "To make a diagnosis of ADHD, the primary care clinician should determine that *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-*

IV-TR) criteria have been met (including documentation of impairment in more than 1 major setting), and information should be obtained primarily from reports from parents or guardians, teachers, and other school and mental health clinicians involved in the child's care. The primary care clinician should also rule out any alternative cause."

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

The grade assigned is B/strong recommendation. The definition of this grade is, "RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies." This level of evidence is based on high- to moderate-quality scientific evidence and a preponderance of benefit over harm.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

The other grades are as follows:

A. Well-designed RCTs or diagnostic studies on relevant population (strong recommendation)
B. RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies (strong recommendation/recommendation)
C. Observational studies (case control and cohort design) (recommendation)
D. Expert opinion, case reports, reasoning from first principles (option)
X. Exceptional situations in which validating studies cannot be performed and there is a clear preponderance of benefit or harm (strong recommendation/recommendation)

A strong recommendation or recommendation statement is based on high- to moderate-quality scientific evidence and a preponderance of benefit over harm. Option-level action statements are based on lesser-quality or limited data and expert consensus or high-quality of evidence with a balance between benefits and harms. A health care provider might or might not wish to implement option recommendations in his or her practice.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

American Academy of Pediatrics, Steering Committee on Quality Improvement. Classifying recommendations for clinical practice guidelines. *Pediatrics*. 2004;114(3):874-877.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The evidence review focused on assessment, diagnosis, and treatment of children with ADHD.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The grade assigned is B. The definition of this grade is, "RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies."

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

The other grades are as follows:

A. Well-designed RCTs or diagnostic studies on relevant population

B. RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies
C. Observational studies (case control and cohort design)
D. Expert opinion, case reports, reasoning from first principles
X. Exceptional situations in which validating studies cannot be performed and there is a clear preponderance of benefit or harm.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1996-2009

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

In the body of evidence, the following study designs are included:

- 4 comparative studies
- 4 cross-sectional studies
- 4 reviews
- 1 case-control study
- 1 Diagnostic and Statistical Manual (Mental Disorders)
- 1 Diagnostic Criteria for ADHD
- 1 longitudinal study

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

There is good evidence that the diagnostic criteria for ADHD can be applied to pre-school children; however, ADHD subtypes detailed in the DSM-IV might not be valid for this population. Literature cited in the 2011 AAP ADHD Guideline indicates that the criteria can reliably and accurately diagnose children with ADHD. There is also strong evidence that the diagnostic criteria for ADHD can be applied to adolescents; however, adolescents are less likely to exhibit overt hyperactivity behavior and care needs to be maintained to establish younger manifestations of the condition that were missed and to consider substance abuse, depression, and anxiety as alternative or co-occurring diagnoses.

Benefits of this recommendation include more uniform categorization of the condition across professional disciplines and while the DSM-IV does not specifically provide for developmental-level differences and might lead to some misdiagnosis; the benefits far outweigh the harm.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

A review of the nosology and epidemiology of ADHD in children 2-5 years of age found in studies using DSM diagnostic criteria, the prevalence of ADHD in preschool children ranges from 2.0-7.9%. Preschoolers with ADHD presented with similar symptoms, features, and prognosis as older children with ADHD affirming the reliability and validity of ADHD diagnosis using DSM-IV-TR criteria. A study of 168 children with behavior problems at age 3 underwent a multimethod assessment of ADHD symptoms and were followed annually for 3 years. Using a diagnostic interview and rating scales at age 3, the authors accurate predicted ADHD diagnosis for 75% of the children. Parent and teacher ratings were collected on 902 and 977 children 3 to 5 years of age, respectively and reliability coefficients ranged from 0.80 to 0.95 indicating good test-retest reliability. Concurrent validity with the Conners Teacher Rating Scales: Revised-Short and Conners Parent Rating Scale: Revised-Short ranged from 0.54 to 0.96.

As approximately, 65% of children diagnosed with ADHD have symptoms persisting into adolescence, it is imperative when diagnosing older children with ADHD that functional impairment is assessed in two or more settings. Furthermore, as symptoms mature with adolescents, understanding and assessing the core symptoms of ADHD is necessary to diagnose an adolescent with ADHD.

The 2011 AAP ADHD Guideline reflects the increased evidence that appropriate diagnosis can be provided for preschool-aged children and adolescents and considers the evidence of previous guidelines focusing on school-aged children to remain unchanged. Studies reported in the 2000 AAP ADHD Practice Guideline indicate that specific questionnaires and rating scaled based on ADHD DSM-IV criteria have been shown to have an odds ration greater than 3.0 (equivalent to a sensitivity and specificity greater than 94%) in studies differentiating children with ADHD from normal, age-matched, community controls.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The DSM-IV system does not specifically provide for development-level differences and might lead to some over diagnosis; however, criteria can appropriately identify children with ADHD and care should be taken to ensure a diagnosis is based on DSM-IV criteria as well as confirmation of functional impairment in at least two other settings.

Adolescents with ADHD may exhibit social impairment, disorganization, inability to follow through on academic tasks, and difficulty sustaining attention for extended academic projects. Similarly, children with ADHD are at increased risk for developing substance abuse as they grow older. Therefore, while there is a slight risk for misdiagnosis of ADHD, the benefits of diagnosing a child or adolescent with ADHD and mitigating symptoms fair outweigh the harms.

It might also be difficult to have separate observers for pre-school aged children and teacher reports for adolescents as pre-school aged children may not spend much time outside the home and adolescents may spend short amounts of time with many teachers. Care must be taken to ensure that children have documentation of impairment in more than one setting.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1.02_ADHD_1_evidence_submission_form_9-28.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
According to statistics provided by the Centers for Disease Control and Prevention, 5 million children (9%) aged 4-17 years have ADHD, the percentage of children with parent-reported ADHD increased by 22% between 2003 and 2007, and rates of ADHD diagnosis increased an average of 3% per year from 1997 to 2006 and an average of 5.5% per year from 2003-2007. (1) In November of 2011, the American Academy of Pediatrics (AAP) published a new evidence based guideline for ADHD diagnosis, follow-up, and treatment, which was based on extensive review of the existing evidence. One recommendation with a high level of evidence indicated that when diagnosing ADHD in children 4-18 years of age, primary care clinicians should determine that "Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition" (DSM-IV) criteria have been met (including documentation of impairment in more than one major setting), with information obtained from reports of parents or guardians, teachers, and other school and mental health clinicians involved in the child´s care. Furthermore, the primary care clinician should rule out any alternative cause and include assessment for other conditions that might coexist or be comorbid or consequent to ADHD, including emotional or behavioral, developmental, and physical conditions (2). Validated tools based on DSM-IV criteria have demonstrated effectiveness for diagnosing ADHD and for distinguishing ADHD from the diagnosis of other conditions which may have the same symptomology and/or impairment. When less rigorous methods are applied to the diagnosis of ADHD, the positive existence of the condition ADHD may be missed, leading to potential social and academic struggle. A diagnosis of ADHD may also be made erroneously when another condition is present that may need immediate attention to prevent increased severity. Either false negative or false positive diagnostic errors can lead to poor quality of care and potential harm.

1. Centers for Disease Control and Prevention. Summary health statistics for U.S. children: National health interview survey, 2009. Vital and Health Statistics Series. 2010;10(247).

2. Subcommittee on Attention-Deficit Disorder/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Pediatrics. 2011; 128(5):1-16.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The performance of this measure as a chart abstraction measure was tested at four outpatient clinician office networks (N=118) in the Chicago area. The relatively small sample size is due to the alignment of this measure with the recommendations made in the 2011 AAP ADHD Guideline, resulting in a short post-Guideline testing period. The performance of the measure at the sites was as follows:

Site 1 - 63.41% (95% CI: 48.67%-78.16%)
Site 2 - 71.88% (95% CI: 56.30%-87.45%)
Site 3 - 90.91% (95% CI: 73.92%-100.00%)
Site 4 0 92.86% (95% CI: 83.32%-100.00%)

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
n/a

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Measure performance was calculated by race/ethnicity group for the data collected during testing. The results are as follows:

Among Asian patients (N=3), the measure performance was 66.67%
Among Black patients (N=31), the measure performance was 54.84%
Among Hispanic patients (N=27), the measure performance was 55.56%
Among White patients (N=32), the measure performance was 81.25%; and
Among patients whose race/ethnicity is Unknown (N=19), the measure performance was 73.68%

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
n/a

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
According to statistics provided by the Centers for Disease Control and Prevention, for children aged 4-17 years, 5 million children (9%) have ADHD. The percentage of children with parent-reported ADHD increased by 22% between 2003 and 2007, and rates of ADHD diagnosis increased an average of 3% per year from 1997-2006 and an average of 5.5% per year from 2003-2007. The process required to sustain appropriate treatments and successful long-term outcomes hinge on accurate and consistent ADHD diagnosis. Hoagwood, et al, found that about 50% of children with ADHD seen in practice settings obtain care that matches guidelines of the found that about 50% of children with ADHD seen in practice setting obtain care that matches guidelines of the American Academy of Child and Adolescent Psychiatry.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
Centers for Disease Control and Prevention. Summary health statistics for U.S. children: National health interview survey, 2009. Vital and Health Statistics Series. 2010;10(247).

Hoagwood K, Kelleher KJ, Feil M, Comer DM. Treatment services for children with ADHD: a national perspective. J Am Acad Child Adolesc Psychiatry. 2000;39(2):198-206.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
n/a

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Behavioral Health : Attention Deficit Hyperactivity Disorder (ADHD)

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Health and Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.ahrq.gov/sites/default/files/wysiwyg/policymakers/chipra/factsheets/chipra_fs_14-p005-1-ef.pdf

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ICD_9_-_10_Codes_-_ADHD_Accurate_Diagnosis.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients whose diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) was based on a clinical exam with a physician or other healthcare professional, as appropriate which includes: confirmation of functional impairment in two or more settings (1) AND assessment of core symptoms of ADHD including inattention, hyperactivity, and impulsivity, either through use of a validated diagnostic tool (2) based on DMS-IV-TR criteria for ADHD or through direct assessment of the patient.

(1) Settings: Includes home, school, and community
(2) Validated diagnostic tool used may include any of the following examples, all of which are based on the DSM-IV criteria for ADHD:
Conners Rating Scales
Barkley ADHD Rating Scale
Vanderbilt Parent and Teacher Assessment Scales
ADHD Rating Scale-IV (DuPaul)
Swanson, Nolan, and Pelham-IV (SNAP IV) Questionnaire

Other ADHD diagnostic tools may be determined valid based on DSM-IV criteria and therefore would be acceptable for this measure and will be added to the list at periodic updates.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back*

to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)
Data can be aggregated for this measure annually.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
n/a

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 4 through 18 years with a diagnosis of ADHD.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Patients with an ADHD diagnosis can be identified by looking for an ADHD diagnostic code in the patient medical record.

ICD-9 ADHD codes include:
314.01 Combined Type
314.01 Predominantly Hyperactive-Impulsive Type
314.00 Predominantly Inattentive Type
314.9 Attention-Deficit/Hyperactivity Disorder NOS

ICD-10 ADHD Codes:
F90-

Patient age can be determined by referencing the patient visit during which the ADHD diagnosis was made, and calculating the patient age at time of diagnosis using the patient date of birth and the date of ADHD diagnosis.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
n/a

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
n/a

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
With a large enough sample, it is possible to stratify the measure results, but it is not required. This measure can be stratified by gender (male, female), race (Asian, Black, Hispanic, White, Unknown), language (English, Spanish), age group (4-5, 6-10, 11-14, 15-18), and Insurance type (Medicaid, Private, Self-Pay, Unknown).

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
To calculate the measure using manual chart abstraction, the following algorithm was followed:

1. Select charts using a simple randomization method for sampling or assess the population by identifying patients diagnosed with ADHD;
2. Review criteria for inclusion including age of child and date of diagnosis;
3. If desired, collect demographics and elements for equity assessment including gender, race/ethnicity, language preference, insurance status/type, and age;
4. Review and document measure elements in the ADHD Measure Abstraction Tool;
5. Record summary of measure elements;
6. Note relevant comments; and
7. Identify which patients met the numerator criteria who also met the denominator criteria. These patients have met the measure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
In testing, the measures were tested in 4 sites. At one site (Lurie), a sample of the population was assessed for this measure. Computer-generated random numbers were used for simple randomization of charts. At the other three sites (Advocate Hope, Advocate Lutheran, and Stroger), the universe was assessed due to the low number of patients that met the denominator criteria in the timeframe provided.

We recommend the use of any simple randomization method for sampling.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
n/a

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
If data required to compute the denominator are missing, the patient is excluded from the measure entirely. As denominator elements include age and a diagnosis of ADHD, we do not expect that many patients who should have been included in the measure will be excluded due to missing data elements. If data required to compute the numerator are missing, the patient in included in the denominator but not the numerator. In this case, the care represented in the chart has not met the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This data was collected using a Chart Abstraction Tool created in MS Excel. The tool is attached in Appendix A.1 in PDF format.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic, Behavioral Health/Psychiatric : Inpatient
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
n/a

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
ADHD_1_nqf_testing_attachment_10-9.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Accurate ADHD Diagnosis
Date of Submission: 9/30/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

N/A

1.3. What are the dates of the data used in testing? December 2011 – June 2012

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

In testing, the measures were tested in 4 sites: At one site (Lurie), a **sample** of the population was assessed for this measure. Computer-generated random numbers were used for simple randomization of charts. At the other three sites

(Advocate Hope, Advocate Lutheran, and Stroger), the **universe** was assessed due to the low number of patients that met the denominator criteria in the timeframe provided. We recommend the use of any simple randomization method for sampling.

Reliability testing was performed in four sites participating in the Chicago Pediatric Quality and Safety Consortium (CPQSC). The CPQSC is a group of Chicago-area hospitals with large pediatric inpatient and ambulatory practices that have come together to improve the quality and safety of medical care delivered to children and their families. Hospitals represented in the CPQSC include a large suburban teaching hospital, a dedicated urban children's hospital, and a large freestanding public safety net hospital. Testing was conducting using paper and electronic medical records from the hospitals ambulatory outpatient clinic networks.

The AAP ADHD Guideline was released in 2011 and testing of this measure was conducted in 2012. Due to the alignment of this measure with the recommendations made in the AAP ADHD Guideline, the team limited the period of testing to post-Guideline release, which resulted in a relatively short time period of retrospective reviews and resulted in a smaller sample size, which can be considered a limitation to the testing.

A total of 118 patient charts were included in the analysis from four test sites in the Chicago area

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
A total of 118 patients were included in the testing and analysis from four sites in the Chicago area. Please see the table below for patient characteristics.

Summary Statistics: ADHD Patients (4 Sites)		
	N (118)	%
Gender		
Male	83	70.34
Female	35	29.66
Current Age Mean, Std Dev (9.23, 3.56)		
4-5	13	11.02%
6-10	78	66.10%
11-14	13	11.02%
15-18	14	11.86%
Insurance		
Medicaid	82	69.49%
Private	30	25.42%
Self Pay	4	3.39%
UTD	2	1.69%
Race		

Asian	3	2.54%
Black	34	28.81%
Hispanic	28	23.73%
Unknown	19	16.10%
White	34	28.81%
Language		
English	109	92.37%
Spanish	9	7.63%

Like the general population, ADHD was more prevalent in males than females and rates of ADHD were high in the Medicaid population.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
Reliability testing was performed in four sites participating in the Chicago Pediatric Quality and Safety Consortium (CPQSC). The CPQSC is a group of Chicago-area hospitals with large pediatric inpatient and ambulatory practices that have come together to improve the quality and safety of medical care delivered to children and their families.

Content validity was assessed for the measure through reconciliation by stakeholders participating in a public comment period as well as members of the ADHD Expert Work Group. Face validity was assessed through a public comment period involving stakeholders and experts in the field.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
Patient level variables include age, race/ethnicity, gender, primary language, and insurance provider as a proxy for socioeconomic status.

2a2. RELIABILITY TESTING
Note*: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
Reliability testing was conducted in four sites and each site identified two research nurses, experienced in chart abstraction, who received training on how to identify and select the charts for inclusion in the testing of the construction of this measure through manual chart abstraction. A chart abstraction tool and algorithm were

developed by the ADHD Quality Measures Leadership Team and were provided to the research nurses during training. The research nurses were trained to abstract measure elements and to use the chart abstraction tool.

The research nurses were instructed to identify a retrospective set of 25-40 charts, between December 2011 and June 2012 that matched the denominator criteria. The research nurses abstracted the relevant elements from the charts regarding demographics, numerator elements, denominator elements, and noted any pertaining exclusions according to the algorithm as follows:

1. Select charts: patients diagnosed with ADHD;
2. Select charts using a simple randomization method for sampling or assess the population by identifying patients diagnosed with ADHD;
3. Review criteria for inclusion: age, date of diagnosis;
4. Collect demographics and elements for equity assessment: gender, race/ethnicity, language preference, insurance status/type, age;
5. Review and document measure elements in the ADHD Chart Abstraction Tool;
6. Record summary of measure elements; and
7. Note relevant comments.

Data analyses included construction of the measure (performance measure score) and assessment of the agreement (critical data elements used in the measure). Analysis of critical elements was performed by comparing the two research nurses responses to whether an element was present in the chart and calculating a percent agreement and kappa.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

For analysis of critical elements, the research nurses agreed 72.88% of the time that there was evidence of an ADHD diagnostic clinical exam by the physician in the chart. The research nurses agreed 76.27% of the time that there was evidence in the chart of an assessment of core symptoms of ADHD including inattention, hyperactivity, and impulsivity through a validated diagnostic tool. The research nurses agreed 72.04% of the time that there was evidence in the chart of assessment of core symptoms of ADHD including inattention, hyperactivity, and impulsivity based on DSM-IV criteria for ADHD through direct assessment of the patient. The research nurses agreed 81.98% of the time that there was evidence in the chart of assessment of impairment in two settings.

Overall, the research nurses agreed 65.77% of the time on the ADHD measure including clinical exam by a physician, evidence of impairment in two settings, and either assessment through a validated tool or direct assessment. Please see the table below:

Agreement ADHD Measure #1 (4 Sites)

	N	Agree %	Both Yes %	Both No %	Kappa
Evidence of ADHD diagnostic clinical exam by physician in the chart (Yes - 1/No - 2)	118	72.88%	61.86%	11.02%	0.2696
Evidence in the chart of assessment of Core symptoms of ADHD including inattention, hyperactivity and impulsivity through a validated diagnostic tool AND through direct assessment of the patient (Yes – 1/No -	118	61.86%	25.42%	23.73%	0.6025

2)					
Evidence in the chart of assessment of impairment in 2 settings (Yes - 1/No - 2)	111	81.98%	74.77%	7.21%	0.3613
Overall ADHD Measure (Clinical Exam by MD, Evidence of Impairment in 2 Settings, and either assessment through validate tool or direct assessment)	111	65.77%	45.95%	19.82%	0.2675

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
As a kappa score of 0.21-0.40 indicates "fair" agreement and a kappa between 0.41-0.60 indicates moderate agreement, the critical data element testing indicates that there are substantial opportunities for performance improvement in the process of accurate diagnosis of ADHD.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☐ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Statement of intent for the selection of ICD-10 codes: Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.
Content Validity
This measure was assessed for content validity through reconciliation of stakeholders and subject matter experts, specifically by the panel of representatives participating in the ADHD Expert Work Group during the measure development process with the evidence in the and the recommendations in the 2011 AAP ADHD Guideline. This subject matter expert panel consisted of 25 members, with representation from pediatricians, pediatric neurologists, social workers, school psychologists, family physicians, school-based learning disability specialists, teachers, parents, consumer representatives, child and adolescent psychologists, occupational therapists, clinical psychologists, pediatric nurses and measure methodologists. Consensus was reached on the significance, definition, and specifcation of the measure elements and their relevance to the 2011 AAP ADHD Guideline including performance, specifications, missing data, and relevance to (per the Guidance for Evaluating Validity Algorithm Box 2).

Face Validity

Input on the face validity of draft measures was obtained through a 21-day public comment period convened by the AMA-PCPI. All comments received were reviewed by the expert work group and the measure was refined as needed. Input from these stakeholders was instrumental in ensuring this measure appropriately addressed the recommendations in the 2011 AAP ADHD Guideline and the needs of children diagnosed with ADHD as well as responded to the needs of children in Medicaid.

Throughout the measure development process, we presented the ADHD measures to the ADHD Expert Work Group and the representative from Illinois Health and Family Services (and the Illinois State Medicaid agency) who oversees quality measure use and application and solicited feedback on importance, relevance, understandability, usability, and performance. The recognized experts determined agreement that the computed measure score as specified can be used to distinguish good and poor quality care (Guidance for Evaluating Validity algorithm Box 4). The experts also agreed that potential threats to validity are not a problem and results are not biased (Guideline for Evaluating Validity algorithm Box 5).

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face Validity
Active participants in the Public Comment period included the Center for Advanced Pediatrics, Children's Health Specialists, Neurology and Sleep Medicine P.C., American Occupational Therapy Association, American Academy of Pediatrics (AAP), AAP members, PMAG, AMA-PCPI Specifications Staff, and many other organizations.

The following key issues and recommendations resulted from the Public Comment period.

Measure	Summary of Responses in Support/Non-Support	Summary of Key Issues, Recommendations and Questions for Work Group
Measure #1: Accurate ADHD Diagnosis	Support – 8 Support w/ definitional modifications to clarify language and intent consistent with the measure – 3 Question asking to consider if there are any exceptions - 1	- Support 1. Recommendation that evaluation should differentiate between co-morbidities **(Center for Advanced Pediatrics and Neurology and Sleep Medicine, P.C.)** 2. Agreement with use of a validated tool or direct assessment of DSM-IV criteria (**PMAG and AAP Individual Member)** 3. Others solely commented their support of the measure **(American Academy of Family Physicians, Children's Hospital of Philadelphia, ISMA, SHIRE)** - **Support w/ definitional modifications to clarify language and intent** 1. Recommendation to revise several points in numerator and denominator to revise language, improve clarity **(AAP)** 2. Language to be more specific to "other clinicians" and include occupational therapists **(American Occupational Therapy Association)** - Question asking to consider if there are any exceptions with no recommendations – 1. This had already been done and the determination was no **(AMA-PCPI Specifications Staff)**

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The Expert Work Group reached consensus that the results of the public comment reflected evidence base that "face value": 1) the elements of the ADHD Accurate Diagnosis measure were in agreement with the 2011 AAP ADHD Guideline, and 2) the process of Accurate Diagnosis was reflected in the measure so the performance of this measure can distinguish good quality ADHD care from poor quality pediatric-related ADHD care.

2b3. EXCLUSIONS ANALYSIS
NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis*

was used)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by risk categories**
☒ **Other,** stratification is not required but may be performed with an adequate sample size. Examples of potential stratification variables include age, sex, race/ethnicity, preferred language, insurance status/provider type

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.
N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
N/A

2b4.4a. What were the statistical results of the analyses used to select risk factors?
N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
N/A
2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
N/A

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
N/A

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
N/A

2b4.9. Results of Risk Stratification Analysis:
N/A**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
N/A

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
The performance of this measure as a chart abstraction measure was tested at four outpatient clinician office networks (N=118) in the Chicago area. The relatively small sample size is due to the alignment of this measure with the recommendations made in the 2011 AAP ADHD Guideline resulting in a short post-Guideline testing period.

The performance of the measure at the sites was as follows:

- Site 1 - 63.41% (95% CI: 48.67% - 78.16%)
- Site 2 – 71.88% (95% CI: 56.30% - 87.45%)
- Site 3 – 90.91% (95% CI: 73.92% - 100.00%)
- Site 4 – 92.86% (95% CI: 83.32% - 100.00%)

Measure performance was calculated by race/ethnicity group for the data collected during testing and the results are as follows:

- Among Asian patients (N=3), the measure performance was 66.67%;
- Among Black patients (N=31), the measure performance was 54.84%;
- Among Hispanic patients (N=27), the measure performance was 55.56%;
- Among White patients (N=32), the measure performance was 81.25%; and
- Among patients whose race/ethnicity is Unknown (N=19), the measure performance was 73.68%.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)The performance of this measure as a chart abstraction measure was tested at four outpatient clinician office networks (N=118) in the Chicago area. The relatively small sample size is due to the alignment of this measure with the recommendations made in the 2011 AAP ADHD Guideline resulting in a short post-Guideline testing period.

The performance of the measure at the sites was as follows:

- Site 1 - 63.41% (95% CI: 48.67% - 78.16%)
- Site 2 – 71.88% (95% CI: 56.30% - 87.45%)
- Site 3 – 90.91% (95% CI: 73.92% - 100.00%)
- Site 4 – 92.86% (95% CI: 83.32% - 100.00%)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
As the results of the measure varied at our four sites from 93.86% to 63.41%, this measure can be used to differentiate the quality of care provided by practices and facilities based on the results of this measure. These results represent significant gaps in performance similar to the ones found in the literature.

Reliability results are likely to be improved with a larger sample size.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
In order to meet the denominator criteria for the measure, all components of the denominator must be present in the patient chart.

In order to meet the numerator criteria for the measure, patients must have a diagnosis of ADHD based on a clinical exam with a physician including confirmation of functional impairment in two or more settings and assessment of core symptoms of ADHD. If data is missing, it is assumed that the care element was not provided and the patient chart does not meet numerator criteria.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Across all sites, 118 charts were abstracted, 112 met denominator criteria and 74 charts met numerator criteria indicating that denominator data was missing in 6 charts and numerator data was missing in 38 charts.

Denominator criteria were missing in approximately 5%of cases and numerator criteria were missing in 34% of cases.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
This measure performed as expected given the gap between guideline recommended ADHD care and common practice as evidenced by the literature and the ADHD Expert Work group's commentary of the measure performance.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

All of the elements for construction of this measure are available in paper records or electronic medical records and the measure can be constructed feasibly and reliably through manual chart abstraction of the elements. There are improvements that could facilitate collection and reporting of this measure through the availability of specific fields or workflow documents to indicate the location of specific elements such as the documentation of the use of a specific validated tool and the setting in which the evaluation took place in queriable fields.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Manual chart abstraction of this measure using either paper or electronic medical records is feasible and reliable. All of the elements for construction of this measure were present and able to be reliably abstracted through manual chart abstraction of paper or electronic medical records. New elements based on the 2011, AAP ADHD Guideline recommendations for a new standard of care for pediatric patients diagnosed with ADHD, such as the documentation of the use of specific validated tools, were indicated in the Notes sections of charts, whether paper or electronic, and the tools themselves were available for abstraction as they were added to the paper charts, scanned into electronic charts, or found in queriable fields.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk*

model, programming code, algorithm).
n/a

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Payment Program	Professional Certification or Recognition Program The American Board of Pediatrics, Performance Improvement Model https://pim.abp.org/adhd_initial/global/demo/ Quality Improvement (Internal to the specific organization) The American Board of Pediatrics, Performance Improvement Model https://pim.abp.org/adhd_initial/global/demo/

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

The American Board of Pediatrics (ABP), Performance Improvement Module, ADHD Initial Diagnosis

The ABP built the ADHD Accurate Diagnosis measure into their Maintenance of Certification (MOC) Part IV ADHD Initial Diagnosis Performance Improvement Model (PIM) for use for both pediatrician performance improvement and as a requirement for professional recertification.

The purpose of the organization of the ABP PIM enables pediatricians to initially assess their own performance for ADHD accurate diagnosis and then to implement improvements in clinical care using quality improvement methods. The PIMs guide pediatricians through the process of collecting and analyzing practice data over time and documenting improved quality of care. The ADHD Initial Diagnosis PIM specifically aims to help pediatricians assess and provide ADHD guideline-based care, learn the basics of quality improvement, and to meet requirements for MOC Part IV.

The ABP PIMs are nationwide and pediatricians across the United States participate. The ADHD Initial Diagnosis PIM was released on August 19, 2013, and since then 313 physicians have selected and completed the PIM using the ADHD Accurate Diagnosis measure.

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
n/a

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6

years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
n/a

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
n/a

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended negative consequences to individuals or populations identified during testing. Furthermore, there is no evidence of unintended negative consequences to individuals or populations since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0108 : Follow-Up Care for Children Prescribed ADHD Medication (ADD)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
While both the NQF endorsed measure and this proposed measure focus on children and adolescents with ADHD diagnoses, the currently endorsed measure is limited to children aged 6-12, while the proposed measure considers children and adolescents ages 4-18. This is because the currently endorsed measure was developed and endorsed prior to the release of the new 2011 AAP ADHD Guideline which adjusts the age groups in focus, and has not been updated since that time to match the recommendations.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** 1.04_S.25._Data_Collection_Instrument.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Academy of Pediatrics
Co.2 **Point of Contact:** Lisa, Krams, lkrams@aap.org, 847-434-4000-7663
Co.3 **Measure Developer if different from Measure Steward:** AHRQ-CMS CHIPRA Pediatric Measurement Center of Excellence (PMCoE) (under the leadership of the Northwestern University Feinberg School of Medicine)
Co.4 **Point of Contact:** Ramesh, Sachdeva, MD, PhD, JD, FAAP, rsachdeva@aap.org, 847-434-4000-7110

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This measure was developed under the leadership of the Northwestern University Feinberg School of Medicine (Site PI: Donna Woods), in their capacity as a member of the PMCoE Consortium (PMCoE PI: Ramesh Sachdeva).

AHRQ-CMS CHIPRA Pediatric Measurement Center of Excellence (PMCoE) Consortium Members:
Medical College of Wisconsin/National Outcomes Center (MCW); American Medical Association-Convened Physician Consortium for Performance Improvement (AMA-PCPI); American Board of Medical Specialties (ABMS); American Board of Pediatrics (ABP); Northwestern University Feinberg School of Medicine (Northwestern); American Academy of Pediatrics (AAP); Thomson-Reuters; TMIT Consulting, LLC; Chicago Pediatric Quality and Safety Consortium

PMCoE Staff:
Ramesh C Sachdeva, PMCoE PI, MCW
Lisa Ciesielczyk, Program Manager, MCW

V. Fan Tait, Site PI, AAP
Keri Thiessen, Project Staff, AAP
Donna Woods, Site PI, Northwestern
Nicole Muller, Project Staff, Northwestern
Lindsay DiMarco, Project Staff, Northwestern
Jin-Shei Lei, Project Staff, Northwestern
Ray Kang, Project Staff, Northwestern
Susan Magasi, Project Staff, Northwestern
Sara Alafogianis, Project Staff, AMA-PCPI
Mark Antman, Project Staff, AMA-PCPI
Amaris Crawford, Project Staff, AMA-PCPI
Kendra Hanley, Project Staff, AMA-PCPI
Molly Siegal, Project Staff, AMA-PCPI
Greg Wozniak, Project Staff, AMA-PCPI

Expert Work Group Members:
Ted Abernathy, Pediatrician, Private Practice of Pediatrics and Adolescent Medicine
Betsy Brooks, Pediatrician, Holyoke Pediatric Associates
Lawrence Brown, Pediatric Neurologist, Children´s Hospital of Philadelphia
Mirean Coleman, Social Worker, National Association of Social Workers
Stephen Downs, Pediatrician, Children´s Health Services Research
George DuPaul, School Psychologist, Lehigh University
Mirian Earls, Developmental-Behavioral Pediatrician, Guildord Child Health
Jeff Epstein, Clinical Psychologist, Cincinnati Children´s Hospital Medical Center
Theodore G Ganiats, Family Physician, University of California San Diego
Jane Hannah, School-based Learning Disability Specialist, Curry Ingram Academy
Romana Hasnain-Wynia, Healthcare Equity Expert, Northwestern University Institute for Healthcare Studies
Steven Kairys, Pediatrician, Jersey Shore Medical Center
Beth Kaplanek, Parent, Children & Adults w/Attention Deficit Disorders (CHADD)
M. Ammar Katerji, Pediatric Neurologist, Advocate Hope Children´s Hospital
Shelly Lane, Occupational Therapist, Virginia Commonwealth University
Nancy Marek, Pediatric Nurse, Advocate Hope Children´s Hospital
Paul Miles, Maintenance of Certification Expert, American Board of Pediatrics
Patrice Mozee-Russell, Teacher, Children & Adults w/Attention Deficit Disorders (CHADD)
Karen Pierce, Child and Adolescent Psychiatrist, Children´s Memorial Hospital/Northwestern University
Sandra Rief, School-based Learning Disability Specialist, Children & Adults w/Attention Deficit Disorders (CHADD)
Clarke Ross, Parent, American Association on Health and Disability
Adrian Sandler, Developmental-Behavioral Pediatrician, Mission Children´s Hospital
Marcia Slomowitz, Child and Adolescent Psychiatrist, Northwestern Memorial Hospital
Laurel Stine, Consumer Representative, Bazelon Center for Mental Health Law
Mark Wolraich, Developmental-Behavioral Pediatrician, University of Oklahoma Child Study Center

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

ICD-9 and ICD-10 Codes

ICD-9 Diagnosis Code	Description ICD-9	ICD-10-CM	Description ICD-10
314.01	Predominantly Hyperactive-Impulse Type	F90.1	Attention-deficit hyperactivity disorder, predominantly hyperactive type
314.01	Combined Type	F90.2	Attention-deficit hyperactivity disorder, combined type
314.00	Predominantly Inattentive Type	F90.0	Attention-deficit hyperactivity disorder, predominantly inattentive type
314.9	Attention-Deficit/Hyperactivity Disorder NOS	F90.9	Attention-deficit hyperactivity disorder, unspecified type

We used www.ICD10Data.com for the conversions

Data Collection Instrument

Patient ID	Race	Ethnicity	Gender	Payer	Preferred Language	Age	
1	White	Non-Hispanic	Female	Medicaid	English	12	
2	Black	Non-hispanic	Male	Medicaid	English	5	
3	White	Hispanic	Male	Private	Spanish	8	
4	Asian Pacific Iland	Non-Hispanic	Male	Private	Chinese	10	
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							

ADHD Diagnosis Date (range 12/11 - 6/12) **Please remove this column before submitting	Patient diagnosed between Dec 2011 and June 2012 (Yes-1/No -2)	Evidence of ADHD diagnostic clinical exam by physician in the chart (Yes - 1/No - 2)	Evidence in the chart of assessment of Core symptomes of ADHD including inattention, hyperactivity and impulsivity through a validated diagnostic tool (Yes - 1/No - 2)	Evidence in the chart of assessment of Core symptomes of ADHD including inattention, hyperactivity and impulsivity based on DSM-IV criteria for ADHD through direct assessment of the patient (Yes - 1/No - 2)	If no, dimensions with no documentation/evidence of assessment
12/12/2011		1	2	1	
3/3/2012		1	1	2	
1/17/2012		2	1	2	
5/6/2012		1	2	2	I1, I5, H4, Im2

Evidence in the chart of assessment of impairment in 2 settings (Yes - 1/No - 2)	Evidence in the chart of assessment of symptoms present for at least 6 months (Yes - 1/No - 2)	Patient is 4 or 5 at time of diagnosis (Yes - 1/No - 2) **Birthdate age ranges should be between: 7/1/2006 - 12/1/2007	ADHD-focused evidence-based behavior therapy prescribed (Yes - 1/No - 2)	Behavior Therapy Component 1 Present: Treatment is directed to parent or caregiver (guarding, teacher, child care worker) (Yes - 1/No - 2)	Behavior Therapy Component 2 Present: Training is provided in parent or caregiver-administered behavior modification (Yes - 1/No - 2)
1	1	2			
1	2	1	1		
1	1	2			
2	2	2			

Behavior Therapy Component 3 Present: Treatment does not involve child-directed play therapy (Yes - 1/No - 2)	ADHD-focused evidence-based behavior therapy prescription date **Please remove this column before submitting	ADHD treatment medication prescribed (Yes - 1/No - 2)	ADHD treatment medication prescription date **Please remove this column before submitting	For patients 4 - 5 behaior therapy was prescribed as first line treatment prior to medication prescription (Yes - 1/No - 2)

4/3/2012 1 4/3-/12 1

ADHD Measure 2 Exclusion - Documentation of medical reason(s) for not prescribing behavior therapy as first line treatment (eg, (eg patient with multiple psychiatric conditions referred to other provider), or patient determined to be at risk for harming themselves or others) (Yes - 1/No - 2)	ADHD Measure 2 Exclusion - Documentation of system reason(s) for not prescribing behavior therapy as first line treatment (eg, lack of access to behavior therapy) (Yes - 1/No - 2)

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2818
Measure Title: ADHD Chronic Care Follow-up
Measure Steward: American Academy of Pediatrics
Brief Description of Measure: Percentage of patients aged 4 through 18 years with a primary or secondary diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) in the year prior to the measurement year who have at least one follow-up visit in the measurement year with ADHD as the primary diagnosis
Developer Rationale: According to statistics provided by the Centers for Disease Control and Prevention, for children aged 4-17 years, 5 million children (9%) have ADHD, the percentage of children with parent-reported ADHD increased by 22% between 2003 and 2007, and rates of ADHD diagnosis increased an average of 3% per year from 1997 to 2006 and an average of 5.5% per year from 2003 to 2007 (1). ADHD has a multidimensional effect on an individual´s daily functioning and can culminate in significant costs attributable to greater health care needs, more frequent unintentional injury, co-occurring psychiatric conditions, and productivity losses. ADHD medications can reduce symptoms but might be associated with side effects and symptoms affecting comorbidity (2). While some core problems evident in young patients with ADHD, such as hyperactivity, generally improve by adulthood; many other symptoms of the disorder may persist into adulthood including impaired social relationships, low self-concept, drug use, and education and occupational disadvantages (3). ADHD continues to cause symptoms and dysfunction in many children who have the condition and available treatments are not usually curative (1). Longitudinal studies have found that frequently, treatments are not sustained despite the fact that long-term outcomes for children with ADHD indicate that they are at greater risk of significant problems if treatment is discontinued (4). ADHD follow-up care and treatment adherence can be enhanced by improving the relationship between parents and health care providers so parents feel both involved and knoledgeable about their child´s health condition and treatment regimen. The medical home and the chronic care model both emphasize patient and family involvement in care and as a result, treating ADHD as a chronic care condition within a medical home is Guideline recommended care. (3). In November, 2011, the American Academy of Pediatrics (AAP) published a new evidence based guideline for ADHD diagnosis, follow-up, and treatment based on extensive review of the existing evidence. One recommendation with a strong level of evidence encouraged primary care clinicians to recognize ADHD as a chronic condition, including managing patient care through follow-up appointments. management of children and youth with special health care needs should follow the principles of the chronic care model and the medical home (4, 5).

There is evidence that ADHD treatment can improve the likelihood of a positive outcome and reduce the negative consequences of ADHD in the short term; however, residual benefits of pharmacological treatment may subside when medication is discontinued (6). Therefore, given that ADHD symptoms may manifest for as long as 8 years after diagnosis and that ADHD treatment has been shown to work in the short term altough it may require many modifications, regular ADHD follow-up care, per the 2011 AAP ADHD guideline, is to ensure that a child is adhering to a treatment plan.

1. Centers for Disease Control and Prevention. Summary health statistics for U.S. children: Naitonal health interview survey, 2009. Vital and Health Statistics Series. 2010; 10(247).

2. Centers for Disease Control and Prevention. Increasing prevalence of parent-reported attention deficit/hyperactivity disorder among children. Morbidity and Mortality Weekly Report. November 12, 2010/59(44); 1439-1443.

3. Ingram S, Hechtman L, Morgenstern G. Outcome issues in ADHD: adolescent and adult long-term outcom. Mental Retardation and Developmental Disabilities Research Reviews. 1999;5:243-250.

4. Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management. ADHD:

clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Pediatrics. 2011; 128(5):1-16.

5. Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents: process of care supplemental appendix. Pediatrics. 2011; SI1-SI21.

6. Barkley R, Fisher M, Edelbrock C, Smallish L. The adolescent outcome of hyperactive children diagnosed by research criteria: 1. an 8-year prospective follow-up study. J Am Acad Child Adolesc Psychiatry. 1990; 29(4):546-557.

Numerator Statement: Patients who attended at least one ADHD follow-up care visit within the calendar year.
Denominator Statement: All patients aged 4 through 18 years with a diagnosis of ADHD.
Denominator Exclusions: Documentation of medical reason(s) for not providing follow-up care (e.g., patient with multiple psychiatric conditions referred to other provider). Please see code list in section S.11.

Documentation of system reason(s) for not providing follow-up care (e.g., patient for whom the follow-up visits were not all with the same practice).

Measure Type: Process
Data Source: Administrative claims
Level of Analysis: Health Plan, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: n/a

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following information for this process measure (Level of Analysis = Health Plan, Population: National):

- The developer links follow-up visits for those with ADHD with increased treatment and, ultimately improvements in function, quality of life, decreased symptoms. Evidence for this process measure should demonstrate that when follow-up visits for ADHD occur, treatment is more likely to occur, which will lead to the desired outcomes of improved quality of life, increased functionality, and reduced symptoms.
- The measure is based on a recommendation from the 2011 American Academy of Pediatrics' *Clinical Practice Guideline for the Diagnosis, Evaluation, and Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents.*
  - "The primary care clinician should recognize ADHD as a chronic condition and, therefore, consider children and adolescents with ADHD as children and youth with special health care needs. Management of children and youth with special health care needs should follow the principles of the chronic care model and the medical home."
  - The recommendation is based on grade B evidence. This grade evidence indicates that it includes RCTs or diagnostic studies with minor limitations and overwhelmingly consistent evidence from observational studies. The recommendation is graded as "strong", meaning that it is based on high- to moderate-quality scientific evidence and a preponderance of benefit over harm.
- The developer reports that the body of evidence underlying the clinical practice guideline included three literature

reviews and one systematic review of evidence for the medical home of at least 30 studies that ranged from 1999 to 2008.

- The developer notes that longitudinal studies have demonstrated that ADHD persists for most patients throughout adolescence and adulthood, and that symptoms of inattention, particularly, continue even if symptoms of hyperactivity and impulsivity decrease over time. The evidence underlying the guideline recommendation indicates improvements in desired outcomes for children treated in a medical home model and for those whose treatment follows the tenets of the chronic care model.
- The developer reports on an additional systematic review since the guideline.
  - This review synthesizes literature on the efficacy and effectiveness of guideline-recommended care, and established the baseline for developing an outcome measure that assesses the quality of care for children with ADHD.
  - 35 studies were reviewed, with 20 rates as good and 15 as fair. Regardless of outcome measure and treatment type, symptom reduction and improvement were relatively large with effect sizes ranging from 0.15-4.57.
  - The review supports the conclusion that core symptoms of ADHD can be improved within 1 year.
  - The developer posits that the long time frame required for improvement along with literature that documents improvements after treatment indicate the need for sustained ADHD care throughout the lifespan, supporting the conclusion that that ADHD should be considered a chronic condition.
  - The developer indicates quantity, quality, and consistency of evidence encompassed by this review, but does not provide information on the grading system or definitions.
- Per the **NQF Algorithm for Evidence**, the eligible ratings are HIGH, MODERATE, or LOW because the developer identifies a systematic review that is graded and that assesses the quantity, quality, and consistency of the evidence (box 3-->4) if the Committee judges the systematic review of medical home literature is applicable. If the Committee assesses that the guideline is not directly appropriate, the additional empirical evidence provided by the developer means the eligible ratings are MODERATE or LOW (box 7-->9)

Questions for the Committee

- *Is the evidence concerning the chronic care model and the medical home model directly applicable to the measure focus (i.e., follow-up visits for ADHD)?*
- *The measure specifies "at least one follow-up visit in the measurement year." Is there evidence to support the frequency stated in the specifications as improving outcomes? Does the medical home literature apply to the specificity of the number of visits?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer provided performance results for this measure using data abstracted from the Truven MarketScan Database.
  - Among Medicaid and commercially-insured patients, 65% and 49%, respectively, had a follow-up visit with ADHD as the primary diagnosis. The developer was unable, however, to estimate individual health plan performance results using this data source.
  - In the measure testing attachment, the developer notes the Truven MarketScan data it used for testing included 569,228 enrollees (166,471 Medicaid patients and 402,757 commercially-insured patients) who met the measure denominator criteria.
- The developer notes that racial/ethnic disparities were found among the Medicaid subpopulation included in the Truven database (i.e., Blacks were more likely than non-Hispanic whites to receive follow-up care, but Hispanics and other minorities were less likely than non-Hispanic whites to receive this care). It was noted that minority children were less likely to be in the measure denominator (i.e., were less likely to have had an AHDA diagnosis). The developer did not provide any disparities information for those children who were commercially-insured.

Questions for the Committee:

- *The developer provides gap information at the commercial vs. public insurance level, but not the health plan level of*

analysis for which the measure is specified. Does the Committee believe the developer has demonstrated there is a a gap in care that warrants a national performance measure?

- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities]).*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- ADHD as a developmental disability is a chronic condition and follow-up to monitor treatment improves likelihood of treatment. Frequency of follow-up is likely dependent on the type of treatment with medication requiring more frequent follow-up for management than other interventions. Even if a child is followed in a specialty clinic, the PCP should pay attention to the presence of the disorder and assure that the child is receiving follow-up.
- The idea that ADHD is a chronic condition is well documented. The ability of care to improve long term outcome is unclear. Most studies of treatment with medications focus on short term outcomes. While benefit of continuity of care within a medical home is clear, impact in this specific chronic problem is anecdotal.
- The evidence available relates to chronic illness in general, not to ADHD; the developers infer that ADHD care will be improved by using this model. The choice of "at least one f/u visit" is also an inference but not specific to the medical home or chronic care model.
- Why only 1 visit? Is this enough?
- Strong evidence from practice guidelines and literature and systematic reviews
- Evidence provided supports that children with ADHD should be treated as a patient with chronic care needs within the medical home. SR with effect size ranging from.15 to 4.57, indicating that symptoms of ADHD can improve within 1 year time.

1b. Performance Gap.

- The performance on the measure is fairly low (~50%) at either insurance type indicating less than recommended levels of care and room for improvement. While this is listed as a plan level measure it could be a practice level measure too. It is somewhat problematic that plan distinctions couldn't be made from the data source. Disparities in care between racial and ethnic groups exist.
- Variation and disparities were established through clams-based analysis.
- Taken from a market database, using the measure as developed, shows fewer than 65% of children have a documented f/u visit with ADHD as the primary dx in the follow up year. This number has face validity.
- Rates and likelihood of diagnosis in different populations?
- Performance gaps are prevalent and some evidence of disparities for minorities
- Provided data abstracted from Truven Market Scan for medicaid and commercially insured patients. However, they were unable to estimate individual health plan performance results using this data. Information provides weak evidence that the gap in care warrants a national performance measure.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

The developer provides the following information:

- The measure Level of Analysis is Health Plan and Population: National.
- The measure uses data from administrative claims.
- This measure captures children ages 4-18 years with a primary or secondary diagnosis of ADHD.
  - Patients who are not continuously enrolled during the measurement year are excluded.

- The measure also excludes patients for "medical reasons" (i.e., those patients also diagnosed with autism, substance abuse anorexia, mood disorders, or anxiety) or for "system reasons" (the only example provided is patients for whom the follow-up visits were not all with the same practice).
- To meet the measure, the primary diagnosis for the follow-up visit must be ADHD.

- ICD-9 and ICD-10 codes to identify patients diagnosed with ADHD (the denominator) are provided. CPT codes and POS codes to identify evaluation and management (E&M) visits (the numerator) are provided. The developer notes in the supplement excel file that the ICD-10 codes were identified through use of www.ICD10data.com.
- The calculation algorithm is relatively detailed and should allow for consistent calculation of the measure.
- This measure is not risk-adjusted. The developer suggests stratification of measure results according to age group, race/ethnicity, and payer type.

Questions for the Committee

- *Are system reasons for exclusions to the measure meaningful for this measure that is specified at the health plan level of analysis?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer provides the following information:

- The developer states it conducted reliability testing at the critical element level. It does not appear, however, that empirical reliability testing was conducted. NQF guidance indicates that if data element validity testing is conducted, then additional reliability testing at the data element level is not required.
  - The developer conducted basic analysis to determine the percentage of patients with various types of E&M visits and compared those frequencies to other sources; this is not empirical reliability testing at the critical element level. ***Only if this analysis is considered an appropriate method of validity testing*** can the Steering Committee's rating of the validity testing be used as the rating for reliability testing. See the validity testing section below.
- Per the **NQF Algorithm for Reliability**: No empirical reliability testing (box 2) --> was empirical validity testing at the data element level conducted? (box 3). ***Based on the Committee's review of validity testing, the eligible ratings are MODERATE or LOW.***

Questions for the Committee

- *Was the analysis conducted an appropriate method of validity testing?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The measure focus is whether or not a follow-up visit for ADHD occurred at some point in the measurement year for children previously diagnosed with ADHD.
- The clinical practice guideline recommendation indicates that ADHD be considered a chronic condition and that those with the diagnosis as having special healthcare needs, whose management should follow the principles of the chronic care model and the medical home model.
- The specifications indicate "at least one" follow-up visit must occur to successfully achieve the measure

Question for the Committee

- *Is the link between the measure focus and specifications consistent with and appropriate for the chronic care model*

and the medical home model for children with special health care needs?

- *Is there any evidence that patients with other mental disorders should be excluded from the measure?*
- *Is there evidence to support the specification of at least one visit will improve outcomes?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer provides the following information:

- The developer describes its empirical testing as determining the percentage of patients with various types of E&M visits. It also states that "*each measure component (numerator, denominator, and exclusions) were tested based on comparison with the total ADHD population in Medicaid/CHIP and commercial insurance, respectively*".
  - Data used for this analysis were abstracted from the Truven MarketScan Database for 2010-2011. This dataset included 569,228 enrollees (166,471 Medicaid patients and 402,757 commercially-insured patients) who met the measure denominator criteria. The developer noted that these patients were representative of the U.S. population diagnosed with ADHD (i.e., similarly age, race/ethnicity).
  - NQF does not consider the analysis described to be empirical validity testing at the data element level.
    - The developer states that the Level of Analysis is Health Plan, but testing was not performed at this level.
    - There is no analysis or result to demonstrate that claims data accurately identify patients with ADHD or the various diagnosis used in the exceptions to the measure.
    - The developer states that the results found in the Truven database "*were in the range of what would be expected from current gaps in ADHD care research and expert opinion*". However, the developer did not specify what data it was comparing the results to, whether it considered these data to be the "gold standard" and why, and how strong the agreement was between the results and the gold standard—i.e., sensitivity, specificity, other measures of agreement.
- The developer reports it conducted face validity
  - NQF guidance indicates that the ***assessment of face validity of the measure score as an indication of quality*** is an acceptable method for measure validation if systematically assessed by recognized experts.
  - The developer notes that a 25-member Expert Panel helped develop the measure. However, it did not describe the process of how the Expert Panel systematically assessed whether the score from this measure will distinguish good from poor quality of care nor the data associated with the assessment, as required by NQF.
    - The developer does provide a couple of statements from the Expert Panel that the commenter believes reflects a commentary on the measure score as an indicator of quality.
    - The developer provides additional statements from the Expert Panel that address performance gap and measure specifications, and thus do not conform to NQF's requirements for face validity.
- The developer also noted that face validity was assessed via a public commenting period, although no description of this process was provided. It is unclear whether this effort opinion of experts, as needed for a face validity assessment. One question asked whether the measure provides for fair comparisons among health plans and the developer provided an answer; however, the developer did not indicate how many commenters agreed with this answer.
- Per the NQF Algorithm for Validity: No empirical testing (box 3) → face validity assessed at the **level of the computed measure score** (box 4), the eligible ratings are MODERATE or LOW.

Questions for the Committee

- *Do you know of any studies that have validated the sensitivity and specificity of administrative claims in identifying those with ADHD and those included in the exceptions to this measure? If so, additional information may be considered so that the comparative frequency analysis may be sufficient to meet the requirements of validity testing at the data element level provided appropriate statistical analyses are performed available (i.e., sensitivity, specificity, other statistical measures of agreement). NOTE: The rating from the Committee's assessment here carries over to reliability testing. If the described testing is not appropriate, the rating of LOW or INSUFFICENT*

carries to reliability.
- *Has the developer assessed face validity at the level of computed measure score, as required by NQF guidance?*
- *Is a Level of Analysis = Health Plan appropriate given the described testing?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- This measure excludes patients for "medical reasons" (i.e., those patients also diagnosed with autism, substance abuse anorexia, mood disorders, or anxiety) or for "system reasons" (the only example provided is patients for whom the follow-up visits were not all with the same practice).
- The developer provided no data to indicate the frequency of exclusions applied to the measure, nor any discussion regarding the need for these exclusions other than noting that they were determined during development of the measure.

Questions for the Committee:
- *Are system reasons for exclusions to the measure meaningful for this measure that is specified at the health plan level of analysis?*
- *Do you agree that patients with other mental disorders should be excluded from the measure?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment:

- This measure is not risk-adjusted.

Questions for the Committee
- *Even though this is a process measure, is there a conceptual reason that it should be risk-adjusted (e.g., for SDS or other factors)?*
- *Do you agree that this measure should not be risk-adjusted?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer presents overall differences in performance between Medicaid and commercially-insured patients (65% and 49%, respectively) but additional statistical results (e.g., confidence intervals) were not provided. The developer does not demonstrate that differences in results among the various Medicaid or commercial plans exist.
- The developer states the Level of Analysis is health plan and population: national. The developer provides no data to indicate the measure identifies meaningful differences among health plans.
- While the developer shows that race is statistically associated with performance at the patient-level, this analysis does not demonstrate that differences among plans can be distinguished.

Question for the Committee:
- *Do you know of other data that demonstrate that follow-up visits for patients with ADHD differ among health plans?*

2b6. Comparability of data sources/methods:

- Because this measure has only one set of specifications (i.e., for claims data), this section is not applicable.

2b7. Missing Data

- The developer notes that just over half of children ages 4-18 in the Truven Database were excluded from the measure due to non-continuous enrollment. However, no data were presented to indicate the extent of missing data for those data elements used to calculate the measure. Typically, however, diagnosis and procedure data are seldom missing in administrative claims data.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- Given that coding errors occur relatively frequently, some evidence that a reasonable number of claims for ADHD follow-up are actually associated with that might have been helpful (e.g. sample chart review or other evidence supporting using claims analysis for this).
- The logic seems clear.
- This measure is likely to be consistently implemented.

2a2. Reliability testing

- The exclusions make sense as they are conditions in which ADHD may be present, but may not be listed as the primary diagnosis for follow-up and therefore not picked up in the data search. Measuring a visit by the claims for requires less interpretation although it cannot assure that the ADHD treatment is appropriate only that it was coded as the reason for the visit. As well discussions about ADHD treatment might occur at other visits such as well child visits, but not be billed as such.
- Concerned that some the codes in the inclusion set are for Inpatient Hospital Visits at Mental Health Centers (99221-99223, 99231-99233, 99238, 99239, 99251-99255) are unlikely to be children with simple ADHD (without a co-morbid condition).
- Doesn't meet validity criteria. Some exclusions seem vague, such as "system reasons" Concerned re numerator that the PRIMARY dx must be ADHD, whereas for diagnosis, it can be the primary or secondary diagnosis. Reliability testing was not adequate per NQF criteria.
- Reliability testing is insufficient.
- Data provided to support reliability testing was weak.

2b1. Validity Specifications

- Treatment of this as a chronic disease and using follow-up as a measure of that works. There are many disorders that are present comorbidity with ADHD and follow-up in those cases doesn't necessarily mean that the follow-up is for that condition. While exclusions were made for "medical" reasons there were also claims that indicated that some disorders that might have been in those exclusions were found in the sample.
- The specifications include visits that do not occur in the context of the medical home, which is inconsistent with the rationale presented for the measure.
- Links between the specific measure and the evidence are very weak, although it makes intuitive sense that follow up should occur.
- The presence of another mental disorder would not preclude the need for follow up. Strength of the evidence for merely one follow-up visit?

2b2. Validity Testing

- Health Plan is not really the appropriate level. This is more at the Commercial insurance versus Medicaid level. There are studies looking at the validity of using claims data for autism, but not for ADHD (or at least not that I have found in a recent search). The "experts" were clearly not all treatment experts (educators and social workers are unlikely to be providing follow-up treatment) and as such unfamiliar with treatment and follow-up as acknowledged in the validity description.
- Validity testing did not include testing at the plan level, and comparison groups were not clearly identified.
- Testing did not occur at the level at which the measure would be used. Face validity was by committee consensus, without clear specification of criteria. No criteria for public comments.
- Validity testing that occurred is not sufficiently described and would be a Low rating.
- Data provided says it was tested at the health plan level but that does not seem to be the case.
- Face validity was conducted per developer but the information provided lacks detail on how that was done.

2b3-2b7. Threats to Validity

- Exclusions based on diagnoses that might confound the reason for the follow-up make sense although most of them would be considered chronic conditions and the model of care with follow-up also appropriate. It is cleaner to measure just ADHD and not some of the other mental health disorders. Risk adjustment doesn't seem relevant since the measure is a process one and dependent on practitioner practice and not SDS. Also given that there might be barriers for some populations that do not exist for others, knowing the actual data without risk adjusting allows for better benchmarking.

- No, since it is pulled from claims data.
- Probably not since it is administrative data needed for billing.
- Patient with ADHD and no other exclusionary conditions--how are they addressed if they are not followed for ADHD at PCP?
- Unlikely to suffer from missing data.
- Data is collected from administrative claims data using diagnostic and procedure data. there is a potential to have missing data due to error.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The data source for this measure is administrative claims.
- All data elements are in defined fields in electronic claims.
- The developer reported that CPT code 96110 was not found to be reliable or valid as a method for assessment of standardized tool use for establishing the ADHD diagnosis.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *How likely is it that the required data elements available in structured fields in EHRs*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Claims data is pretty standard which is helpful. CPT code 96110 is general developmental screening and used in multiple ways and so is not specific to ADHD or behavior assessment. 96127 is more specific to behavioral assessment that might be appropriate in ADHD, but also is not used exclusively for that. Primary diagnosis is also somewhat problematic especially in children with other more significant diagnoses. As well visits in which follow-up for ADHD occurred simultaneously with another event may not be captured.
- All are generated routinely.
- All in administrative system. Miscoding would be the only problem. However requirement for f/u visit to have a primary diagnosis of ADHD while dx visit can have it as a secondary dx seems odd.
- Easily extracted due to coding.
- Data source is administrative claims data which is usually in EMR.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- This measure is not in use.
- The developer has not indicated any specific plans for the measure's use in public reporting and payment programs.
- The developer stated no unintended consequences to individuals or populations were identified during testing.

Questions for the Committee*:*

- *Can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Is this measure appropriate for accountability purposes?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- This measure broadens ADHD follow-up to those children who might not be treated with stimulant medication,

but there are some problems with assuring that is what is being captured. Follow-up is important regardless of stimulant medication use (which is measure #0108) as stimulant medication is not the only appropriate treatment for ADHD and there are children for which it may not be the appropriate first line treatment (children <6 years of age) or for whom is medically contraindicated (some kids with heart rhythm conditions).

- Not yet publicly reported. Concerned that the measure as specified does not capture continuity of care, which is an essential component of chronic disease management in the PCMH.
- A follow-up visit would increase health care consumption, but the benefits of evaluating treatment outweigh the risks.
- Not currently being used.

Criterion 5: Related and Competing Measures
• 0108 : Follow-Up Care for Children Prescribed ADHD Medication (ADD) (NQF-endorsed) is related to this measure. • Both measures focus on children and adolescents ADHD follow-up, however, this measure considers children and adolescents ages 4-18 and focuses on accurate diagnosis. Measure # 0108 considers children ages 6-12 with a new prescription for ADHD medication who had at least three follow-up care visits within a 10-month period, one of which is within 30 days of when the first ADHD medication was dispensed.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: ADHD Chronic Care Follow-up

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting

PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Chronic care follow-up for children diagnosed with ADHD

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

This measure is based on the evidence that ADHD continues to cause symptoms and dysfunction in many children who have the condition over long periods of time, even into adulthood, and that the treatments available address symptoms and function but are usually not curative. Longitudinal studies have found that, frequently, treatments are not sustained despite the fact that long term outcomes for children with ADHD indicate that they are at greater risk of significant problems if they discontinue treatment. Patients with ADHD who receive follow-up visits are more likely to receive treatment, which in turn can improve function, quality of life, and reduce symptoms. The primary care clinician should recognize ADHD as a chronic condition and therefore, consider children and adolescents with ADHD as children and youth with special health care needs. Management of children and youth with special health care needs should follow the principles of the chronic care model and the medical home per the 2011 AAP ADHD Guidelines. This includes appropriate follow-up care which encourages sustained treatment, improves function and quality of life, and reduces symptoms.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management, Wolraich M, Brown L, Brown RT, DuPaul G, Earls M, Feldman HM, Ganiats TG, Kaplanek B, Meyer B, Perrin J, Pierce K, Reiff M, Stein MT, Visser S. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. *Pediatrics*. 2011;128(5):1007-1022.

http://pediatrics.aappublications.org/content/early/2011/10/14/peds.2011-2654

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

The guideline recommendation of interest is Action Statement 4 on page 8, "The primary care clinician should recognize ADHD as a chronic condition and, therefore, consider children and adolescents with ADHD as children and youth with special health care needs. Management of children and youth with special health care needs should follow the principles of the chronic care model and the medical home."

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

The grade assigned is B/strong recommendation. The definition of this grade is, "RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies." This level of evidence is based on high- to moderate-quality scientific evidence and a preponderance of benefit over harm.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

The other grades are as follows:

A. Well-designed RCTs or diagnostic studies on relevant population (strong recommendation)

B. RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies (strong recommendation/recommendation)
C. Observational studies (case control and cohort design) (recommendation)
D. Expert opinion, case reports, reasoning from first principles (option)
X. Exceptional situations in which validating studies cannot be performed and there is a clear preponderance of benefit or harm (strong recommendation/recommendation)

A strong recommendation or recommendation statement is based on high- to moderate-quality scientific evidence and a preponderance of benefit over harm. Option-level action statements are based on lesser-quality or limited data and expert consensus or high-quality of evidence with a balance between benefits and harms. A health care provider might or might not wish to implement option recommendations in his or her practice.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

American Academy of Pediatrics, Steering Committee on Quality Improvement. Classifying recommendations for clinical practice guidelines. *Pediatrics*. 2004;114(3):874-877.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The evidence review focused on assessment, diagnosis, and treatment of children with ADHD.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The grade assigned is B/strong recommendation. The definition of this grade is, "RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies."

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

The other grades are as follows:

A. Well-designed RCTs or diagnostic studies on relevant population
B. See above
C. Observational studies (case control and cohort design)
D. Expert opinion, case reports, reasoning from first principles
X. Exceptional situations in which validating studies cannot be performed and there is a clear preponderance of benefit or harm.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1999-2008

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

In the body of evidence, the following study designs are included:

- 3 literature reviews
- 1 systematic review

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

A research review of long-term follow-up studies indicated very strongly that ADHD is a chronic condition with symptoms continuing through adolescence and adulthood. However, the authors of this review admit certain limitations, namely the reclassification of the disorder over the years and differences in longitudinal study designs that make it difficult to replicate findings. Despite these difficulties, results have consistently indicated that while the core symptoms of hyperactivity impulsivity may decrease over time, symptoms of inattention persist.

A systematic review of the evidence for the medical home, which advocates for long-term follow-up of individuals with chronic conditions, for children with special health care needs found 33 articles reporting on 30 distinct studies of which 10 were comparison-group studies. While none of the studies examined the medical home in its entirety, many had weak designs, inconsistent definitions and extent of medical home attributes, and inconsistent outcome measures, the majority of evidence indicates a positive relationship between the medical home and outcomes such as better health status, timeliness of care, family centeredness, and improved family functioning.

Therefore, primary care clinicians should recognize ADHD as a chronic condition and consider children and adolescents with ADHD as children and youth with special health care needs who require comprehensive follow-up treatment.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

A research review consisting of long-term follow-up studies indicated that overall, in adolescence, most patients (70-80%) continue to show symptoms and meet the diagnostic criteria for ADHD. Furthermore, in adulthood, many patients continue to be symptomatic (60%). Additional difficulties often develop in individuals with ADHD in adolescence or adulthood including low self-esteem, poor academic performance, and poor interpersonal skills.

In a systematic review of the evidence for the medical home for children with special health care needs, 9 studies found long-term significant improvements in health and functional status, 4 studies found long-term significant improvement in family function, and one study found positive but non-significant improvement in long-term cost. Similarly, a different systematic review found that 32 of 39 studies indicated that chronic care

model components improved at least 1 process or outcome measure for patients and 18 of 27 studies demonstrated reduced health care costs or lower use of health care services.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

While implementing the chronic care model takes time and resources, case studies have shown that it is feasible and improves primary care for patients with chronic illness. Overall, this recommendation describes the coordinated services most appropriate for managing the condition and while providing additional services might be more costly, there is preponderance of benefit over harm.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1) Woods D, Wolraich M, Pierce K, DiMarco L, Muller N, Sachdeva R. Considerations and evidence for an ADHD outcome measure. *Acad Pediatr*. 2014;14(5 Suppl):S54-60.
    a. This systematic review established the baseline for developing an outcome measure that assesses the quality of care for children with ADHD. This review synthesizes literature on the efficacy and effectiveness of guideline-recommended care.
    b. The systemic review results in 35 studies with 20 rates as good and 15 as fair. Regardless of outcome measure and treatment type, symptom reduction and improvement were relatively large with effect sizes ranging from 0.15-4.57.
    c. This study supports the conclusion that core symptoms of ADHD can be improved within 1 year which could satisfy the requirements for an outcome measure. The long time frame required for improvement along with the vast improvements seen after treatment indicate the need for sustained ADHD care throughout the lifespan. This supports the conclusions made in the systematic review that ADHD should be considered a chronic condition.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
evidence_attachment_ADHD_3_09-28.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
According to statistics provided by the Centers for Disease Control and Prevention, for children aged 4-17 years, 5 million children (9%) have ADHD, the percentage of children with parent-reported ADHD increased by 22% between 2003 and 2007, and rates of ADHD diagnosis increased an average of 3% per year from 1997 to 2006 and an average of 5.5% per year from 2003 to 2007 (1). ADHD has a multidimensional effect on an individual´s daily functioning and can culminate in significant costs attributable to greater health care needs, more frequent unintentional injury, co-occurring psychiatric conditions, and productivity losses. ADHD medications can reduce symptoms but might be associated with side effects and symptoms affecting comorbidity (2). While some core problems evident in young patients with ADHD, such as hyperactivity, generally improve by adulthood; many other symptoms of the disorder may persist into adulthood including impaired social relationships, low self-concept, drug use, and education and occupational disadvantages (3). ADHD continues to cause symptoms and dysfunction in many children who have the condition and available treatments are not usually curative (1). Longitudinal studies have found that frequently, treatments are not sustained despite the fact that long-term outcomes for children with ADHD indicate that they are at greater risk of significant problems if treatment is discontinued (4). ADHD follow-up care and treatment adherence can be enhanced by improving the relationship between parents and health care providers so parents feel both involved and knoledgeable about their child´s health condition and treatment regimen. The medical home and the chronic care model both emphasize patient and family involvement in care and as a result, treating ADHD as a chronic care condition within a medical home is Guideline recommended care. (3). In November, 2011, the American Academy of Pediatrics (AAP) published a new evidence based guideline for ADHD diagnosis, follow-up, and treatment based on extensive review of the existing evidence. One recommendation with a strong level of evidence encouraged primary care clinicians to recognize ADHD as a chronic condition, including managing patient care through follow-up appointments. management of children and youth with special health care needs should follow the principles of the chronic care model and the medical home (4, 5).

There is evidence that ADHD treatment can improve the likelihood of a positive outcome and reduce the negative consequences of ADHD in the short term; however, residual benefits of pharmacological treatment may subside when medication is discontinued (6). Therefore, given that ADHD symptoms may manifest for as long as 8 years after diagnosis and that ADHD treatment has been shown to work in the short term altough it may require many modifications, regular ADHD follow-up care, per the 2011 AAP ADHD guideline, is to ensure that a child is adhering to a treatment plan.

1. Centers for Disease Control and Prevention. Summary health statistics for U.S. children: Naitonal health interview survey, 2009. Vital and Health Statistics Series. 2010; 10(247).

2. Centers for Disease Control and Prevention. Increasing prevalence of parent-reported attention deficit/hyperactivity disorder among children. Morbidity and Mortality Weekly Report. November 12, 2010/59(44); 1439-1443.

3. Ingram S, Hechtman L, Morgenstern G. Outcome issues in ADHD: adolescent and adult long-term outcom. Mental Retardation and Developmental Disabilities Research Reviews. 1999;5:243-250.

4. Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. Pediatrics. 2011; 128(5):1-16.

5. Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management. ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents: process of care supplemental appendix. Pediatrics. 2011; SI1-SI21.

6. Barkley R, Fisher M, Edelbrock C, Smallish L. The adolescent outcome of hyperactive children diagnosed by research criteria: 1. an 8-year prospective follow-up study. J Am Acad Child Adolesc Psychiatry. 1990; 29(4):546-557.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
When testing the results in the Truven MarketScan Database, approximately 63% of Medicaid enrollees and 49% of Commercial enrollees who were in the denominator met the numerator criteria (met the measure overall). Unfortunately, plan-level information was unavailable so we could not test the performance of this measure across health care plans or providers.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
n/a

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The ADHD Chronic Care Follow-up measure was tested in multiple racial/ethnic groups using the Medicaid data in the Truven MarketScan database. There were sufficient numbers of cases to assess disparities according to the Affordable Care Act Classification for the following race/ethnicity categories: White, Black, Hispanic, Other (American Indian or Alaska Native, Asian or Pacific Islander; Missing/Unknown; Native Hawaiian or Other Pacific Islander; Other) by plan or geography. Results of these analyses for children diagnosed with ADHD, in the Medicaid covered population, indicate that Black children are slightly more likely to recieve Chronic Care Follow-up than non-Hispanic White children (65.3% vs 63.6%), but Hispanic (53.2%) and other minority children (57.4%) are less likely to receive appropriate Chronic Care Follow-up than non-Hispanic White children.

Although Black children are more likely to be in the numerator than non-Hispanic While children, they are less likely to be in the measure denominator (9% vs 14%) and the other groups are also less likely (4% for Hispanics and 11% for other minorities).

All differences are statistically significant and represent disparities in care. However, bias may be introduced from the artificial loss of some potentially eligible children diagnosed with ADHD in the construction of the measure.

The measure was tested in both commercial and Medicaid populations as a proxy for Socioeconomic Status (SES). Since there were no other SES variables available in the MarketScan data, if SES information is availalbe and there are sufficient number, then there are no additional issues with measures SES-based disparities.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
n/a

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

1c.2. If Other:

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
n/a

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Behavioral Health : Attention Deficit Hyperactivity Disorder (ADHD)

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Health and Functional Status

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
tbd

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ICD_9_-_10_Codes_-_ADHD_Chronic_Care_Follow-up.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who attended at least one ADHD follow-up care visit within the calendar year.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The time period for the data is one year.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target*

process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.

For patients that meet the denominator criteria, determine whether they had an evaluation and management visit with a primary ADHD diagnostic code during the measurement year.

ICD-9 ADHD Diagnosis Codes: 314.00, 314.01
ICD-10 ADHD Diagnosis Code*: F90-

The following CPT codes can be used to identify outpatient follow-up visits:
90804-90815
96150-96154
98960-98962
99078
99201-99205
99211-99215
99217-99220
99241-99245
99341-99345
99347-99350
99383
99384
99393
99394
99401-99404
99411
99412
99510

The following CPT codes can be used to identify outpatient follow-up visits with places of service (POS)** 03, 05, 07, 09, 11, 12, 13, 14, 15, 20, 22, 33, 49, 50, 52, 53, 71, 72:
90801
90802
90816-90819
90821-90824
90826-90829
90845
90847
90849
90853
90857
90862
90875
90876

The following CPT codes can be used to identify outpatient follow-up visits with place of service (POS) 52, 53:
99221-99223
99231-99233
99238
99239
99251-99255

*In ICD-10 codes, the (-) should be treated like the (xx) in ICD-9 codes. We can no longer use "x" because that letter is used in actual ICD-10 codes.

**POS codes specify the entity where service(s) were rendered, i.e. a Federally Qualified Health Center, Community Mental Health

Center, etc.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 4 through 18 years with a diagnosis of ADHD.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
In order to be eligible to meet this measure, a patient must have continuous enrollment during all days of the measurement year and 1 or more days in the prior year (identification year).

ICD-9 Diagnosis Codes that can be used to identify cases of ADHD include:
314.01 Combined Type
314.01 Predominantly Hyperactive-Impulsive Type
314.00 Predominantly Inattentive Type
314.9 Attention-Deficit/Hyperactivity Disorder NOS

ICD-10 Diagnosis Code*: F90-

*In ICD-10 codes, the (-) should be treated like the (xx) in ICD-9 codes. We can no longer use "x" because that letter is used in actual ICD-10 codes.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Documentation of medical reason(s) for not providing follow-up care (e.g., patient with multiple psychiatric conditions referred to other provider). Please see code list in section S.11.

Documentation of system reason(s) for not providing follow-up care (e.g., patient for whom the follow-up visits were not all with the same practice).

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
ICD-9 Diagnosis Codes which can be used to identify cases to exclude include the following:
Autism: 299.xx
Substance Abuse: 303.xx, 304.xx, 305.xx
Anorexia: 307.1
Mood Disorders: 296.00-296.06, 296.10-96.16, 296.22, 296.24, 296.32-296.34, 296.4, 296.5, 296.6, 296.7, 296.8
Anxiety: 300.01, 300.10-300.19, 300.21, 300.22, 300.5-300.9

Corresponding ICD-10-CM*:
Autism: F84-
Substance Abuse: F10.1-, F10.2-, F11.1-, F11.2-, F12.1-, F12.9-, F13.1-, F13.2-, F14.1-, F14.2-, F15.1-, F15.2-, F16.1-, F16.2-, F17,2-, F18.1-, F19.1-, F19.2-
Anorexia: F50.0
Mood Disorders: F30.10-F30.13, F30.2-F30.4, F30.8, F31.10-F31.13, F31.2, F31.73, F31.74, F31.30-F31.32, F31.4, F31.5, F31.60-F31.64, F31.75, F31.76, F31.77, F31.78, F31.81, F31.9, F32.1, F32.3, F32.8, F33.1-F33.3
Anxiety: F40.00, F40.01, F40.02, F41.0, F44.0, F44.1, F44.4, F44.6, F44.81, F44.89, F44.9, F48.8, F48.9, F68.11, F68.8

Presence of any one of these codes is considered an exclusion. The patient does not need to have multiple comorbidities.

*In ICD-10 codes, the (-) should be treated like the (xx) in ICD-9 codes. We can no longer use "x" because that letter is used in actual ICD-10 codes.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables,*

definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)
It is possible to stratify the measure results, but it is not required. This measure may be stratified by race/ethnicity, age group (4-5, 6-12, 13-18) and payer type (Medicaid vs Commercial).

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
n/a

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Using outpatient administrative claims data, calculate the measure using the following steps:

1. Identify children with complete coverage in the measurement year and 1 or more days of coverage in the prior year (identification year).

2. Identify all patients age 4 through 18 at the time of the primary or secondary ADHD diagnosis (ICD-9: 314.0; ICD-10: F90-) at a visit (2a, below) during the identification year. Remove patients meeting the exclusion criteria (2b, below). This remaining group of patients is the measure denominator.

2.a. CPT codes for visits:
2.a.i. The following CPT codes can be used to identify outpatient follow-up visits: 90804-90815, 96150-96154, 98960-98962, 99078, 99201-99205, 99211-99215, 99217-99220, 99241-99245, 99341-99345, 99347-99350, 99383, 99384, 99393, 99394, 99401-99404, 99411, 99412, 99510

2.a.ii. The following CPT codes can be used to identify outpatient follow-up visits with POS 03, 05, 07, 09, 11, 12, 13, 14, 15, 20, 22, 33, 49, 50, 52, 71, 72: 90801, 90802, 90816-90819, 90821-90824, 90826-90829, 90845, 90847, 90849, 90853, 90857, 90862, 90875, 90876

2.a.iii. The following CPT codes can be used to identify outpatient follow-up visits w/POS 52, 53: 99221-99223, 99231-99233, 99238, 99239, 99251-99255

2.b The following ICD Diagnosis codes can be used to identify cases to exclude include the following:
2.b.1. ICD-9 codes

Autism: 299.xx
Substance Abuse: 303.xx, 304.xx, 305.xx
Anorexia: 307.1
Mood Disorders: 296.00-296.06, 296.10-296.16, 296.22, 296.24, 296.32-296.34, 296.4, 296.5, 296.6, 296.7, 296.8
Anxiety: 300.01, 300.10-300.19, 300.21, 300.22, 300.5-300.9

2.b.2 ICD-10 codes*
Autism: F84-
Substance Abuse: F10.1-, F10.2-, F11.1-, F11.2-, F12.1-, F12.9-, F13.1-, F13.2-, F14.1-, F14.2-, F15.1-, F15.2-, F16.1-, F16.2-, F17,2-, F18.1-, F19.1-, F19.2-
Anorexia: F50.0
Mood Disorders: F30.10-F30.13, F30.2-F30.4, F30.8, F31.10-F31.13, F31.2, F31.73, F31.74, F31.30-F31.32, F31.4, F31.5, F31.60-F31.64, F31.75, F31.76, F31.77, F31.78, F31.81, F31.9, F32.1, F32.3, F32.8, F33.1-F33.3
Anxiety: F40.00, F40.01, F40.02, F41.0, F44.0, F44.1, F44.4, F44.6, F44.81, F44.89, F44.9, F48.8, F48.9, F68.11, F68.8

3. For these patients, determine the number of children with an evaluation and management visit with a primary ADHD diagnostic code during the measurement year (this group meets the numerator criteria).

*In ICD-10 codes, the (-) should be treated like the (xx) in ICD-9 codes. We can no longer use "x" because that letter is used in actual ICD-10 codes.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
n/a

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
n/a

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
If data required to compute the denominator are missing, the patient is excluded from the measure entirely. As denominator elements include age and a diagnosis of ADHD, we do not expect that many patients who should have been included in the measure will be excluded due to missing data elements. If data required to compute the numerator are missing, the patient is included in the denominator but not the numerator. In this case, the care represented in the chart has not met the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data source is an administrative claims database.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
n/a

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
nqf_testing_attachment_ADHD_3_10-12.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: ADHD Chronic Care Follow-up
Date of Submission: 9/30/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

The dataset used for this testing was the Truven MarketScan database which includes administrative claims from both Medicaid/CHIP and Commercial claims. The MarketScan database is the largest of its kind in the industry with data on more than 200 million unique patients since 1995. The database contains fully integrated patient-level data including inpatient, outpatient, drug, lab, health and productivity management, health risk assessment, dental, and benefit design from commercial, Medicare supplemental, and Medicaid populations that reflect real-world treatment patterns and costs. The database has rigorous validation methods to ensure that claims and enrollment data are complete, accurate, and reliable.

1.3. What are the dates of the data used in testing? 2010-2011

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: Population: National	☒ other: Population: National

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The Truven MarketScan database is a national database of administrative claims. There were a total of 17,753,011 enrollees who were in the designated age range; 8,653,053 who had complete coverage in 2011 and 1 or more days in 2010; and 569,228 who were diagnosed with ADHD and met the denominator criteria of this measure. Of the population that met the denominator criteria, 166,471 were in Medicaid claims and 402,757 were in Commercial claims.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

The analysis included 569,228 enrollees who had complete coverage in 2011 and 1 or more days in 2010 and who were diagnosed with ADHD and did not meet any exclusion criteria.

Of the group that met the denominator criteria, and therefore were included in the measure, 37,398 were aged 4-5 years, 347,358 were aged 6-12 years, and 184,472 were aged 13-18 years. Race/ethnicity data was only available for Medicaid claims but of the Medicaid population, 693,210 patients were Non-Hispanic White, 498,472 were Black, 143,151 were Hispanic, and 142,953 listed other race/ethnicity groups. This is representative of the US population diagnosed with ADHD.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

While reliability and measure performance analysis were conducted in the Truven MarketScan database, validity testing (face validity) was performed through the ADHD Expert Work Group. The measure was assessed for content validity by looking for agreement among subject matter experts, specifically the panel of stakeholder representatives participating in the ADHD Expert Work Group during the measure development process. The Expert Work Group consisted of 25 members including pediatricians, pediatric neurologists, social workers, school psychologists, family physicians, school-based learning disability specialists, teachers, parents, consumer representatives, child and adolescent psychologists, occupational therapists, clinical psychologists, pediatric nurses, and measure methodologists.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Race/ethnicity was available in claims data for the Medicaid population and the measure was tested and analyzed looking across different race/ethnicity groups. Socioeconomic status was unavailable; however, Medicaid and commercial claims could be used as a proxy for socioeconomic status and data was widely available in the database. Geographic identifiers and language proficiency were unavailable.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

For critical data elements testing, each measure component (numerator, denominator, and exclusions) were tested through implementation. Results were reviewed and reliability was assessed based on comparison with the total ADHD population in Medicaid/CHIP and commercial insurance, respectively.. Results of the analyses of the measure led to substantial changes to the initially proposed specifications. The components were iteratively tested until results indicated that the measure specifications were capturing the correct population.

For performance measure score, the measure was implemented in the Truven MarketScan database and performance was compared to the performance of the Initial Core ADHD Follow-up Measure.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Administrative claims are claims submitted to Medicaid and private insurers for care delivered to patients and are generally assumed to be reliable within a margin of error given the large number of patients and patient claims to be analyzed.

In the critical data elements testing of the Medicaid population, 22.52% of the denominator population had a valid specific psychiatric E&M visit with an ADHD diagnosis code in the measurement year (2011). Similarly, 13.43% of the denominator population had a valid other psychiatric E&M visit with ADHD diagnosis code in the measurement year (2011), 46.83% of the denominator had a valid non-psychiatric E&M visit with an ADHD diagnosis code in the measurement year (2011). In the critical data elements testing of the Commercial claims, 13.80% of the denominator population had a valid specific psychiatric E&M visit with ADHD diagnosis code in the measurement year, 6.48% of the denominator population had a valid other psychiatric E&M visit with an ADHD diagnosis code in the measurement year, and 38.62% had a valid non-psychiatric E&M visit with an ADHD diagnosis code in the measurement year.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The results of ADHD Chronic care Follow-up demonstrated that when implemented, the measure results were in the range of what would be expected from current gaps in ADHD care research and expert opinion.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Statement of intent for selection of ICD-10 codes: Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

The measure was assessed for face validity by having the topic, language, specifications and results reviewed by an Expert Workgroup which included a broad range of stakeholders. The PMCoE ADHD Measures Leadership Team convened a twenty-four member multi-stakeholder advisory panel with representation from a wide range of stakeholders, including consumers, pediatricians, family physicians, adolescent medicine physicians, psychiatrists, teachers, state Medicaid agencies and researchers. Input from these stakeholders was instrumental in ensuring this measure addressed the needs of children diagnosed with ADHD and responded to the needs of children in Medicaid. Throughout the measure development process, we presented the ADHD measures to this Expert Workgroup technical panel and the person within Illinois Health and Family Services (and the Illinois State Medicaid agency) who oversees quality measure use and application and solicited feedback on importance, relevance, understandability, and usability.

Face validity testing was performed on administrative claims for payment. Administrative claims are claims submitted to Medicaid and private insurers for care delivered to patients and are generally assumed to be reliable within a margin of error given the large number of patients and patient claims to be analyzed.

Face validity comments regarding the ability to identify statistically significant and meaningful differences in performance (NQF Algorithm 3. Guidance for Evaluating Validity – box 2b5) included:

"While I am an educator and do not have responsibilities of care as that of the primary care clinician, I do feel strongly that the ADHD treatment follow-up care should be the focus with the understanding that ADHD is a chronic condition. If treatment (which may change over time) is not sustained, negative outcomes are certainly more likely to occur."

"The measure and documentation seem appropriate—I would say that the measure would be a minimum standard and that optimal care would require more than one visit a year."

Face validity comments regarding multiple sets of specifications (NQF Algorithm 3. Guidance for Evaluating Validity – box 2b6) included:
"Overall, the measure and specifications are good!"

"…..everything looks good to me. Congratulations."

Face validity comments regarding computer performance measure score (NQF Algorithm 3. Guidance for Evaluating Validity – boxes 4 & 5) included:
"I have reviewed these and have no edits or suggestions for revision. As a non-physician, I am somewhat limited in commenting on some aspects of these materials (e.g., visit codes), but overall the measure and data collection procedures make perfect sense to me. And if I am interpreting the data correctly, it is sobering to note that a large percentage of children and adolescents with ADHD are not receiving follow-up care. This is not surprising, but is sobering nonetheless."

"I agree with the intention to use a chronic care model in order to enhance treatment adherence and improve the quality of follow up care. It is sobering to see how low the percentages are for any 1 follow up visit in 1 year (63% Medicaid, 49% Commercial)."

In addition, some minor edits to the measure language were suggested from an entity that would attempt to program the measure as written. These were applied in the finalized version of the measure.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Aside from just the numbers of children in each component, we also conducted complementary analyses to measure the validity. We implemented the existing CHIPRA Initial Core Measure of ADHD and compared it to our proposed version. Since the measure requires a follow-up visit with an ADHD diagnosis code, we examined the most frequent diagnosis codes for visits without an ADHD code to determine if there was a pattern for these non-ADHD related visits. We also examined the likelihood that children met the follow-up requirement with a psychiatric E&M visit vs. a non-psychiatric visit.

According to the NQF Algorithm 3. Guidance for Evaluating Reliability, the following questions were posed and answered through the Public Comment validity testing period:

1. How strong is the scientific evidence supporting the validity of this measure as a quality measure(box 2b5)?

   The scientific evidence is determined to be strong, Level B, according to the 2011 AAP ADHD Guideline which recommends pediatric ADHD patients be considered as having a Chronic Condition and treated using the Chronic Care Model, requiring regular follow-up visits to ensure presence of condition and monitor treatment plans.

2. Are all inviduals in the denominator equally eligible for inclusion in the numerator (box 2b3, 2b7)?

   Yes, except for those in the exclusion categories which include: Mood Disorders (296.xx), Autism (299.xx), Anxiety (300.xx), Substance Abuse (303.xx, 304.xx, 305.xx), and Anorexia (307.1).

3. How well do the measure specifications capture the event that is the subject of the measure (box 2b6)?

   Results of testing of the new specifications of the enhanced ADHD Follow-up measure to assess Chronic Care Follow-up were strong (Attachment 6.8). High level results include that 63% of Medicaid enrollees and 49% of Commercial enrollees who had sufficient coverage (complete coverage in the measure year and at least 1 day of coverage in the prior year) and were diagnosed with ADHD in 2010 had any valid E&M visit with ADHD diagnosis code in the measurement year.

   A question arose regarding the inclusion of psychiatric codes in the E&M code list specified in the measure. A list of psych codes was provided to ensure that these were included: 90804-90807; 90862-90863 (medication management).

   Data analyses results using the Truven MarketScan database determined that the measure has strong face validity for the measurement of ADHD ChronicCare Follow-up, both for the Medicaid/CHIP population and in the Comercial insurance population.

4. Does the measure provide for fair comparisons of the performance of providers, facilities, health plans, or geographic areas (box 2b5) ?

As specified, the measure is simple to construct and should provide a fair comparison of health plans and geographic areas.

5. Does the measure allow for adjustment of the measure excluding patients with rare performance-related characteristics when appropriate (box 2b4)?

   The specified exclusion criteria already takes this issue into account and additional criteria should not be necessary.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
Overall, the comments received during Public Comment support and reaffirm the need to tread ADHD as a chronic condition.

_____

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

While measure exclusions were not tested, they were determined through the ADHD Expert Work Group and Leadership Team based on evidence-based practice.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

N/A

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

N/A

_____

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Ethnicity **risk categories**
☐ **Other,**

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

N/A

2b4.4a. What were the statistical results of the analyses used to select risk factors?

N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

The ADHD Chronic Care Follow-up measure was tested in multiple racial/ethnic groups using the Medicaid data in the Truven MarketScan database. There were sufficient numbers of cases to assess disparities according to the Affordable Care Act Classification for the following race/ethnicity categories: White, Black, Hispanic, Other (American Indian or Alaska Native Asian or Pacific Islander; Missing/Unknown; Native Hawaiian or Other Pacific Islander; Other) by plan or geography. Results of these analyses for children diagnosed with ADHD, in the Medicaid covered population, indicate that Black children are slightly more likely to receive Chronic Care Follow-up than non-Hispanic White children (65.3% vs. 63.6%), but Hispanic (53.2%) and other minority children (57.4%) are less likely to receive appropriate Chronic Care Follow-up than non-Hispanic White children.

Although Black children are more likely to be in the numerator than non-Hispanic White children, they are less likely to be in the measure denominator (9% vs. 14%) and the other groups are also less likely (4% for Hispanics and 11% for other minorities).

All differences are statistically significant and represent disparities in care. However, bias may be introduced from the artificial loss of some potentially eligible children diagnosed with ADHD in the construction of the measure.

% Children (Age 4-18) the ADHD Chronic Measure Numerator by Race

	Non-Hispanic White	Black	Hispanic	Other
N	693,210	498,472	143,151	142,953
%	63.55%	65.30%***	53.22%***	57.36%***

We use chi-square tests to measure disparities * p<0.05 ** p<0.01 *** p<0.001

% Children (Age 4-18) the ADHD Chronic Measure Denominator by Race

	Non-Hispanic White	Black	Hispanic	Other
N	693,210	498,472	143,151	142,953
%	14.38%	9.14%***	3.77%***	11.06%***

We use chi-square tests to measure disparities * p<0.05 ** p<0.01 *** p<0.001

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

In order to determine if results were statistically significant, the Medicaid claims data was divided into race/ethnicity groups and then chi-square tests were used to measure disparities in performance of the measure. A p-value of less than 0.05 was used to determine significance.

The Truven MarketScan database was divided in such a way that we could not assess the performance of the measure across all children with continuous coverage and as a result, we tested the measure in the Medicaid population and Commercial population separately and report two performance scores below.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Approximately 63% of Medicaid enrollees and 49% of Commercial enrollees who were in the denominator population met the measure.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

Medicaid enrollees were more likely to meet the measure than Commercial enrollees and a larger percentage of Medicaid enrollees had sufficient coverage than Commercial enrollees.

The results represent the populations of both Medicaid and Commercial enrollees and therefore represent actual and meaningful differences and can distinguish good from poor ADHD care.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than***

one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

In order to meet the denominator criteria for the measure, all components of the denominator must be present in the claims data. Missing data will be most likely due to non-continuous coverage during the measurement year.

In order to meet the numerator criteria for the measure, care must be provided throughout the measurement year. If data is missing, it is assumed that care was not provided and the enrollee does not meet the numerator criteria.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

There are 17,753,011 enrollees in both Commercial and Medicaid claims who are between the ages of 4 and 18 and were enrolled during 2010-2011. However, after imposing a requirement of continuous enrollment through 2011 with 1 of more days of coverage in 2010, 9,099,955 enrollees fall out.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

There are challenges and limitations with the use of administrative claims-based data for the measurement of

care quality. The measure requires full enrollment and a one year look back period to ensure a fair assessment of clinicians' performance on this measure. A number of patients fall out of the measure if they do not meet the criteria for continuous enrollment. While a large number of enrollees do drop out, this measure is a measure of chronic follow-up care which requires, at minimum, one year of follow-up. The results of the testing, however, gave us confidence as this measure performed better than the ADHD CHIPRA core set measure, were similar to the results described in the literature, and were in an expected range for the participants in the Expert Work Group.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Through many iterations of testing, we have learned the following:

-CPT code 96110 may be used to document the use of a validated, standardized screening tool to assess ADHD symptoms in follow-up visits as an enhancement to the Initial Core ADHD Follow-up measure. However, usage of CPT code 96110 was not found to be reliable or valid as a method for assessment of standardized tool use for establishing the ADHD diagnoses as only certain states use this code for reimbursement and many physicians do not use this code when reimbursed. Through this, we concluded that CPT code 96110 is underutilized in administrative claims for hte use of a standardized tool and should not be used in the specification of an ADHD measure for accountability at this time.

-To assess the parameters of follow-up visits, date of initial diagnosis must be identified. A longer diagnosis period was proposed and tested in administrative claims by extending the denominator look-back period to one year to identify any prior diagnoses. This resulted in an optimal requirement of 16 months of eligibility and approximately 60% of the total population was lost, leading to limitations in the reliability and validity of the measure. The reliability and performance of different look-back periods was tested by assessing the loss percentage of enrollees with extended continuous coverage requirement incrementally and results showed relatively stable denominator numbers with continuous reduction over time. A 4 month clean period was not adequate to define an initial diagnosis. We concluded that it was difficult to reliably determine the initial diagnosis of ADHD in order to assess appropriate visit timeframes following diagnosis as many children with ADHD and potentially eliglble for inclusion in the measure were excluded

with an extended look-back period. As a result, the measure specifications were changed to assess continuous chronic care follow-up for ADHD, using a one year look-back period, which was conceptualy consistent with the 2011 AAP ADHD guideline for recommended care, and significantly improves the reproducibility of this measure in administrative claims-based data systems.

-When an evaluation & management and an ADHD diagnosis code in the measurement year was considered, a larger percentage of enrollees who were in the denominator met the measure than for the previous iteration of the specifications. The performance was consistent with performance in the literature and expert opinion.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
n/a

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Payment Program Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This measure is not yet endorsed.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

This measure is being submitted for endorsement for use by public and private health plans, Medicaid, and CHIPRA to assess the quality of chronic care follow-up for children diagnosed wtih ADHD for public reporting and quality improvement.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
n/a

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended negative consequences to individuals or populations identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0108 : Follow-Up Care for Children Prescribed ADHD Medication (ADD)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

The currently endorsed ADHD follow-up measure is limited to children 6-12 years of age with newly prescribed ADHD medications, while the PMCoE proposed measure targets children 4-18 years old with a primary or secondary diagnosis of ADHD. Therefore, while both measures are ADHD follow-up measures, the ADHD Chronic Care Measure has a broader age range for the target population and includes all children diagnosed with ADHD, regardless of whether they have been prescribed medication. The ADHD Chronic Care Measure measures performance of care recommended in the AAP ADHD guideline (B level evidence). The difference in age range is due to the fact that the current measure was developed and endorsed prior to the release of the new 2011 AAP ADHD guideline which adjusts the age groups in focus, and has not been updated since that time to match the recommendations. This enhances the current ADHD Follow-up measure as it includes children 4-5 year of age who should be receiving Behavior Treatment (as first-line treatment) and would fall out of the currently endorsed measure. Furthermore, the currently endorsed measure as specified requires a DEA number for the clnician visit, which is associated with an individual and not a clinic, and as Federally Qualified Health Centes (FQHC) cannot bill using a DEA number, children seeking care at FQHCs who might otherwise be eligible may fall out of the measure as well. As the proposed measure is specified using administrative claims, its addition should not increase data collection burden.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American Academy of Pediatrics
Co.2 **Point of Contact:** Lisa, Krams, lkrams@aap.org, 847-434-4000-7663
Co.3 **Measure Developer if different from Measure Steward:** AHRQ-CMS CHIPRA Pediatric Measurement Center of Excellence (PMCoE)
Co.4 **Point of Contact:** Ramesh, Sachdeva, MD, PhD, JD, FAAP, rsachdeva@aap.org, 847-434-4000-7110

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
This measure was developed under the leadership of the Northwestern University Feinberg School of Medicine (Site PI: Donna Woods), in their capacity as a member of the PMCoE Consortium (PMCoE PI: Ramesh Sachdeva).

AHRQ-CMS CHIPRA Pediatric Measurement Center of Excellence (PMCoE) Consortium Members:
Medical College of Wisconsin/National Outcomes Center (MCW); American Medical Association-Convened Physician Consortium for Performance Improvement (AMA-PCPI); American Board of Medical Specialties (ABMS); American Board of Pediatrics (ABP); Northwestern University Feinberg School of Medicine (Northwestern); American Academy of Pediatrics (AAP); Thomson-Reuters; TMIT Consulting, LLC; Chicago Pediatric Quality and Safety Consortium

PMCoE Staff:
Ramesh C Sachdeva, PMCoE PI, MCW

Lisa Ciesielczyk, Program Manager, MCW
V. Fan Tait, Site PI, AAP
Keri Thiessen, Project Staff, AAP
Donna Woods, Site PI, Northwestern
Nicole Muller, Project Staff, Northwestern
Lindsday DiMarco, Project Staff, Northwestern
Jin-Shei Lei, Project Staff, Northwestern
Ray Kang, Project Staff, Northwestern
Susan Magasi, Project Staff, Northwestern
Sara Alafogianis, Project Staff, AMA-PCPI
Mark Antman, Project Staff, AMA-PCPI
Amaris Crawford, Project Staff, AMA-PCPI
Kendra Hanley, Project Staff, AMA-PCPI
Molly Siegal, Project Staff, AMA-PCPI
Greg Wozniak, Project Staff, AMA-PCPI
Emily Ehrlich, Project Staff, Truven Analytics

Expert Work Group Members:
Ted Abernathy, Pediatrician, Private Practice of Pediatrics and Adolescent Medicine
Betsy Brooks, Pediatrician, Holyoke Pediatric Associates
Lawrence Brown, Pediatric Neurologist, Children´s Hospital of Philadelphia
Mirean Coleman, Social Worker, National Association of Social Workers
Stephen Downs, Pediatrician, Children´s Health Services Research
George DuPaul, School Psychologist, Lehigh University
Mirian Earls, Developmental-Behavioral Pediatrician, Guildord Child Health
Jeff Epstein, Clinical Psychologist, Cincinnati Children´s Hospital Medical Center
Theodore G Ganiats, Family Physician, University of California San Diego
Jane Hannah, School-based Learning Disability Specialist, Curry Ingram Academy
Romana Hasnain-Wynia, Healthcare Equity Expert, Northwestern University Institute for Healthcare Studies
Steven Kairys, Pediatrician, Jersey Shore Medical Center
Beth Kaplanek, Parent, Children & Adults w/Attention Deficit Disorders (CHADD)
M. Ammar Katerji, Pediatric Neurologist, Advocate Hope Children´s Hospital
Shelly Lane, Occupational Therapist, Virginia Commonwealth University
Nancy Marek, Pediatric Nurse, Advocate Hope Children´s Hospital
Paul Miles, Maintenance of Certification Expert, American Board of Pediatrics
Patrice Mozee-Russell, Teacher, Children & Adults w/Attention Deficit Disorders (CHADD)
Karen Pierce, Child and Adolescent Psychiatrist, Children´s Memorial Hospital/Northwestern University
Sandra Rief, School-based Learning Disability Specialist, Children & Adults w/Attention Deficit Disorders (CHADD)
Clarke Ross, Parent, American Association on Health and Disability
Adrian Sandler, Developmental-Behavioral Pediatrician, Mission Children´s Hospital
Marcia Slomowitz, Child and Adolescent Psychiatrist, Northwestern Memorial Hospital
Laurel Stine, Consumer Representative, Bazelon Center for Mental Health Law
Mark Wolraich, Developmental-Behavioral Pediatrician, University of Oklahoma Child Study Center

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

ICD-9 and ICD-10 Codes

ICD-9 Diagnosis Code	Description ICD-9	ICD-10-CM	Description ICD-10
314.01	Predominantly Hyperactive-Impulse Type	F90.1	Attention-deficit hyperactivity disorder, predominantly hyperactive type
314.01	Combined Type	F90.2	Attention-deficit hyperactivity disorder, combined type
314.00	Predominantly Inattentive Type	F90.0	Attention-deficit hyperactivity disorder, predominantly inattentive type
314.9	Attention-Deficit/Hyperactivity Disorder NOS	F90.9	Attention-deficit hyperactivity disorder, unspecified type
299.00	Autistic disorder, current or active state	F84.0	Autistic Disorder
299.01	Autistic disorder, residual state	F84.0	Autistic Disorder
303.00-303.03	Acute alcoholic intoxication in alcoholism (unspecified or continuous or episodic or in remission)	F10.229	Alcohol dependence with intoxication, unspecified
303.90-303.92	Other and unspecified alcohol dependence (unspecified or continuous or episodic)	F10.20	Alcohol dependence, uncomplicated
303.93	Other and unspecified alcohol dependence, in remission	F10.21	Alcohol dependence, in remission
304.00-304.02	Opioid type dependence (unspecified or continuous or episodic)	F11.20	Opioid dependence, uncomplicated
304.03	Opioid type dependence, in remission	F11.21	Opioid dependence, in remission
304.10-304.12	Sedative, hypnotic or anxiolytic dependence (unspecified or continuous or episodic)	F13.20	Sedative, hypnotic, or anxiolytic dependence, uncomplicated
304.13	Sedative, hypnotic or anxiolytic dependence, in remission	F13.21	Sedative, hypnotic, or anxiolytic dependence, in remission
304.20-304.22	Cocaine dependence (unspecified or continuous or episodic)	F14.20	Cocaine dependence, uncomplicated
304.23	Cocaine dependence, in remission	F14.21	Cocaine dependence, in remission
304.30-304.32	Cannabis dependence (unspecified or continuous or episodic)	F12.20	Cannabis dependence, uncomplicated
304.33	Cannabis dependence, in remission	F12.21	Cannabis dependence, in remission
304.40-304.42	Amphetamine and other psychostimulant dependence (unspecified or continuous or episodic)	F15.20	Other stimulant dependence, uncomplicated
304.43	Amphetamine and other psychostimulant dependence, in remission	F15.21	Other stimulant dependence, in remission
304.50-304.52	Hallucinogen dependence (unspecified or continuous or episodic)	F16.20	Hallucinogen dependence, uncomplicated
304.53	Hallucinogen dependence, in remission	F16.21	Hallucinogen dependence, in remission
304.60-304.62	Other specified drug dependence (unspecified or continuous or episodic)	F19.20	Other psychoactive substance dependence, uncomplicated
304.63	Other specified drug dependence, in remission	F19.21	Other psychoactive substance dependence, in remission
304.70-304.72	Combinations of opioid typ drug with any other drug dependence (unspecified or continuous or episodic)	F19.20	Other psychoactive substance dependence, uncomplicated

304.73	Combinations of opioid typ drug with any other drug dependence, in remission	F19.21	Other psychoactive substance dependence, in remission
304.80-304.82	Combinations of drug dependence excluding opioid type drug (unspecified or continuous or episodic)	F19.20	Other psychoactive substance dependence, uncomplicated
304.83	Combinations of drug dependence excluding opioid type drug, in remission	F19.21	Other psychoactive substance dependence, in remission
304.90-304.92	Unspecified drug dependence (unspecified or continuous or episodic)	F19.20	Other psychoactive substance dependence, uncomplicated
304.93	Unspecified drug dependence, in remission	F19.21	Other psychoactive substance dependence, in remission
305.00-305.0	Nondependent alcohol abuse (unspecified or continuous or episodic or in remission)	F10.10	Alcohol abuse, uncomplicated
305.1	Tobacco use disorder	F17.2000	Nicotine dependence, unspecified, uncomplicated
305.20-305.23	Nondependent cannabis abuse (unspecified or continuous or episodic or in remission)	F12.10 OR F12.90	Cannabis abuse, uncomplicated OR Cannabis use, unspecified, uncomplicated
305.30-305.33	Nondependent hallucinogen use (unspecified or continuous or episodic or in remission)	F16.10	Hallucinogen abuse, uncomplicated
305.40-305.43	Nondependent sedative, hypnotic or anxiolytic abuse (unspecified or continuous or episodic or in remission)	F13.10	Sedative, hypnotic or anxiolytic abuse, uncomplicated
305.50-305-53	Nondependent opioid abuse (unspecified or continuous or episodic or in remission)	F11.10	Opioid abuse, uncomplicated
305.60-305.63	Nondependent cocaine abuse (unspecified or continuous or episodic or in remission)	F14.10	Cocaine abuse, uncomplicated
305.70-305-73	Nondependent amphetamine or related acting sympathomimetic abuse (unspecified or continuous or episodic or in remission)	F15.10	Other stimulant abuse, uncomplicated
305.80-305.83	Nondependent antidepressant type abuse (unspecified or continuous or episodic or in remission)	F19.10	Other psychoactive substance abuse, uncomplicated
305.90-305.93	Nondependent other mixed or unspecified drug abuse (unspecified or continuous or episodic or in remission)	F18.10	Inhalant abuse, incomplicated
307.1	Anorexia nervosa	F50.00	Anorexia nervosa, unspecified
296.00-296.03	Bipolar I disorder, single manic episode (unspecified or mild or moderate or severe, without mention of psychotic behavior)	F30.10-F30.13	Manic episode without psychotic symptoms (unspecified or mild or moderate or severe, without psychotic symptoms)
296.04	Bipolar I disorder, single manic episode, severe, specified as with psychotic behavior	F30.2	Manic episode, severe with psychotic symptoms
296.05	Bipolar I disorder, single manic episode, in partial or unspecified remission	F30.3	Manic episode in partial remission

296.06	Bipolar I disorder, single manic episode, in full remission	F30.4	Manic episode in full remission
296.10-296.13	Manic affective disorder, recurrent episode (unspecified or mile or moderate or severe, without mention of psychotic behavior)	F30.10-F30.13	Manic episode without psychotic symptoms
296.14	Manic affective disorder, recurrent episode, severe, specified as with psychotic behavior	F30.2	Manic episode, severe with psychotic symptoms
296.15	Manic affective disorder, recurrent episode, in partial or unspecified remission	F30.3	Manic episode in partial remission
296.16	Manic affective disorder, recurrent episode, in full remission	F30.4	Manic episode in full remission
296.22	Major depressive affective disorder, single episode, moderate	F32.1	Major depressive disorder, single episode, moderate
296.24	Major depressive affective disorder, single episode, severe, specified as with psychotic behavior	F32.3	Major depressive disorder, single episode, severe with psychotic features
296.32	Major depressive affective disorder, recurrent episode, moderate	F33.1	Major depressive disorder, recurrent, moderate
296.33	Major depressive affective disorder, recurrent episode, severe, without mention of psychotic behavior	F33.2	Major depressive disorder, recurrent severe without psychotic features
296.34	Major depressive affective disorder, recurrent episode, severe, specified as with psychotic behavior	F33.3	Major depressive disorder, recurrent, severe with psychotic symptoms
296.40-296.43	Bipolar I disorder, most recent episode (or current) manic (unspecified or mild or moderate or severe, without mention of psychotic behavior)	F31.10-F31.13	Bipolar disorder, current episode manic without psychotic features (unspecified or mild or moderate or severe)
296.44	Bipolar I disorder, most recent episode (or current) manic, severe, specified as with psychotic behavior	F31.2	Bipolar disorder, current episode manic sever with psychotic features
296.45	Bipolar I disorder, most recent episode (or current) manic, in partial or unspecified remission	F31.73	Bipolar disorder, in partial remission, most recent epidose manic
296.46	Bipolar I disorder, most recent episode (or current) manic, in full remission	F31.74	Bipolar disorder, in full remission, most recent episode manic
296.50-296.52	Bipolar I disorder, most recent episode (or current) depressed (unspecified or mild or moderate)	F31.30-F31.32	Bipolar disorder, current episode depressed (mild or moderate severity, unspecified or mild or moderate)
296.53	Bipolar I disorder, most recent episode (or current) depressed, severe, without mention of psychotic behavior	F31.4	Bipolar disorder, current episode depressed, severe, without psychotic features
296.54	Bipolar I disorder, most recent episode (or current) depressed, severe, specified as with psychotic behavior)	F31.5	Bipolar disorder, current episode depressed, severe, with psychotic features
296.55	Bipolar I disorder, most recent episode (or current) depressed, in partial or unspecified remission	F31.75	Bipolar disorder, in partial remission, most recent episode depressed
296.56	Bipolar I disorder, most recent episode (or current) depressed, in full remission	F31.76	Bipolar disorder, in full remission, most recent episode depressed

296.60-296.64	Bipolar I disorder, most recent episode (or current) mixed (unspecified or mild or moderate or severe, without mention of psychotic behavior or severe, specified as with psychotic behavior)	F31.60-F31.64	Bipolar disorder, current episode mixed (unspecified or mild or moderate or severe, without psychotic features or severe, with psychotic features)
296.65	Bipolar I disorder, most recent episode (or current) mixed, in partial or unspecified remission	F31.77	Bipolar disorder, in partial remission, most recent episode mixed
296.66	Bipolar I disorder, most recent episode (or current) mixed, in full remission	F31.78	Bipolar disorder, in full remission, most recent episode mixed
296.7	Bipolar I disorder, most recent episode (or current) unspecified	F31.9	Bipolar disorder, unspecified
296.80	Bipolar disorder, unspecified	F31.9	Bipolar disorder, unspecified
296.81	Atypical manic disorder	F30.8	Other manic episodes
296.82	Atypical depressive disorder	F32.8	Other depressive episodes
296.89	Other bipolar disorders	F31.81	Bipolar II disorder
300.01	Panic disorder without agoraphobia	F41.0	Panic disorder [episodic paroxysmal anxiety] without agoraphobia
300.10	Hysteria, unspecified	F44.9	Dissociative and conversion disorder, unspecified
300.11	Conversion disorder	F44.4 OR F 44.6	Conversion disorder with motor symptom or deficit OR Converstion disorder with sensory symptom or deficit
300.12	Dissociative amnesia	F44.0	Dissociative amnesia
300.13	Dissociative fugue	F44.1	Dissociative fugue
300.14	Dissociative identify disorder	F44.81	Dissociative identity disorder
300.15	Dissociative disorder or reaction, unspecified	F44.9	Dissociative and conversion disorder, unspecified
300.16	Factitious disorder with predominantly psychological signs and symptoms	F44.89 OR F68.11	Other dissociative and conversion disorders OR Factitious disorder with predominantly psychological signs and symptoms
300.19	Other and unspecified factitious illness	F68.8	Other specified disorders of adult personality and behavior
300.21	Agoraphobia with panic disorder	F40.01	Agoraphobia with panic disorder
300.22	Agoraphobia without mention of panic attacks	F40.02	Agoraphobia without panic disorder
300.5	Neurasthenia	F48.8	Other unspecified nonpsychotic mental disorders
300.6	Depersonalization disorder	F48.1	Depersonalization-derealization syndrome
300.7	Hypochondriasis	F45.21 OR F45.22	Hypochondriasis OR Body dysmorphic disorder
300.81	Somatization disorder	F45.0	Somatization disorder
300.82	Undifferentiated somatoform disorder	F45.1 OR F45.9	Undifferentiated somatoform disorder OR Somatoform disorder, unspecified
300.89	Other somatoform disorders	F45.8 OR F48.8	Other somatoform disorders OR Other specified nonpsychotic mental disorders

300.9	Unspecified nonpsychotic mental disorder	F48.9 OR F99	Nonpsychotic mental disorder, unspecified OR Mental disorder, not otherwise specified

We used www.ICD10Data.com for the conversions

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2820
Measure Title: Pediatric Computed Tomography (CT) Radiation Dose
Measure Steward: University of California, San Francisco
Brief Description of Measure: The measure requires hospitals and output facilities that conduct Computed Tomography (CT) examinations in children to: 1. Review their CT radiation dose metrics, 2. calculate the distribution of the results, and 3.compare their results to benchmarks. This would then imply a fourth step to investigate instances where results exceed a trigger value for underlying cause, such as issues with protocol, tech, equipment, patient, etc.

It is important to review doses of radiation used for CT, as the doses are far higher than conventional radiographs (x-rays), the doses are in the same range known to be carcinogenic (Pearce, Lancet, 2012; Ozasa, Radiation Research, 2012), and the higher the doses, the greater the risk of subsequent cancer (Miglioretti, JAMA Pediatrics, 2013) Thus the goal of the measure is to provide a framework where facilities can easily assess their doses, compare them to benchmarks, and take corrective action to lower their doses if they exceed threshold values, as per specifications in benchmarks.

The measure calls for assessment of doses for the most frequently conducted CT examination types, and compare these doses to published benchmarks. The measure calls for the assessment of radiation doses within four anatomic areas (CT's of the head, chest, abdomen/pelvis and combined chest/abdomen/pelvis.) The measure provides a simple framework for how facilities can assess their dose, compare their doses to published benchmarks (Smith-Bindman, Radiology, 2015) and identify opportunities to improve if their doses are higher than the benchmarks. For example, If a hospital finds their doses are higher than published benchmarks, they can review the processes and procedures they use for performance of CT in children and take corrective action, and follow published guidelines for how to lower doses (such as "child sizing" the doses, reducing multiple phase scans, and reducing scan lengths).

Published benchmarks for radiation dose in children exist (Smith-Bindman, Radiology, 2015) and additional benchmarks are under development and will be published within the year by us. (Kumar, 2015) Other groups have also published benchmarks (Goeske) or in the process of doing so.

Our work and that of others have shown that institutional review of dose metrics as outlined in this measure results in a significant lowering of average and outlier doses. (Demb, 2015; Greenwood, RadioGraphics, 2015; Miglioretti, JAMA Pediatrics, 2013; Keegan, JACR, 2104; Wilson, ARRS, 2015).

This measure is being proposed for diagnostic CT in children, but can also be used for CT in adults, and CT used in conjunction with radiation therapy for cancer. Whenever context the doses are used, the doses should be compared with appropriate benchmarks.

A similar measure (#0739) was previously endorsed by the NQF in 2011. The NQF did not provide ongoing endorsement when the measure was up for renewal in 2015, primarily because there was no evidence that assessing doses as called for in the measure would result in an improvement in outcomes (i.e. patient dose). Since that time, there has been additional research that has shown that assessing doses using the format outlined in the measure does indeed result in lower doses, and thus we are re-submitting a similar although updated measure.

Of note, the surrogate measure we are using for outcomes is radiation dose. The true outcome of interest is the number of cancers that result from imaging. Because of the lag time between exposure to radiation and cancer development (years to decades) it is not feasible to use cancer cases as the outcome of a quality improvement effort. Thus while there is ample evidence that radiation causes cancer (sited below), and evidenced that cancer risk is proportional to dose, there are no direct data that suggest that

lowering doses lowers cancer risk. However, we have used mathematical modeling to try to understand the relationship between lowering doses and cancers and estimated that if the top quartile of doses were reduced in children (i.e. the very high doses are brought down the average doses), the number of cancer cases would be reduced by approximately 43%, the equivalent to preventing 4,350 cancer cases / year in the US among children (Miglioretti, JAMA Pediatrics 2013).

Cited in this section:

Demb J, manuscript under preparation. CT Radiation Dose Standardization Across the University of California Medical Centers Using Audits to Optimize Dose. 2015.
Following an in-person meeting regarding CT radiation dose, radiologists, technologists and medical physicists from University of California medical centers strategized how to best optimize dosing practices at their sites, which were then analyzed for effectiveness and success after implementation.

Greenwood T, Lopez-Costa R, Rhoades P, et al. CT Dose Optimization in Pediatric Radiology: A Multiyear Effort to Preserve the Benefits of Imaging While Reducing the Risks. RadioGraphics. Jan 2015;35(5):1539-1554
"This systematic approach involving education, streamlining access to magnetic resonance imaging and ultrasonography, auditing with comparison with benchmarks, applying modern CT technology, and revising CT protocols has led to a more than twofold reduction in CT radiation exposure between 2005 and 2012…" – Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014
Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Kumar K, manuscript under preparation. Radiation Dose Benchmarks in Children.
This paper will describe dose metrics among 29,000 children within age strata <1, 1-4 years, 5-9 years, 10-14 years, and 15-19 years. 2015.

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311, 2013.
Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered.

Miglioretti, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. Journal of the American College of Radiology: JACR 2014.
"Personalized audit feedback and education can change technologists´ attitudes about, and awareness of, radiation and can lower patient radiation exposure from CT imaging." – Conclusion statement from Abstract

Ozasa K, Shimizu Y, Suyama A, et al. Studies of the mortality of atomic bomb survivors, Report 14, 1950-2003: an overview of cancer and noncancer diseases. Radiation Research; 177(3):229-243. Mar 2012
Fourteenth follow-up report on the lifetime health effects from radiation on atomic bomb survivor showing that: 58% of the 86,611 LSS cohort members with DS02 dose estimates have died, 17% more cancer deaths especially among those under age 10 at exposure (58% more deaths).

Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet;380(9840):499-505. Aug 4 2012
"Use of CT scans in children to deliver cumulative doses of about 50 mGy might almost triple the risk of leukaemia and doses of about 60 mGy might triple the risk of brain cancer… although clinical benefits should outweigh the small absolute risks, radiation doses from CT scans ought to be kept as low as possible" – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology 2015:277: 134–141
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Wilson N. CT Radiation Dose Standardization Across the Five University of California Medical Centers. ARRS: Annual Toronto Meeting presentation. April 19-24, 2015

Understanding the reasons for variation in commonly performed CT procedures, and figuring out how to standardize them.
Developer Rationale: Radiologists and other physicians who perform CT in general are not aware of the doses they use, and there is tremendous variation in the doses they use even for patients seen for the same clinical indication. Even when clear standards around optimal doses exist, facilities do not routinely assess whether they use appropriate doses. For example, we conducted a 15 center randomized controlled trial of patients with suspected kidney stones seen in one of 15 U.S. emergency rooms (Smith-Bindman, NEJM, 2014),. The primary purpose of this study was to assess whether CT or ultrasound should be used as the first diagnostic test in these patients. As a secondary aim, we assessed the radiation doses of patients who received CT scans as part of this trial (Smith-Bindman, Jama IM). It is well established that patients with suspected kidney stones should undergo CT using a low dose, renal stone protocol CT, which delivers a dose of around 4 mSv or lower, as it is equally diagnostic to routine abdominal CT but uses around 1/3 the amount of radiation without any loss of diagnostic accuracy. (ACR, DIR, 2014) Nonetheless, when we assessed the doses that were actually used among the patients in our cohort, fewer than 10% of patients received low doses, the average dose was 12 mSv (three times higher) and some patients received doses as high as 75mSv (Smith-Bindman, JAMA, 2012; Smith-Bindman, Radiology, 2015). Of note, all of these patients were at high risk for stone disease, and at low risk for alternative diagnoses, and thus all should have received low dose examinations. These results closely paralleled the results of the American College of Radiology Dose Index Registry where the doses for Stone protocol CTs were assessed, and only 2% of exams used low dose. Of not, none of the participating sites in my 15 center trial were aware of their doses, and our quantification of these doses was the first step for facilities to try to optimize. If all doses were at the appropriate dose level, the doses would have been around 40% lower.

The lack of local practice assessment as highlighted by our STONE trial leads to dramatic practice variation that introduces unnecessary harm from excessive radiation dosing, and many publications have demonstrated profound variation in doses when a patient goes to different facilities to obtain a CT, or variation within institution when studies are obtained at different times of the day (ACR, DIR, 2014; Hausleiter, JAMA 2009; Keegan, JACR 2014; Miglioretti, JAMA Pediatrics, 2013; Parker, Pediatrics, 2015; Smith-Bindman, Arch Int Med, 2009; Smith-Bindman, JAMA, 2012; Smith-Bindman, JACR 2014).

In our JAMA Pediatrics paper (Miglioretti, JAMA IM 2013) using statistical modeling and observed CT doses, we modeled what would occur if the highest dose patients (those above the 75% benchmark) came down to the median dose. vThe dominant two indications for imaging in this cohort was imaging with CT for minor trauma and imaging with CT for appendicitis. Using current exposures, we would expect that due to CT exposures in children age 15 and younger in the US in 2010, 9,820 future cancers will occur. If the highest exposed individuals instead had doses at the median, 44% of these cancers would be prevented.

Furthermore, since information on radiation is reported differently across the different types of CT machines, and data are pooled in various ways, it is difficult for physicians to easily standardize their practice without a common and simple framework for doing so. Currently, physicians do not know the typical radiation doses received by their patients. This tool provides the framework for measurement – the first step towards quality improvement.

Creation of a simple standard for collection of radiation dose information would help facilities understand their current practice, would allow understanding changes in practice over time (Keegan, JACR, 2014; Greenwood, RadioGraphics, 2015) would allow comparisons to local and national standards, and would indicate to facilities whether their is a need to improve. There is currently a high level of interest in this area - facilities are being asked by their patients and governing boards to report whether they are performing CT safely - and this measure is an ideal starting point for facilities to assemble this information to answer these questions. If facilities collect dose information, it is the first step towards trying to compete on a measure of safety and to lower the doses they use.

The measure will contribute to the creation of broadly applicable expected range, and UCSF and other professional organizations will contribute to their creation. This will lead to dose awareness and inevitable improvements as it will enable physicians to consider dose as an important measure.

We compared several methods of assessing doses as outlined in this measure, including automated and manual dose assessment. While automatic approaches have obvious advantages, it is feasible to collect these data manually with minimal time(Keegan, JACR, 2014).

Cited in this section:

American College of Radiology (ACR). Dose Index Registry (DIR). 2014.
http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/DIR%20Measures.pdf
Registry designed to showcase measures for certain CT procedure types.

Greenwood T, Lopez-Costa R, Rhoades P, et al. CT Dose Optimization in Pediatric Radiology: A Multiyear Effort to Preserve the Benefits of Imaging While Reducing the Risks. RadioGraphics. Jan 2015;35(5):1539-1554
"This systematic approach involving education, streamlining access to magnetic resonance imaging and ultrasonography, auditing with comparison with benchmarks, applying modern CT technology, and revising CT protocols has led to a more than twofold reduction in CT radiation exposure between 2005 and 2012…" – Conclusion statement from Abstract

Hausleiter, J., T. Meyer, et al. Estimated radiation dose associated with cardiac CT angiography. JAMA 301(5): 500-7. 2009
"Median doses of CCTA differ significantly between study sites and CT systems. Effective strategies to reduce radiation dose are available but some strategies are not frequently used. The comparable diagnostic image quality may support an increased use of dose-saving strategies in adequately selected patients."– Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014
Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311, 2013.
Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered.

Parker M, Shah S, Hall M, et al. Computed Tomography and Shifts to Alternate Imaging Modalities in Hospitalized Children. Pediatrics. 2015-0995.
"For the 10 most common All-Patient Refined Diagnosis Related Groups (APR-DRGs) for which children received CT in 2004, a decrease in CT utilization was found in 2012. Alternative imaging modalities for 8 of the diagnoses were used." – Conclusion statement from Abstract

Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA;307:2400-9. 2012
" Within integrated health care systems, there was a large increase in the rate of advanced diagnostic imaging and associated radiation exposure between 1996 and 2010." – Conclusion statement from Abstract

Smith-Bindman R, Aubin C, Bailitz J, et al. Ultrasonography versus Computed Tomography for Suspected Nephrolithiasis. N Engl J Med (NEJM); 371:1100-1110. 2014
"Initial ultrasonography was associated with lower cumulative radiation exposure than initial CT, without significant differences in high-risk diagnoses with complications, serious adverse events, pain scores, return emergency department visits, or hospitalizations." – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology; 277: 134–141. 2015
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med;169:2078-86. 2009
" Radiation doses from commonly performed diagnostic CT examinations are higher and more variable than generally quoted, highlighting the need for greater standardization across institutions." – Conclusion statement from Abstract

Lukasiewicz A, Bhargavan-Chatfield M, Coombs L, et al. Radiation Dose Index of Renal Colic Protocol CT Studies in the United States: A Report from the American College of Radiology National Radiology Data Registry. Radiology. May 2014;271(2):445-451.
"Reduced-dose renal protocol CT is used infrequently in the United States. Mean dose index is higher than reported previously, and institutional variation is substantial." – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Griffey RT, et al. Computed Tomography Radiation Dose in Patients With Suspected Urolithiasis. JAMA internal medicine. Aug 1 2015;175(8):1413-1416.

Smith-Bindman 2015, Predictors of Computed Tomography Radiation Dose and Their Impact on Patient Care. In Press, Radiology

Numerator Statement: Radiation Dose metrics among consecutive patients, who have undergone CT of the head, chest, abdomen/pelvis, or chest/abdomen/pelvis. The metrics are 1) mean dose as measured using DLP, CTDIvol, and SSDE: within age strata. And 2) the proportion of exams with doses greater than the 75th percentile of the benchmark you are comparing with for the same anatomic area strata (Kumar, 2015; Smith-Bindman, Radiology, 2015; Goske, Radiology, 2013)

The CTDIvol and DLP are directly reported by the scanner using an "industry wide" standardized dose report (DICOM Radiation Dose Structured Report). The data should be assembled for the entire CT examination. If there are several series, the CTDIvol values should be averaged, and the DLP values should be added.

SSDE can be calculated using any dose monitoring software product, or using published multiplier coefficients which are highly valid.

These different metrics are highly correlated, but nonetheless reveal important differences regarding radiology practice and performance and are thus complimentary. However, if a practice only assesses data from a single metric, there is substantial opportunity for data-driven improvement.

CTDIvol reflects the average dose per small scan length. Modern CT scanners directly generate this.

DLP reflects the CTDIvol x scan length, and is directly generated by modern CT scanners.

SSDE is a modified measure of CTDIvol that takes into account the size of the patient scanned and is useful for scaling dose to patient size. Several current radiation tracking software tools directly report SSDE.

Cited in this section

Goske MJ, Strauss KJ, Coombs LP, et al. Diagnostic reference ranges for pediatric abdominal CT. Radiology. Jul 2013;268(1):208-218.
"Calculation of reference doses as a function of BW (body weight) for an individual practice provides a tool to help develop site-specific CT protocols that help manage pediatric patient radiation doses." – Conclusion statement from Abstract

Kumar K, manuscript under preparation. Radiation Dose Benchmarks in Children.
This paper will describe dose metrics among 29,000 children within age strata <1, 1-4 years, 5-9 years, 10-14 years, and 15-19 years. 2015.

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology 2015:277: 134–141
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Smith-Bindman R, Miglioretti DL. CTDIvol, DLP, and Effective Dose are excellent measures for use in CT quality improvement. Radiology. Dec 2011;261(3):999; author reply 999-1000.
An explanation as to why these radiation dose metrics are useful in calculating a patient's absorbed doses.

Huda W, Ogden KM, Khorasani MR. Converting dose-length product to effective dose at CT. Radiology. Sep 2008;248(3):995-1003.
"This article describes a method of providing CT users with a practical and reliable estimate of adult patient EDs by using the DLP displayed on the CT console at the end of any given examination." – Conclusion statement from Abstract

Denominator Statement: Consecutive sample of CTs conducted in the head, chest, abdomen/pelvis and chest/abdomen/pelvis. No examinations should be excluded

Denominator Exclusions: CT examinations conducted in anatomic areas not included above (such as CTs of the extremities or lumbar spine) or that combine several areas (head and chest) should not be included. In children, these four included categories will reflect approximately 80% of CT scans.

Examinations performed as part of diagnostic procedures – such as biopsy procedures – should not be included. CT examinations performed as part of surgical planning or radiation therapy should not be included.

Examinations that are considered "limited abdomen" or "limited pelvis" studies should be included in the abdomen and pelvis

category. Any examinations that include any parts of the abdomen and or pelvis should count in the abdomen/pelvis category.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Registry
Level of Analysis: Facility, Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. The evidence requirements for a *health outcomes measure* include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

The developer appears to intend this as a single measure with two components:

- mean dose as measured using DLP, CTDIvol, and SSDE (within age strata) among consecutive patients who have undergone CT of the head, chest, abdomen/pelvis, or chest/abdomen/pelvis; and
- the proportion of CT exams with doses greater than the 75th percentile of the benchmark for the same anatomic area strata

The evidence should support both components—i.e., that standardizing the mean dose improves outcomes and that the 75th percentile, specifically, improves outcomes. The developer provides the following evidence (Level of Analysis=facility-, health plan-, and integrated delivery system):

- Radiation is a well-studied carcinogen, and the relationship between dose and cancer in the range of CT scanning is linear (in the range of CT), where the higher the dose, the higher the risk.
  - The relationship between dose and risk is thought to be linear in the lower dose range of chest x-rays, and the model describing the relationship between dose and risk is often called the linear, no threshold model, meaning no dose is safe.
  - The linear low dose threshold model does not pertain to doses in the range of CT. In the range of CT, there is directly observed epidemiological data that cancer risks are proportional to dose, and lowering the dose would result in an expected reduction in cancers, especially for children.
  - Extensive epidemiologic and biological evidence supports that radiation doses in the range delivered by medical imaging with CT increase cancer risk.
  - The Board of Radiation Effects Research Division on Earth and Life Sciences *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2 Washington, D.C.* report conducted a review of the literature and concluded that no dose of radiation should be considered completely safe, and attempts should be made to keep radiation doses as low as possible.
- The developer states that facilities are currently using higher doses of radiation for medical imaging with CT than needed for diagnosis, and they are in general unaware of the doses they routinely use for their patients.
- The developer states the measure addresses lack of standardized documentation of radiation doses in children

(while the measure is stratified by age, the measure does not specify an upper age limit) for computed tomography (CT) scans conducted in the head, chest, abdomen/pelvis and chest/abdomen/pelvis.
    - Radiation dose given to patients is generally unknown to physicians and providers. Doses can vary up to 50-fold across institutions for patients imaged for the same clinical reason.
    - Radiation is reported differently across different types of CT machines, which makes it difficult to standardize.
    - A 2010-2011 study within an integrated health care system found that abdominal CT DLPs decreased by 3%-12% at facilities that received dose audit reports and education on dose-reduction strategies.
    - Adoption of standardized reporting of summary dose would allow comparison across providers/facilities and ultimately result in quality improvement and improved patient safety.
    - The developer provided a list of previously conducted studies demonstrating the high variability in radiation doses across facilities. While the evidence from each study was not summarized, the developer did note that in one study, a range of 4.8 to 137 mSv in effective dose for an abdominal CT in children aged 1-4 years.
    - The standard proposed in this measure to collect radiation doses has been studied and standardized in the United Kingdom and Europe for more than 10 years, with a UK study reporting doses to be 50% lower than doses used in the United States.
- The developer provides evidence on the risks associated with CT dose in children:
    - Retrospective, population-based cohort studies compared children in the United Kingdom who received two or more CTs to children who underwent a single CT. Those with multiple CTs had a small but significant increased risk of leukemia and brain cancer.
    - One 2013 study has estimated that the reduction in outlier doses (i.e., doses > 75th percentile in distribution) could reduce the burden of radiation-related cancers in children by 40%.
- The developer indicates this is a measure of an intermediate clinical outcome. If the Committee concurs, per the NQF Evidence Algorithm, the eligible ratings are PASS or NO PASS rating (box 2). If the Committee determines it is a process measure, the evidence is not based on a systematic review, but empirical evidence has been submitted without systematic review and grading. Per the NQF Evidence Algorithm the eligible ratings are MODERATE and LOW (boxes 7-->9). NQF guidance on outcome, intermediate clinical outcome, and process measures is provided here.

Questions for the Committee

- *Is the measure an outcomes measure or a process measure?*
- *Should the components be separated into two distinct measures for separate voting at the in-person meeting?*
- *Does the evidence for each component of the measure support the relationship to outcomes?*
- *The stratification variables include age as a proxy for weight. Does the Committee wish to further discuss with the developer the evidence for using age as a proxy for weight?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following information:

- In the last 10 years, children are receiving 5-fold more abdominal CTs and 50% more head CTs. At these rates, 1 in 3 children will undergo at least one CT scan before their 18th birthday.
- Although children are smaller than adults, and the risk of radiation related to causing cancer is proportional to the radiation per unit of tissue, there is no evidence that radiation doses given to children are lower than the ones given to adults. A 2013 study found variation in doses used in children.
- Pediatric hospitals tend to use lower dose techniques, while adult hospitals do not tailor their CT doses when performing CT scans on children.
- The developer compared the performance of county hospitals that provide care to underserved populations versus non-county hospitals. The county hospitals were found to perform routine CT using doses many times higher than the best performing hospitals that tend to have newer technologies. The developer notes that

county hospitals are more likely to have older equipment that does not allow for reduction in dosage.

- Information on radiation is reported differently across the different types of CT machines, and data are pooled in various ways, making it is difficult for physicians to easily standardize their practice without a common and simple framework for doing so. Currently, physicians do not know the typical radiation doses received by their patients. Creation of a simple standard for collection of radiation dose information would help facilities understand their current practice, would allow understanding changes in practice over time (Keegan 2014; Greenwood 2015) would allow comparisons to local and national standards, and would indicate to facilities whether there is a need to improve.
- Dose metrics collected from 2010-2012 showed a 30-50% decrease in variability of doses after an earlier version of this measure was put into use. Five University of California hospitals reported 0-18% reduction after being given strategies to optimize CT doses. Doses have declined 10-30% across all published studies, with the greater reduction shown among sites with higher doses. The developer provides several tables in this regard.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- No evidence that dose monitoring improved reductions in dosing in any large scale trial, RCT. I am not convinced that measurement in this case will necessarily result in improvement of dosing. I am not convinced that the composite is valid. It is not clear how this measure is intended--intermediate clinical outcome (not really from a poem perspective--patient oriented evidence that matters). Seems more process to me.
- Evidence supports measure directly. The intermediate outcome can be causally linked to the desired (undesired) outcome.
- There is good evidence of dose related risk of radiation. The measure is a process measure to measure and compare dose of pediatric CT to benchmarks. The assumptions are that this will result in more appropriate radiation doses for various procedures that will in the long run reduce risk/incidence of radiation induced malignancy.

1b. Performance Gap

- There certainly is variation, but is it of national significance as constructed--not convinced.
- Gap was identified by variability. Population subgroup by size was not provided except in attached literature. Other subgroups were not identified.
- Yes... there is good evidence that there is significant variation in radiation doses for CTs for children and adults.
- The performance data provided, specifically pediatric CT rates, 2013 study, and performance of county versus non-county hospitals, demonstrate a gap in care (increased abdominal and head CT results compared to adults, increased radiation exposure in county hospitals). Disparities in care are related to CT imaging techniques in county hospitals providing care to underserved populations.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

The developer provides the following information:

- Level of Analysis: Facility, Health Plan, Integrated Delivery System
- The numerator is defined as: Radiation Dose metrics among consecutive patients, who have undergone CT of

the head, chest, abdomen/pelvis, or chest/abdomen/pelvis. The metrics are 1) mean dose as measured using DLP, CTDIvol, and SSDE: within age strata, andnd 2) the proportion of exams with doses greater than the 75th percentile of the benchmark you are comparing with for the same anatomic area strata

- The denominator is defined as: Consecutive sample of CTs conducted in the head, chest, abdomen/pelvis and chest/abdomen/pelvis.
- Variables provided are: Age strata: infant (<1); small child (1-5); medium child (>5 - 10); large child (>10-15) and adult (>15). Age groups were chosen based on required radiation dose depending on patient size, with age used as a surrogate for size. Anatomic area strata: head, chest, abdomen/pelvis; these account for 75% of all CT examinations performed in children.
- The developer states, "The length of time needed to accrue a sufficient number of CT scans to generate sufficient precision will vary by the size of the facility, but for average sized practices, will include review of data from several months. The sample size to generate sufficient precision in each category is 25 CTs within each anatomic and age stratum."
- No type of score is provided, and the developer does not specify if higher or lower = better quality. NQF staff infer that for part 1, lower mean dose is better/higher quality and for part 2, fewer exams greater than the 75th percentile is better/higher quality.
- The measure is not risk-adjusted.
- A calculation algorithm is not included, but the developer includes information on how to extract the numerator information in several different ways.

Questions for the Committee :

- *Are all the data elements clearly defined?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

The developer reports the following:

- Empirical testing was performed at seven integrated health systems and five hospitals during the period 2012-2014, depending on the site. The developer also indicates a Level of Analysis=health plan, but no testing information is provided at the health plan level, as required by NQF.
  - Group Health Research Institute, a large integrated Health System in the Pacific Northwest; CT examinations on more than 10,000 examinations were assembled.
  - A consortium of six integrated health care systems with data from more than 5,000 CT examinations.
  - Five University of California medical centers with data on more than 100,000 CT examinations.
- Reliability testing was done at the level of data elements using several metrics reflecting CT dose indices including DLP, CTDIvol, and SSDE.
  - DLP and CTDI are calculated automatically by all current CT scanners, without variability. Reliability of CT radiation dose metric abstraction (DLP and CTDIvol) was tested through both manual and automated data abstraction, both yielding identical results, perfect Kappa statistics.
  - SSDE is a calculated variable that is automatically calculated by dose monitoring programs. Errors from manual calculation were not tested.
- The measure was tested using data from clinical database/registry and data abstracted from electronic health records. Additional data included data extracted from stored CT images.
- The reliability testing results had Kappas greater than 95%.
- Patient-level sociodemographic (SDS) variables were not available nor tested.

Questions for the Committee:

- *Is the empirical reliability testing methodology appropriate?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified at all three levels of analyses (facility, integrated care delivery system, health plan)?*
- *Does the Committee wish to discuss further the lack of reliability testing data at the health plan level?*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The goal of the measure is to improve care for pediatric patients receiving CTs by measuring 1) mean dose using DLP, CTDIvol, and SSDE (within age strata) among consecutive patients who have undergone CT of the head, chest, abdomen/pelvis, or chest/abdomen/pelvis; and 2) the proportion of CT exams with doses greater than the 75th percentile of the benchmark for the same anatomic area strata.
- The developer identifies a stratification scheme by age as a proxy for weight.
- The developer states evidence for a structured framework is the current lack of lack of standardized documentation of radiation doses in children, radiation dose dose given to patients is generally unknown to physicians and providers, doses can vary up to 50-fold across institutions for patients imaged for the same clinical reason. radiation is reported differently across different types of CT machines. The developer further notes increased adverse outcomes (leukemia and head cancers) are associated with multiple CTs and data from one 2013 study estimated that the reduction in outlier doses (i.e., doses > 75th percentile in distribution) could reduce the burden of radiation-related cancers in children by 40%.

Question for the Committee:

- *Are the specifications consistent with the evidence?*
- *Does the age range, in particular the upper age limit, need to be clarified?*
- *Does the Committee wish to discuss the stratification scheme of age as a proxy for weight with the developer?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer reports the following information:

- Empirical testing was performed at the performance measure score.
- The developer indicated that a study was conducted comparing each of the dose metrics with measures of absorbed dose among a sample of 10,000 CT examinations showed a "high correlation," > 90%.
  - The developer does not summarize results in the NQF Testing form in a narrative, but provides an appendix with a dose report.
  - The developer also provides additional analyses on pre- and post-implementation of the measure at five University of California medical centers.

Questions for the Committee

- *Was the empirical validity testing methodology appropriate?*
- *Was the testing of the measure as specified?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made at the facility level? at the integrated care delivery system level? at the health plan level?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The developer provides the following information:

- The developer indicates on the Testing Form that there are no exclusions.
- Elsewhere, however, the developer states the measure has the following exclusions:
  - CT examinations conducted in anatomic areas not included above (such as CTs of the extremities or lumbar spine) or that combine several areas (head and chest) should not be included. In children, these four included categories will reflect approximately 80% of CT scans, but no specific data are provided from each testing site.
  - Examinations performed as part of diagnostic procedures – such as biopsy procedures – should not be included. CT examinations performed as part of surgical planning or radiation therapy should not be included. No additional data are provided on frequency or effect on
  - Examinations that are considered "limited abdomen" or "limited pelvis" studies should be included in the abdomen and pelvis category. Any examinations that include any parts of the abdomen and or pelvis should count in the abdomen/pelvis category. No additional data are provided on frequency or effect on performance score.

Questions for the Committee

- *Are these appropriate exclusions?*
- *Does the Committee wish to discuss with the developer additional analyses re: the exclusions?*
- *Does the Committee believe there are other threats to validity?*

2b4. Risk adjustment:

The developer provides the following information:

- The measure is stratified by three anatomic areas and five pediatric age groups:
  - Anatomic areas: head, chest, abdomen/pelvis, Chest/abdomen/pelvis
  - Age groups: infant (<1); small child (1-5); medium child (>5 - 10); large child (>10-15) and adult (>15)
- The developer states that the measure should not be stratified for clinical indication or protocol because the reason for scanning will affect how a scan is performed.
- The developer states that it is not important to adjust by patient size because that does not ultimately affect the dosage very much: “weight differences are not relevant at the facility level, as while patient size may influence dose by 2-fold (between the smallest and largest patients) other factors, can influence the dose by up to 100 fold (based on our data), and these factors, rather than individual patient weight, will drive the facility level dose indices measures. Even if a facility had ALL patients of a size <25%, versus all patients over the 75% the influence would be very modest.”
- The developer indicates that recent publications provide ways to account for size when reporting radiation dose. The developer has incorporated this new measurement, SSDE, into the measure to assist with greater adoption of the measure. SSDE is SSDE is a modified measure of CTDIvol that takes into account the size of the patient scanned and is useful for scaling dose to patient size. Several current radiation tracking software tools directly report SSDE.

Questions for the Committee

- *Is there any evidence that contradicts the developer’s rationale and analysis?*
- *Are the variables included in the risk adjustment model adequately described for the measure to be implemented?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provides the following information on identifying meaningful differences:

- Comparing institutional performance to benchmarks permits identification of outlying performance. Because the metric is based on summarizing dose for a large number of individuals (> 100 within each strata) and comparison to benchmarks, the comparisons are stable at identifying outlying performance.
- The developer provides an attachment that illustrates the result of comparing institutions (using t-tests and quantile regression). Information on meaningful differences is not called out for integrated care systems or health plans in this appendix.

Question for the Committee

- *Does this measure identify meaningful differences in quality among hospitals? among integrated care systems? among health plans?*

2b7. Missing Data

- The developer provides the following information regarding missing data: The measure calls for collecting consecutive scans so that participants cannot choose their best or most optimum dose metrics to quantify. The data will be available, or can be calculated from essentially all (>95%) of CT scans .

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- There are several ways to measure dose.. automated and manual. There undoubtedly will be issues with interrater reliability... but as the intent of the measure is to create a database for measuring variation and establishing benchmarks, this will likely not prevent achievement of the purpose of the measure. And by increasing the awareness of the issue, CT practices related to using appropriate dosing for a given procedure (or patient size) will likely improve.

2a2. Reliability testing

- Don't see this has been done adequately for all levels of analysis.
- Reliability testing should Kappas greater than 95%... but for a limited number of sites. Not sure if this degree of reliability will be achieved if measured at all hospitals.
- I am not clear how much variability might be introduced with the dose measurement, why the chosen strata were made. Not sure if all areas are equal--ie, head and abdomen.
- Three dosing algorithms are proposed with one needing calculation and the others with varying degrees of automatic reporting. Consecutive sampling is proposed to assure continuous results.
- Specifications are not present for older machines which are likely a significant part of the high dose issue.
- The specifications seem consistent with the evidence provided. The data elements are clearly defined. It may not be appropriate, in some children, to use age as a substitute for size (particularly in children who are small for their age). Need more evidence to assure that metrics to measure radiation dose, particularly SSDE, can be consistently implemented.

2b1. Validity Specifications

- Age/weight proxy? Age limit/variation
- This is a dose per test measure and not a dose per patient measure. The measure hence does not measure patient specific effects of repeated high dosing (e.g. CTs of the head for shunt malfunction).
- Codes for exclusions are not offered or for some inclusions (limited CT) (unless further along in the document.

2b2. Validity Testing

- Validity testing was conducted at the performance measure level. An adequate study size of 10,000 CT examinations was reported showing a high correlation. Additional analyses on pre- and post-implementation of the measure at 5 medical centers is also provided. Data provided support that conclusions about quality can be made at the facility level but further application may be challenging.

2b3-2b7. Threats to Validity

- Exclusions needs to be clarified, and validated. Risk adjustment and variation by site and age is not clear. Does there need to be outlier analysis have to be performed at a individual or unit level? Maybe a single operator is the biggest issue, for example.
- As above missing data from older machines.
- As the developer identifies that there are both "no exclusions" and "the measure has the following exclusions", this would need further clarification. Would request additional rationale for the exclusion criteria.
- The measure would identify meaningful differences in quality among hospitals but may be difficult at the integrated care system or health plan level.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer provides the following information:

- Two of the specified metrics (CTDIvol and DLP) are generated as part of clinical CT examinations. The two additional metrics can be calculated from these two primary metrics, and these calculations are done within existing software products or can be done manually, or using various additional approaches.
- Nearly all facilities that perform CT examinations can collect all the measure elements (three dose metrics: DLP, CTDI and SSDE).
- DLP and CTDIvol are available on nearly all (>95%) of CT scans conducted in the United States.
- SSDE can be calculated manually and is currently calculated by many vendors who developed software to extract radiation dose metrics from CT machines.
- CT manufacturers have agreed to adopt the same standard for reporting the radiology dose data, with all machines built after late 2010 equipped with this feature. Several excel based programs can be used to calculate the measure for facilities with older machines.
- Results from testing demonstrate that the measure is feasible for clinicians to report.

Questions for the Committee

- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments

Criteria 3: Feasibility

- Seems like this could be feasible, although smaller facilities might have more work.
- Two of the specified metrics are routinely generated during care delivery however SSDE often requires a manual calculation which may limit the operational use of the measure.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

The developer provides the following information:

- The measure is planned to be used for public reporting, and both internal and external quality improvement benchmarking for healthcare facilities. Starting in 2015, the Joint Commission is asking facilities it oversees to begin assessing radiation doses. The developer leads a patient safety project based at UCSF with 150 member hospitals/facilities that could begin providing data for benchmarking.
- This is the revision of a previously NQF-endorsed measure. Use of the prior measure assessing facility-level and

provider-level radiation dose metrics demonstrated substantial improvements in dose over time (Keegan, 2014; Demb, 2015; Duncan, JACR 2013) and as the result of a randomized trial of an educational intervention and process, whereby technologists were shown their performance using the dose summary that was designed to follow the previously endorsed measure (Miglioretti, 2014).

- The developer does not report any unintended consequences, but does highlight two potential limitations:
    - Patient size: One factor that influences the radiation dose in CT is patient size. In general higher doses are used in large patients in order to maintain the same image quality as can be achieved with lower doses in smaller patients. It simply takes higher doses of radiation to penetrate (get through) larger sized patients. Thus the recorded radiation doses in part will reflect the size of the patients seen. If a facility sees a very high proportion of obese patients, their doses will be higher than a facility that sees very thin patients. This issue will be important when facilities compare their dose indices to normative data (to the diagnostic reference level data), as they should compare their actual data to data of facilities that assess similar patients.
        - To address this issue, state should be reported; this diagnostic reference data should reflect geographic differences and be appropriate to the typical patients seen in a given area, as called for in the FDA white paper on radiation safety.
        - The developer also notes that patient size will be a relatively small impact on overall dosage and that none of the current quality programs assess patient size in conjunction with dose, for feasibility issues.
    - CT protocols: The way CT studies are conducted (the "protocols" using the language of CT) leads to the radiation doses patients will receive. These are the specific instructions the radiologist or other physician and technologists program into the CT machine at the time of scanning. If a larger anatomic area is imaged, the dose the patient receives will be higher. If a multiphase study is done (meaning a single anatomic area is imaged many times) the dose will be higher than if a single-phase study is done. If a facility chooses to use multiphase protocols frequently, or to scan large anatomic areas frequently, their doses will be higher than facilities that try to minimize the area imaged or number of scans taken.
        - The two ways to collect and compare CT dose index information would be first to compare doses WITHIN the specific study type - thus compare doses for routine single phase studies and compare doses for multiphase studies, or second, to compare typical doses for all patients who undergo a CT within a single anatomic area (ignoring considering of the specific protocol used).
        - The developer notes, however, there are no evidenced based guidelines about when particular protocols should be used. In particular the multiphase, higher dose protocols are not clearly indicated in particular clinical situation, studies have not shown they lead to improved diagnoses or quantified the potential harm in their use, and differences reflect practice variation more than any objective criteria of the need for these multiphase, studies. While higher dose protocols do have value – but decisions about when to use different protocols are more based on physician preferences that patient outcomes, and choosing to frequently use these higher dose protocols should be reflected in the radiation dose quality metrics generated at a facility.

Questions for the Committee

- *Do the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Seems likely the use of single versus multiphase could significantly influence results, and local (albeit non-evidence-based differences in protocol) may be a significant factor influencing variation.
- This measure identifies one aspect of the imaging conundrum, the dose of individual tests, but does not look at the changes to overall imaging to move from ionizing radiation to other less risky forms for diagnostic accuracy.

Criterion 5: Related and Competing Measures
• This measure, #2820, is an update to a previously endorsed measure, NQF 0739: Radiation Dose of Computed Tomography (CT).

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Measure similar but not the same was 0739

Measure Title: Pediatric Computed Tomography (CT) Radiation Dose

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 9/28/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☑Health outcome:

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☑Intermediate clinical outcome (*e.g., lab value*): Distribution in radiation dose metrics (i.e. mean, median, and percent of exams greater than the 75% benchmark values for the following specific CT radiation dose metrics: CTDIvol, DLP and SSDE) associated with computed tomography (CT) examinations of the head, chest, and abdomen/pelvis and chest/abdomen/pelvis performed among children (within specified age strata). These metrics are calculated at the facility or health plan level or institutional level.

☐Process:

☐Structure: Click here to name the structure

☐Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Brief Summary: Radiation is a well-studied carcinogen, and the relationship between dose and cancer in the range of CT scanning is linear (in the range of CT), where the higher the dose, the higher the risk. Of note, the

relations ship between dose and risk is thought to be linear in the lower dose range of chest xrays, and the model describing the relationship between dose and risk is often called the linear, no threshold model, meaning no dose is safe (BEIR VII, 2006.) However, the linear low dose threshold model does not pertain to doses in the range of CT. In the range of CT, there is directly observed epidemiological data that cancer risks are proportional to dose, and lowering the dose would result in an expected reduction in cancers, especially for children.

Because radiation it is a known carcinogen, it must be used in the safest way possible. Facilities are currently using higher doses of radiation for medical imaging with CT then needed for diagnosis *(Smith-Bindman JAMA Int Med 2009; JAMA 2012; JAMA Pediatrics 2013.)* Further, they are in general unaware of the doses they routinely use for their patients. The adoption of a standard metric for summarizing dose at the facility level would allow facilities to compare their performance to other facilities, and pooling dose data created can further be used to generate benchmarks for CT. This process of assessment of dose and comparison to benchmarks would enable facilities to lower the doses they use and thereby reduce this important potential harm of imaging. Miglioretti et al (*JAMA Pediatrics 2013*) has estimated that the reduction in the outlier doses (i.e., doses > 75th percentile in distribution) could reduce the burden of radiation related cancers in children by 40%. Radiologists determine how the CT tests are performed. However, there are few national guidelines on how these studies should be conducted and, therefore, there is great potential for practice variation that could introduce unnecessary harm from excessive radiation dosing. Furthermore, since information on radiation is reported differently across the different types of CT machines, it is difficult for radiologists to standardize their practice. Currently, radiologists do not know the typical radiation doses received by their patients. Almost certainly non-radiologists who are conducting CT studies also do not know the radiation doses delivered to their patients. Facilities that complete the data analysis as part of this measure would rapidly understand the doses they use and how they compare to other facilities, and would motivate improvement. Further, if this measure was adopted by quality organizations, assessment of facility processes of reviewing dose could further improve quality.

Details of Rational for Measure

<u>Radiation can be harmful:</u> Radiation is one of the most heavily studied carcinogens, and extensive epidemiologic and biological evidence supports that radiation doses in the range delivered by medical imaging with CT increase cancer risk. The epidemiologic evidence comes from studies indicating cancer development among survivors of environmental and accidental exposures, populations repeatedly irradiated for benign conditions or diagnostic imaging, patients receiving radiotherapy for malignant disease, and people who received occupational exposure, such as radiologists and nuclear power workers. *(BEIR VII Report)*. The literature on the health effects of exposure to ionizing radiation is summarized in the BEIR VII phase 2 report (Board of Radiation Effects Research Division on Earth and Life Sciences "Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2 Washington, D.C." The National Academies Press, 2006. The BEIR VII committee, the most widely sited source on the topic, concluded after an exhaustive review of the literature that no dose of radiation should be considered completely safe, and attempts should be made to keep radiation doses as low as possible. As part of their report, The BEIR VII report presented the best risk estimates for exposure to low-dose, radiation in human subjects, which largely rely in large part on results of the Life Span Study (LSS), the study of the 120,000 survivors of the atomic bombings in Hiroshima and Nagasaki Japan. Organ specific radiation doses are linked with organ specific risks of cancer and cancer mortality. Researchers have used these data to estimate the risk of exposure to a single medical imaging study. For example, Einstein and colleagues estimated the risk of cancer associated with the radiation exposure from a single 64-slice computed tomography coronary angiography was as high as a 1/114. Smith-Bindman found in our work that the risk of cancer could be as high as 1/80. (JAMA Internal Medicine 2009)

Direct studies of CT Studies have directly assessed cancer risk associated with CT. Retrospective, population-based cohort studies by Pearce et al. *(Lancet 2012)* compared children in the UK who received two or more CTs to children who underwent a single CT. Those with multiple CTs had a small but significant increased risk of leukemia and brain cancer. Thus Radiation in the same dose range as used with Computed Tomography is known to be carcinogenic.

The risk of radiation induced cancer, is widely believed to be approximately proportional to the level of radiation exposure. Reduction in radiation exposure will be associated with reduction in cancer risk *(BEIR VII Phase 2, 2006; JAMA Internal Medicine 2009, Berrington de Gonzales 2009; Miglioretti JAMA Pediatrics 2013)*

Currently no formal program of oversight

Although radiation dose information is not currently collected in the US, programs exist in many European countries, Canada and Asia, for collecting the dose information using the indices specified in this measure. They have found the doses can be reduced through data collection and reporting. These programs have collected and reported dose information for many years, largely using voluntary programs, and this has resulted in a lowering of typical radiation dose. The most well-known and published program is run through the National Radiological Protection Board (NRPB) in the United Kingdom. The most recent report, NPRB-W67, describes a snapshot of patient CT dose. (Doses from Computed Tomography (CT) Examinations in the UK - 2003 Review. Shrimpton PC et al. National Radiological Protection Board, Childton, Didcot, Oxon, ISBN 0 859515567, http://www.mendeley.com/research/nrpbw67-doses-from-computed-tomography-ct-examinations-in-the-uk-2003-review/) The doses described in this report are on average approximately 50% lower than the doses used in the US. The near absence of widely collected data on current doses in the US, agreed upon standards for how the CTs should be programmed (meaning how these complex machines should be instructed to conduct the examinations), or an agreed upon metric whereby data could be collected and analyzed across facilities has led to the current situation where each facility decides on how to set up their individual CT scans. Further, the absence of widely published guidelines for acceptable ranges of dose in the US would make it difficult for an institution to know if they are doing well in minimizing this important harm of CT.

Oversight of CT is limited and highly fragmented, with no single organization assigned responsibility to ensure the standardization of CT dose when used in clinical practice. For example, while the FDA monitors the manufacture of CT machines, they do not assess how they are used in routine practice and they do not collect information on actual clinical practice. However, the FDA, have recently highlighted in their white paper on minimizing radiation dose the pressing need to collect dose information associated with the most common types of diagnostic CT and to use these data to generate standards for targeted dose.

The Joint Commission has recently instituted oversight of CT, and for hospitals and outpatient hospital facilities that have certification through the Joint Commission, the oversight will help facilities measure and report doses. In contrast, outpatient facilities can easily fall through this oversight of the Joint Commission.

Radiation doses used in clinical practice are highly variable: CT radiation doses are higher and more variable than widely reported, and can vary up to 50-fold across institutions for patients imaged for the same clinical reason *(Miglioretti JAMA Pediatrics 2013; Smith-Bindman JAMA 2012; Smith-Bindman JAMA Internal Medicine 2009).* We have published extensively on this variation. For example, we found a range of 4.8 to 137 mSv in effective dose for an abdominal CT in children aged 1-4 years. Only a small part of the variation is due

to appropriate accommodation of patients of difference sizes; most variation reflects physician and technologist preferences, rather than doses needed for improved diagnosis.

The doses used for CT can be readily reduced, thereby reducing the risks of imaging, by 40% or more without loss of diagnostic accuracy.

The first step towards reducing dose is for facilities to quantify their doses. The NQF endorsed measure provides the only simply way for facilities to compare their doses to national norms, and thereby reduce the high doses they use in their patients. The comparison to benchmarks had been done in the UK as part of the National Health Service Health Protection Agency Program for over 10 years. Two recent papers used the endorsed NQF measure as the framework for assessing the doses they used *(Keegan Journal of the American College of Radiology 2014; Miglioretti, Journal of the American College of Radiology 2014*)

Adoption of a simple standard for collection of radiation dose information would help facilities understand their current practice, would allow comparisons to local and national standards, and would indicate to facilities whether there is a need to improve. There is currently a high level of interest in this area - facilities are being asked by their patients and governing boards to report whether they are performing CT safely - and this measure is an ideal starting point for facilities to assemble this information to answer these questions. If facilities collect dose information, it is the first step towards trying to compete on a measure of safety and I envision facilities will begin to do all they can to lower the doses they use.

The measure will facilitate to the creation of regional and national diagnostic reference levels, improve dose awareness and inevitable improvements, as it will enable physicians to consider dose as an important measure.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐Yes → ***complete section 1a.7***

☐No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence <u>with definition</u> of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Several observational studies and a randomized trial have shown that review of the doses as defined in this measure lead to a reduction in dose

One study, conducted within an integrated health care system from November 2010 to October 2011, used ten technologists at 2 facilities, who received personalized dose audit reports and education on dose-reduction strategies. A control of 9 technologists at another facility received no interventions. Technologists were then surveyed before and after the intervention. And it was found that abdominal CT DLPs decreased by 3% to 12% at intervention facilities, but not at the control facility, 7% to 12% at one intervention facility for brain CT DLPs, and one control facility even increased their DLPs. It was ultimately found that technologists were more likely to report always thinking about radiation exposure and associated cancer risk and optimizing settings to reduce exposure after they have personalized audits and intervention strategies. (Miglioretti, JACR, 2014)

Following the passing of NQF Measure #0739, manual and electronic scans were collected for their dose metric statistics. These collection processes were timed and evaluated for their effectiveness. Fifty manual scans required 2 hours and 15 minutes, whereas the dose extraction tool eXposure compiled the data in an hour. All dose metrics, which were abstracted from 2010 to 2012, showed a 30-50% decrease in their variability of doses. Thus, it was found that this measure's passing facilitated the facility's dose reduction, while it was in effect (Keegan, JACR, 2014).

Following an in person meeting, five University of California hospitals were given specific strategies on how they could optimize CT doses. Those strategies were made during the meeting, then evaluated for their effectiveness and sustainability at several time periods afterwards. It has been found that there has been a general reduction of 0-18% doses due to this (Demb, 2015)

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

There are many studies demonstrating that doses are highly variable across facilities and that physicians are not aware of these doses. The papers described above have shown that facilities or technologists who have been provided with their doses in turn lower their doses.

American College of Radiology (ACR). Dose Index Registry (DIR). 2014. http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/DIR%20Measures.pdf

Registry designed to showcase measures for certain CT procedure types.

Calvert C, Strauss KJ, Mooney DP. Variation in computed tomography radiation dose in community hospitals. Journal of pediatric surgery. Jun 2012;47(6):1167-1169.

Demb J, manuscript under preparation. CT Radiation Dose Standardization Across the University of California Medical Centers Using Audits to Optimize Dose. 2015.

Following an in-person meeting regarding CT radiation dose, radiologists, technologists and medical physicists from University of California medical centers strategized how to best optimize dosing practices at their sites, which were then analyzed for effectiveness and success after implementation.

Dorfman AL, Fazel R, Einstein AJ, et al. Use of Medical Imaging Procedures With Ionizing Radiation in Children: A Population-Based Study. Arch Pediatr Adolesc Med. Jan 3 2011.

Duncan J, Street M, Strother M, et al. Optimizing Radiation Use During Fluoroscopic Procedures: A Quality and Safety Improvement Project. J Am Coll Radiol. 2013;10:847-853

Einstein AJ, Henzlova MJ, Rajagopalan S. Estimating risk of cancer associated with radiation exposure from 64-slice computed tomography coronary angiography. JAMA 2007;298:317-23.

Greenwood T, Lopez-Costa R, Rhoades P, et al. CT Dose Optimization in Pediatric Radiology: A Multiyear Effort to Preserve the Benefits of Imaging While Reducing the Risks. RadioGraphics. Jan 2015;35(5):1539-1554

“This systematic approach involving education, streamlining access to magnetic resonance imaging and ultrasonography, auditing with comparison with benchmarks, applying modern CT technology, and revising CT protocols has led to a more than twofold reduction in CT radiation exposure between 2005 and 2012…” – Conclusion statement from Abstract

Hausleiter, J., T. Meyer, et al. Estimated radiation dose associated with cardiac CT angiography. JAMA 301(5): 500-7. 2009

“Median doses of CCTA differ significantly between study sites and CT systems. Effective strategies to reduce radiation dose are available but some strategies are not frequently used. The comparable diagnostic image quality may support an increased use of dose-saving strategies in adequately selected patients.”– Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315.

http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014

Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Kumar K, manuscript under preparation. Radiation Dose Benchmarks in Children.

This paper will describe dose metrics among 29,000 children within age strata <1, 1-4 years, 5-9 years, 10-14 years, and 15-19 years. 2015.

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311. 2013.

Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered.

Miglioretti, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. Journal of the American College of Radiology: JACR 2014.

“Personalized audit feedback and education can change technologists' attitudes about, and awareness of, radiation and can lower patient radiation exposure from CT imaging.” – Conclusion statement from Abstract

Nationwide Evaluation of X-ray Trends: NEXT 2005-2006. This presentation was given by David Spelic, physicist with the Food and Drug Administration (FDA), to the 39th Conference of Radiation Control Program Directors (CRCPD) annual meeting, held in Spokane Washington, May 21-24, 2007.

Morin, R. L. (2006). "CT dosimetry--an enigma surrounded by a conundrum." J Am Coll Radiol 3(8): 630.

Morin, R. L. (2006). "What are the national radiation doses?" J Am Coll Radiol 3(12): 956.

Parker M, Shah S, Hall M, et al. Computed Tomography and Shifts to Alternate Imaging Modalities in Hospitalized Children. Pediatrics. 2015-0995.

"For the 10 most common All-Patient Refined Diagnosis Related Groups (APR-DRGs) for which children received CT in 2004, a decrease in CT utilization was found in 2012. Alternative imaging modalities for 8 of the diagnoses were used." – Conclusion statement from Abstract

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med 2009;169:2078-86.

" Radiation doses from commonly performed diagnostic CT examinations are higher and more variable than generally quoted, highlighting the need for greater standardization across institutions." – Conclusion statement from Abstract

Smith-Bindman R. Is computed tomography safe? N Engl J Med 2010;363:1-4.

Smith-Bindman R. Environmental causes of breast cancer and radiation from medical imaging: findings from the Institute of Medicine report. Arch Intern Med 2012;172:1023-7.

Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012;307:2400-9.

" Within integrated health care systems, there was a large increase in the rate of advanced diagnostic imaging and associated radiation exposure between 1996 and 2010." – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology 2015:277: 134–141

"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Wilson N. CT Radiation Dose Standardization Across the Five University of California Medical Centers. ARRS: Annual Toronto Meeting presentation. April 19-24, 2015

Understanding the reasons for variation in commonly performed CT procedures, and figuring out how to standardize them.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Doses have declined 10-30% across all published studies, with the greater reduction shown among sites with higher doses

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

Theoretically doses could be made too low. We have not seen that in our work or any publication. Radiologists are extremely sensitive to the quality of their images and would be expected to complain about the doses if they became too low.

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

Radiation dose associated with CT has been studied and standardized in the UK and Europe for over 10 years, although no comparable work has been done in the US. The metrics to measure dose have been well established, and data in the UK has been collected in a format that parallels the method proposed in this measure.

The recommended measure is a technique for summarizing dose and is simple and straightforward. Facilities summarize the doses they use in consecutive patients so that they can compare their doses to normative data. Many of the sources shown in 1a.7.6 also showcase methods and the need for lowering CT dose.

1a.8.2. Provide the citation and summary for each piece of evidence.

Amis ES, Jr., Butler PF, Applegate KE, et al. American College of Radiology white paper on radiation dose in medicine. J Am Coll Radiol 2007;4:272-84.

Hausleiter, J., T. Meyer, et al. Estimated radiation dose associated with cardiac CT angiography. JAMA 301(5): 500-7. 2009

"Median doses of CCTA differ significantly between study sites and CT systems. Effective strategies to reduce radiation dose are available but some strategies are not frequently used. The comparable diagnostic image quality may support an increased use of dose-saving strategies in adequately selected patients."– Conclusion statement from Abstract

Hricak H, Brenner DJ, Adelstein SJ, et al. Managing Radiation Use in Medical Imaging: A Multifaceted Challenge. Radiology 2010.

Board of Radiation Effects Research Division on Earth and Life Sciences National Research Council of the National Academies. Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2 Washington, D.C.: The National Academies Press; 2006.

Demb J, manuscript under preparation. CT Radiation Dose Standardization Across the University of California Medical Centers Using Audits to Optimize Dose. 2015.

Following an in-person meeting regarding CT radiation dose, radiologists, technologists and medical physicists from University of California medical centers strategized how to best optimize dosing practices at their sites, which were then analyzed for effectiveness and success after implementation.

Einstein AJ, Henzlova MJ, Rajagopalan S. Estimating risk of cancer associated with radiation exposure from 64-slice computed tomography coronary angiography. JAMA 2007;298:317-23.

"… estimates derived from our simulation models suggest that use of 64-slice CTCA is associated with a nonnegligible LAR (lifetime attributable risk) of cancer. This risk varies markedly and is considerably greater for women, younger patients..."– Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014

Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Mathews J, Forsythe A, Brady Z, al. e. Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians. BMJ 2013;346 doi: http://dx.doi.org/10.1136/bmj.f2360

"Future CT scans should be limited to situations where there is a definite clinical indication, with every scan optimised to provide a diagnostic CT image at the lowest possible radiation dose." – Conclusion statement from Abstract

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311 2013.

Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered.

Miglioretti, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. Journal of the American College of Radiology: JACR 2014.

"Personalized audit feedback and education can change technologists' attitudes about, and awareness of, radiation and can lower patient radiation exposure from CT imaging." – Conclusion statement from Abstract

Nationwide Evaluation of X-ray Trends: NEXT 2005-2006. This presentation was given by David Spelic, physicist with the Food and Drug Administration (FDA), to the 39th Conference of Radiation Control Program Directors (CRCPD) annual meeting, held in Spokane Washington, May 21-24, 2007.

Morin, R. L. (2006). "CT dosimetry--an enigma surrounded by a conundrum." J Am Coll Radiol 3(8): 630.

An explanation of the difficulties surrounding CT dosing and estimations of its harmful effects.

Morin, R. L. (2006). "What are the national radiation doses?" J Am Coll Radiol 3(12): 956.

An explanation of why benchmarks or national measures are so difficult to set (related to the article listed above).

Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet 2012;380:499-505.

"Use of CT scans in children to deliver cumulative doses of about 50 mGy might almost triple the risk of leukaemia and doses of about 60 mGy might triple the risk of brain cancer… although clinical benefits should outweigh the small absolute risks, radiation doses from CT scans ought to be kept as low as possible" – Conclusion statement from Abstract

Preston DL, Ron E, Tokuoka S, et al. Solid cancer incidence in atomic bomb survivors: 1958-1998. Radiat Res 2007;168:1-64.

Preston RJ. Update on linear non-threshold dose-response model and implications for diagnostic radiology procedures. Health Phys 2008;95:541-6.

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med 2009;169:2078-86.

" Radiation doses from commonly performed diagnostic CT examinations are higher and more variable than generally quoted, highlighting the need for greater standardization across institutions." – Conclusion statement from Abstract

Smith-Bindman R. Is computed tomography safe? N Engl J Med 2010;363:1-4.

An explanation of the harmful effects of CT overdose, and why its diagnostic purposes are often misused.

Smith-Bindman R. Environmental causes of breast cancer and radiation from medical imaging: findings from the Institute of Medicine report. Arch Intern Med 2012;172:1023-7.

“ The IOM's conclusion of a causal relation between radiation exposure and cancer is consistent with a large and varied literature showing that exposure to radiation in the same range as used for computed tomography will increase the risk of cancer.” – Conclusion statement from Abstract

Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012;307:2400-9.

“ Within integrated health care systems, there was a large increase in the rate of advanced diagnostic imaging and associated radiation exposure between 1996 and 2010.” – Conclusion statement from Abstract

Tables from Specification Section:

Table 2

Radiation Dose Metrics in Children

Area and Examination Type	No. of Examinations	$CTDI_{vol}$ (mGy) 25th Percentile	$CTDI_{vol}$ (mGy) 50th Percentile	$CTDI_{vol}$ (mGy) 75th Percentile	DLP (mGy · cm) 25th Percentile	DLP (mGy · cm) 50th Percentile	DLP (mGy · cm) 75th Percentile	Effective Dose (mSv) 25th Percentile	Effective Dose (mSv) 50th Percentile	Effective Dose (mSv) 75th Percentile
Head										
Single phase	1116				290	420	570	1	2	3
Multiphase	166				540	870	1310	3	4	6
All	1282	22	30	38	310	450	650	1	2	4
Chest										
Single phase	292				50	90	130	2	3	4
Multiphase	63				70	150	210	2	6	7
All*	355	2 (3)	3 (4)	5 (6)	60	90	150	2	3	5
Abdomen										
Single phase	625				80	140	230	3	4	6
Multiphase	83				120	210	330	4	6	10
All*	708	2 (4)	4 (5)	5 (9)	90	140	230	3	4	7
Chest and abdomen										
Single phase	49				110	240	380	3	6	12
Multiphase	35				210	300	840	7	9	20
All*	84	3 (4)	4 (6)	8 (12)	130	270	510	5	9	15
Sinus										
Single phase	153				110	270	460	<0.5	1	2
Multiphase	32				270	570	850	1	2	4
All	185	9	18	28	150	310	500	1	1	2
Neck										
Single phase	103				90	140	340	2	3	6
Multiphase	16				170	270	590	4	6	9
All	119	5	6	13	100	160	340	2	4	6
All other areas	1138									

Note.—Examinations were performed in children younger than 1 year (n = 483 [12.5%]), 1–4 years (n = 949 [24.5%]), 5–9 years (n = 991 [25.6%]), and 10–14 years (n = 1448 [37.4%]).

* Numbers in parentheses are SSDEs, which reflect an adjusted $CTDI_{vol}$ measurement.

Table 2 from Smith-Bindman, Radiology, 2015 showing the dose benchmarks in children, and where the hospitals within that paper fell, within those percentiles.

Tables from Importance Section of Application

Table 2. Summary of dose metrics including the 25th, 50th and 75th percentiles, by anatomic area and study year

	$CTDI_{vol}$ (mGy) 2010	$CTDI_{vol}$ (mGy) 2012	$CTDI_{vol}$ (mGy) % change	DLP (mGy cm) 2010	DLP (mGy cm) 2012	DLP (mGy cm) % change	E (mSv) 2010	E (mSv) 2012	E (mSv) % change
Head									
25%	49	28	−44%	1,127	491	−56%	2.1	1.0	−52%
50%	57	33	−41%	1,205	645	−46%	2.4	1.3	−46%
75%	68	49	−29%	1,394	945	−32%	2.8	1.9	−32%
Chest									
25%	5.1	3.5	−31%	174	119	−32%	3.5	2.4	−31%
50%	8.1	5.5	−32%	282	189	−33%	5.2	3.6	−31%
75%	15.5	9.5	−39%	491	333	−32%	9.7	6.1	−37%
Abdomen/Pelvis									
25%	7.6	4.5	−41%	496	247	−50%	8.5	4.1	−52%
50%	11.7	6.5	−44%	808	490	−39%	14.1	8.1	−43%
75%	17.5	12.2	−30%	1,304	890	−32%	21.6	14.9	−31%

Values may appear to be off because of rounding. $CTDI_{vol}$ = volume of CT dose index; DLP = dose-length product.

Table 2 from Keegan, JACR, 2014 that shows a reduction in dose metrics after the prior NQF Measure was included

Table 3. Summary of $CTDI_{vol}$ and SSDE for abdomen and pelvis by study year

	$CTDI_{vol}$ (mGy)			SSDE (mGy)		
Percentile	2010	2012	% change	2010	2012	% change
25%	7.6	4.5	−41%	10.5	6.4	−39%
50%	11.7	6.5	−44%	14.9	7.9	−47%
75%	17.5	12.2	−30%	19.8	13.7	−31%

$CTDI_{vol}$ = volume of CT dose index; SSDE = size-specific dose estimate.

Table 3 from Keegan, JACR, 2014 that shows a reduction in dose metrics after the prior NQF Measure was included

Table 2. Characteristics of CT examinations and patients

	Intervention Facility 1		Intervention Facility 2		Control Facility	
Variable	Pre	Post	Pre	Post	Pre	Post
Total	589	554	380	390	661	555
CT type						
Abdominal and pelvic	173 (29%)	142 (26%)	123 (32%)	103 (26%)	234 (35%)	162 (29%)
Brain	143 (24%)	145 (26%)	99 (26%)	102 (26%)	189 (29%)	149 (27%)
Chest	145 (25%)	140 (25%)	91 (24%)	101 (26%)	145 (22%)	146 (26%)
Maxillofacial/sinus	128 (22%)	127 (23%)	67 (18%)	84 (22%)	93 (14%)	98 (18%)
Age (y)						
15–29	60 (10%)	60 (11%)	29 (8%)	34 (9%)	53 (8%)	43 (8%)
30–49	110 (19%)	115 (21%)	90 (24%)	97 (25%)	136 (21%)	111 (20%)
50–74	299 (51%)	262 (47%)	199 (52%)	202 (52%)	297 (45%)	278 (50%)
≥75	120 (20%)	117 (21%)	62 (16%)	57 (15%)	175 (26%)	123 (22%)
Sex						
Female	344 (58%)	343 (62%)	233 (61%)	241 (62%)	401 (61%)	327 (59%)
Male	234 (40%)	211 (38%)	147 (39%)	149 (38%)	260 (39%)	228 (41%)
Body mass index						
Underweight	13 (2%)	8 (2%)	5 (1%)	6 (2%)	13 (2%)	12 (2%)
Normal weight	194 (34%)	186 (36%)	88 (24%)	81 (21%)	202 (31%)	157 (29%)
Overweight	197 (35%)	168 (32%)	110 (29%)	139 (36%)	221 (34%)	188 (35%)
Obese	163 (29%)	156 (30%)	170 (46%)	157 (41%)	213 (33%)	179 (33%)
Missing	22	36	7	7	12	19

Table 2 from Miglioretti, JACR, 2014 showing a reduction in CTs performed after an individualized intervention strategy was in place.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Demb_and_Kumar-_2015_tables.pdf,NQF_Evidence_document_2015_10_12.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Radiologists and other physicians who perform CT in general are not aware of the doses they use, and there is tremendous variation in the doses they use even for patients seen for the same clinical indication. Even when clear standards around optimal doses exist, facilities do not routinely assess whether they use appropriate doses. For example, we conducted a 15 center randomized controlled trial of patients with suspected kidney stones seen in one of 15 U.S. emergency rooms (Smith-Bindman, NEJM, 2014),. The primary purpose of this study was to assess whether CT or ultrasound should be used as the first diagnostic test in these patients. As a secondary aim, we assessed the radiation doses of patients who received CT scans as part of this trial (Smith-Bindman, Jama IM). It is well established that patients with suspected kidney stones should undergo CT using a low dose, renal stone protocol CT, which delivers a dose of around 4 mSv or lower, as it is equally diagnostic to routine abdominal CT but uses around 1/3 the amount of radiation without any loss of diagnostic accuracy. (ACR, DIR, 2014) Nonetheless, when we assessed the doses that were actually used among the patients in our cohort, fewer than 10% of patients received low doses, the average dose was 12 mSv (three times higher) and some patients received doses as high as 75mSv (Smith-Bindman, JAMA, 2012; Smith-Bindman, Radiology, 2015). Of note, all of these patients were at high risk for stone disease, and at low risk for alternative diagnoses, and thus all should have received low dose examinations. These results closely paralleled the results of the American College of Radiology Dose Index Registry where the doses for Stone protocol CTs were assessed, and only 2% of exams used low dose. Of not, none of the participating sites in my 15 center trial were aware of their doses, and our quantification of these doses was the first step for facilities to try to optimize. If all doses were at the appropriate dose level, the doses would have been around 40% lower.

The lack of local practice assessment as highlighted by our STONE trial leads to dramatic practice variation that introduces unnecessary harm from excessive radiation dosing, and many publications have demonstrated profound variation in doses when a patient goes to different facilities to obtain a CT, or variation within institution when studies are obtained at different times of the day (ACR, DIR, 2014; Hausleiter, JAMA 2009; Keegan, JACR 2014; Miglioretti, JAMA Pediatrics, 2013; Parker, Pediatrics, 2015; Smith-Bindman, Arch Int Med, 2009; Smith-Bindman, JAMA, 2012; Smith-Bindman, JACR 2014).

In our JAMA Pediatrics paper (Miglioretti, JAMA IM 2013) using statistical modeling and observed CT doses, we modeled what would occur if the highest dose patients (those above the 75% benchmark) came down to the median dose. vThe dominant two indications for imaging in this cohort was imaging with CT for minor trauma and imaging with CT for appendicitis. Using current exposures, we would expect that due to CT exposures in children age 15 and younger in the US in 2010, 9,820 future cancers will occur. If the highest exposed individuals instead had doses at the median, 44% of these cancers would be prevented.

Furthermore, since information on radiation is reported differently across the different types of CT machines, and data are pooled in various ways, it is difficult for physicians to easily standardize their practice without a common and simple framework for doing so. Currently, physicians do not know the typical radiation doses received by their patients. This tool provides the framework for measurement – the first step towards quality improvement.

Creation of a simple standard for collection of radiation dose information would help facilities understand their current practice, would allow understanding changes in practice over time (Keegan, JACR, 2014; Greenwood, RadioGraphics, 2015) would allow comparisons to local and national standards, and would indicate to facilities whether their is a need to improve. There is currently a

high level of interest in this area - facilities are being asked by their patients and governing boards to report whether they are performing CT safely - and this measure is an ideal starting point for facilities to assemble this information to answer these questions. If facilities collect dose information, it is the first step towards trying to compete on a measure of safety and to lower the doses they use.

The measure will contribute to the creation of broadly applicable expected range, and UCSF and other professional organizations will contribute to their creation. This will lead to dose awareness and inevitable improvements as it will enable physicians to consider dose as an important measure.

We compared several methods of assessing doses as outlined in this measure, including automated and manual dose assessment. While automatic approaches have obvious advantages, it is feasible to collect these data manually with minimal time(Keegan, JACR, 2014).

Cited in this section:

American College of Radiology (ACR). Dose Index Registry (DIR). 2014.
http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/DIR%20Measures.pdf
Registry designed to showcase measures for certain CT procedure types.

Greenwood T, Lopez-Costa R, Rhoades P, et al. CT Dose Optimization in Pediatric Radiology: A Multiyear Effort to Preserve the Benefits of Imaging While Reducing the Risks. RadioGraphics. Jan 2015;35(5):1539-1554
"This systematic approach involving education, streamlining access to magnetic resonance imaging and ultrasonography, auditing with comparison with benchmarks, applying modern CT technology, and revising CT protocols has led to a more than twofold reduction in CT radiation exposure between 2005 and 2012…" – Conclusion statement from Abstract

Hausleiter, J., T. Meyer, et al. Estimated radiation dose associated with cardiac CT angiography. JAMA 301(5): 500-7. 2009
"Median doses of CCTA differ significantly between study sites and CT systems. Effective strategies to reduce radiation dose are available but some strategies are not frequently used. The comparable diagnostic image quality may support an increased use of dose-saving strategies in adequately selected patients."– Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014
Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311, 2013.
Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered.

Parker M, Shah S, Hall M, et al. Computed Tomography and Shifts to Alternate Imaging Modalities in Hospitalized Children. Pediatrics. 2015-0995.
"For the 10 most common All-Patient Refined Diagnosis Related Groups (APR-DRGs) for which children received CT in 2004, a decrease in CT utilization was found in 2012. Alternative imaging modalities for 8 of the diagnoses were used." – Conclusion statement from Abstract

Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA;307:2400-9. 2012
" Within integrated health care systems, there was a large increase in the rate of advanced diagnostic imaging and associated radiation exposure between 1996 and 2010." – Conclusion statement from Abstract

Smith-Bindman R, Aubin C, Bailitz J, et al. Ultrasonography versus Computed Tomography for Suspected Nephrolithiasis. N Engl J Med (NEJM); 371:1100-1110. 2014
"Initial ultrasonography was associated with lower cumulative radiation exposure than initial CT, without significant differences in high-risk diagnoses with complications, serious adverse events, pain scores, return emergency department visits, or hospitalizations." – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology; 277: 134–141. 2015
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med;169:2078-86. 2009
" Radiation doses from commonly performed diagnostic CT examinations are higher and more variable than generally quoted, highlighting the need for greater standardization across institutions." – Conclusion statement from Abstract

Lukasiewicz A, Bhargavan-Chatfield M, Coombs L, et al. Radiation Dose Index of Renal Colic Protocol CT Studies in the United States: A Report from the American College of Radiology National Radiology Data Registry. Radiology. May 2014;271(2):445-451.
"Reduced-dose renal protocol CT is used infrequently in the United States. Mean dose index is higher than reported previously, and institutional variation is substantial." – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Griffey RT, et al. Computed Tomography Radiation Dose in Patients With Suspected Urolithiasis. JAMA internal medicine. Aug 1 2015;175(8):1413-1416.

Smith-Bindman 2015, Predictors of Computed Tomography Radiation Dose and Their Impact on Patient Care. In Press, Radiology

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
N/A

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
Assessment of radiation doses as outlined in this measure was shown to reduce the radiation doses by 10-30%. The reductions were seen in average doses, and high doses using the techniques of assessing dose as outlined in this measure. Of note, the improvements in doses occurred primarily at sites or among technologists who had the greatest need for improvement (Demb, 2015; Keegan, JACR, 2014; Miglioretti, JACR, 2014; Wilson, ARRS, 2015)

See all the tables attached at the bottom of the Evidence Document, "Demb and Kumar, 2015" attached above.

Table from Keegan, JACR, 2014 shows a reduction in dose metrics after the prior NQF Measure was included. These tables (Table 2 from the paper) showed the summary of dose metrics in 2010 vs 2012 and their percent changes, which all lowered between 29-52% after the previous NQF Measure 0739 was endorsed.

Table 3 from Keegan, JACR, 2014 also shows the percentile reduction in dose metrics after the prior NQF Measure was included. These percentiles and their percent changes range from -30% to -47%.

Table 2 from Miglioretti, JACR, 2014 showing a reduction in CTs performed after an individualized intervention strategy was in place.These percent changes show that intervention facilities has higher reduction rates of the number of CTs performed, as opposed to a control facility.

Tables from Kumar, 2015 show the differences in DLP, CTDIvol, and SSDE regarding CT dosing and their confidence intervals between US and Non US sites.

Cited in this section:

American College of Radiology (ACR). Dose Index Registry (DIR). 2014.
http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/NRDR/DIR/DIR%20Measures.pdf
Registry designed to showcase measures for certain CT procedure types.

Calvert C, Strauss KJ, Mooney DP. Variation in computed tomography radiation dose in community hospitals. Journal of pediatric surgery. Jun 2012;47(6):1167-1169.
"Radiation exposure is a concern among those who evaluate injured children…This study identified a thirty-times range of radiation dosage for CT scans performed across 40 different hospitals." – Conclusion statement from Abstract

Demb J, manuscript under preparation. CT Radiation Dose Standardization Across the University of California Medical Centers Using Audits to Optimize Dose. 2015.
Following an in-person meeting regarding CT radiation dose, radiologists, technologists and medical physicists from University of California medical centers strategized how to best optimize dosing practices at their sites, which were then analyzed for effectiveness and success after implementation.

Dorfman AL, Fazel R, Einstein AJ, et al. Use of Medical Imaging Procedures With Ionizing Radiation in Children: A Population-Based Study. Arch Pediatr Adolesc Med. Jan 3 2011.
" Exposure to ionizing radiation from medical diagnostic imaging procedures may occur frequently among children. Efforts to optimize and ensure appropriate use of these procedures in the pediatric population should be encouraged." – Conclusion statement from Abstract

Duncan J, Street M, Strother M, et al. Optimizing Radiation Use During Fluoroscopic Procedures: A Quality and Safety Improvement Project. J Am Coll Radiol. 2013;10:847-853
"A systematic approach to improving radiation use during procedures led to a substantial and sustained reduction in risk with no reduction in benefits. Data were readily captured by both manual and automated processes." – Conclusion statement from Abstract

Einstein AJ, Henzlova MJ, Rajagopalan S. Estimating risk of cancer associated with radiation exposure from 64-slice computed tomography coronary angiography. JAMA 2007;298:317-23.
"… estimates derived from our simulation models suggest that use of 64-slice CTCA is associated with a nonnegligible LAR (lifetime attributable risk) of cancer. This risk varies markedly and is considerably greater for women, younger patients..."– Conclusion statement from Abstract

Greenwood T, Lopez-Costa R, Rhoades P, et al. CT Dose Optimization in Pediatric Radiology: A Multiyear Effort to Preserve the Benefits of Imaging While Reducing the Risks. RadioGraphics. Jan 2015;35(5):1539-1554
"This systematic approach involving education, streamlining access to magnetic resonance imaging and ultrasonography, auditing with comparison with benchmarks, applying modern CT technology, and revising CT protocols has led to a more than twofold reduction in CT radiation exposure between 2005 and 2012…" – Conclusion statement from Abstract

Hausleiter, J., T. Meyer, et al. Estimated radiation dose associated with cardiac CT angiography. JAMA 301(5): 500-7. 2009
"Median doses of CCTA differ significantly between study sites and CT systems. Effective strategies to reduce radiation dose are available but some strategies are not frequently used. The comparable diagnostic image quality may support an increased use of dose-saving strategies in adequately selected patients."– Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014
Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Kumar K, manuscript under preparation. Radiation Dose Benchmarks in Children.
This paper will describe dose metrics among 29,000 children within age strata <1, 1-4 years, 5-9 years, 10-14 years, and 15-19 years. 2015.

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311. 2013.
Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered.

Miglioretti DL, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. Journal of the American College of Radiology: JACR 2014.
"Personalized audit feedback and education can change technologists´ attitudes about, and awareness of, radiation and can lower

patient radiation exposure from CT imaging." – Conclusion statement from Abstract

Nationwide Evaluation of X-ray Trends: NEXT 2005-2006. This presentation was given by David Spelic, physicist with the Food and Drug Administration (FDA), to the 39th Conference of Radiation Control Program Directors (CRCPD) annual meeting, held in Spokane Washington, May 21-24, 2007.

Morin, R. L. (2006). CT dosimetry--an enigma surrounded by a conundrum. J Am Coll Radiol 3(8): 630.
An explanation of the difficulties surrounding CT dosing and estimations of its harmful effects.

Morin, R. L. (2006). What are the national radiation doses? J Am Coll Radiol 3(12): 956.
An explanation of why benchmarks or national measures are so difficult to set (related to the article listed above).

Parker M, Shah S, Hall M, et al. Computed Tomography and Shifts to Alternate Imaging Modalities in Hospitalized Children. Pediatrics. 2015-0995.
"For the 10 most common All-Patient Refined Diagnosis Related Groups (APR-DRGs) for which children received CT in 2004, a decrease in CT utilization was found in 2012. Alternative imaging modalities for 8 of the diagnoses were used." – Conclusion statement from Abstract

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med 2009;169:2078-86.
" Radiation doses from commonly performed diagnostic CT examinations are higher and more variable than generally quoted, highlighting the need for greater standardization across institutions." – Conclusion statement from Abstract

Smith-Bindman R. Is computed tomography safe? N Engl J Med 2010;363:1-4.
An explanation of the harmful effects of CT overdose, and why its diagnostic purposes are often misused.

Smith-Bindman R. Environmental causes of breast cancer and radiation from medical imaging: findings from the Institute of Medicine report. Arch Intern Med 2012;172:1023-7.
" The IOM´s conclusion of a causal relation between radiation exposure and cancer is consistent with a large and varied literature showing that exposure to radiation in the same range as used for computed tomography will increase the risk of cancer." – Conclusion statement from Abstract

Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012;307:2400-9.
" Within integrated health care systems, there was a large increase in the rate of advanced diagnostic imaging and associated radiation exposure between 1996 and 2010." – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology 2015:277: 134–141
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Wilson N. CT Radiation Dose Standardization Across the Five University of California Medical Centers. ARRS: Annual Toronto Meeting presentation. April 19-24, 2015
Understanding the reasons for variation in commonly performed CT procedures, and figuring out how to standardize them.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
There are two principle areas on known, or suspected, disparity. The first involves children as compared with adults. In general, it is believed that the harm of radiation with respect to the potential to cause future cancer is proportional to the radiation per unit of tissue, as well as the age at exposure. Because children are smaller than adults, and because doses have not been reliably reduced in children, the same radiation dose will be more harmful in children because of their smaller size (ie greater radiation per unit tissue) and also because children are more radio sensitive. This has been known for many years (Brenner, NEJM 2007) and in fact the FDA issued a warning in 2012 asking physicians to lower the doses they use in children; however, there is no evidence that this has been widely done. We found profound variation in doses used in children (Miglioretti, JAMA Pediatrics, 2013), and a related abstract found

that while exams in pediatric hospitals tend to use lower dose technique, doses used on children in adult hospitals (where most CT scans in children occur) are not tailored. Image Gently, a large social marketing campaign, has focused attention on this issue, but awareness and agreement are only the first two steps in Pathman et al's model of clinical guideline compliance. Regrettably, the Image Gently campaign lacks processes that assess adoption and adherence.

The second potential area of disparity has to do with socioeconomic status. In general, newer technologies of CT allow dose to be reduced, and public hospitals are less likely to have these newer machines. To support this hypothesis, we assessed the CT radiation doses in the STONE trial, and stratified the results by whether the hospital was a county hospital that provides care to the under-served. The country hospitals had doses that were higher than non county hospitals (significant in univariate analysis and boder line significant in multivariate analysis) and on average delivered doses of radiation for routine CT that were many times times higher than the best performing hospitals in the sample. Other studies have shown that pediatric specialty centers also have different care standards than community hospitals (Parker, Pediatrics, 2015). Thus this observation merits further study.

The last area of disparity has to do with differences in the care provided by pediatric specialty centers and children's hospitals (Agarwal et al, AJR in Press)

There are no additional data describing demographic, or racial or ethnic, or insurance or SES disparity

Cited in this section:

Brenner D, Hall E. Computed Tomography — An Increasing Source of Radiation Exposure. NEJM. 357:2277-2284. 2007
Study showing that the marked increase in CT scans across the US has increased the radiation exposure of the general public.

Parker M, Shah S, Hall M, et al. Computed Tomography and Shifts to Alternate Imaging Modalities in Hospitalized Children. Pediatrics. 2015-0995.
"For the 10 most common All-Patient Refined Diagnosis Related Groups (APR-DRGs) for which children received CT in 2004, a decrease in CT utilization was found in 2012. Alternative imaging modalities for 8 of the diagnoses were used." – Conclusion statement from Abstract

Pathman D, Konrad T, Freed G, et al. The Awareness-to-Adherence Model of the Steps to Clinical Guideline Compliance: The Case of Pediatric Vaccine Recommendations. Medical Care 34: 873-889 Sept 1996
Showcase of the model that should be followed in order to help compliance for awareness and adherence to our measure.

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, et al. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311. 2013.
Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Other
1c.2. **If Other:** In the last ten years, children are receiving fivefold more abdominal CTs and 50% more head CTs. At these rates, 1 in 3 children will undergo at least one CT scan before their 18th birthday and this is only start since rates of CT scans tend to increase with age. The number of individuals exposed to radiation from medical imaging is extremely high and has been increasing steadily for the past 20 years (Smith-Bindman, JAMA, 2012; Miglioretti, JAMA Pediatrics, 2013.) In the last ten years, children are receiving fivefold more abdominal CTs and 50% more head CTs. At these rates, 1 in 3 children will undergo at least one CT scan before their

18th birthday and this is only start since rates of CT scans tend to increase with age (Miglioretti, JAMA Pediatrics, 2013). Such numbers show a marked increase in the lifetime attributable risk of developing cancers related to this extra radiation, especially for children with chronic diseases. Overall, in the US, the exposure to radiation has increased 600% in the last 20 years, and the average annual exposure to radiation from all sources has doubled due to the radiation from medical imaging (NCRP, 2007) Thus this is a high priority issue due to the numbers of individuals involved. Although the analysis was based on adults, and not children, the Institute of Medicine, in their recent report on environmental causes of breast cancer, concluded that avoiding unnecessary exposure to CT was the single most important step that women could take to reduce their risk of breast cancer (Smith-Bindman, Arch Intern Med, 2012.) In 2009, 10% of patients underwent a CT annually, and thus the number of people who are be impacted by the quality and safety of CT is extremely high. (Smith-Bindman, JAMA, 2012) Cited in this section: Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311. 2013. Radiation-induced cancers in children could be dramatically reduced if the highest quartile of CT radiation doses were lowered. National Council on Radiation Protection (NCRP). Report No. 157 - Radiation Protection in Educational Institutions. http://www.ncrppublications.org/Reports/157. 2007 "The purpose of this Report is to provide guidance for the safe use of ionizing- and nonionizing-radiation sources in educational institutions, including both teaching and research activities." – Introduction statement from Abstract Smith-Bindman R. Environmental causes of breast cancer and radiation from medical imaging: findings from the Institute of Medicine report. Arch Intern Med 2012;172:1023-7. " The IOM´s conclusion of a causal relation between radiation exposure and cancer is consistent with a large and varied literature showing that exposure to radiation in the same range as used for computed tomography will increase the risk of cancer." – Conclusion statement from Abstract Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012;307:2400-9. " Within integrated health care systems, there was a large increase in the rate of advanced diagnostic imaging and associated radiation exposure between 1996 and 2010." – Conclusion statement from Abstract Smith-Bindman R. Environmental causes of breast cancer and radiation from medical imaging: findings from the Institute of Medicine report. Arch Intern Med. Jul 9 2012;172(13):1023-1027

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

The use of diagnostic imaging has increased dramatically over the past decade, contributing to medical exposure to ionizing radiation. The largest growth has been in the utilization of computed tomography (CT). The total number of CT examinations performed annually in the United States has risen from approximately 3 million in 1980 to nearly 80 million today, and this is 6 fold higher than in 1980. Integrating CT into routine care has improved patient health care. However, CT delivers much higher radiation doses than do conventional diagnostic x-rays. For example, a chest CT typically delivers 500- 1000 times the radiation dose of chest x-ray. Further, radiation exposure from individual CT examinations has also increased, in part due to the increased speed of image acquisition allowing vascular, cardiac, and multiphase examinations, all associated with higher doses. Thus, greater utilization of CT and higher exposure per examination has resulted in a substantial increase in the US population's exposure to radiation from medical imaging. The National Counsel on Radiation Protection reported that the US population's exposure to radiation from medical imaging increased 600 fold over the last 20 years.

Further, recent research conducted by our group has documented significant variation in the radiation doses associated with specific CT examinations, between facilities and patients, raising concerns that the doses may be higher than necessary and potentially un-safe. Further several egregious errors in the use of CT and its associated radiation dose– identified in several California hospitals including Cedar's Sinai and in Huntsville, Alabama where doses were delivered that were as high as radiation used to treat brain cancer – further highlighted concerns about the radiation doses that can be delivered (either deliberately or accidentally through CT) can be extremely high. These errors led to levels of radiation exposure comparable to those delivered by radiation therapy for brain cancer

Exposure to ionizing radiation is of concern, because extensive evidence has linked exposure to ionizing radiation at doses used in medical imaging to the development of cancer. While there are some uncertainties in the exact quantification of risk, the overwhelmingly supported view is that it is prudent to limit radiation to the degree possible.

Recognizing the potential risks associated with CT, The FDA has announced plans to increase their oversight of radiation from CT – including their call that for facilities to begin to assess the radiation used in examinations, and call for creation of diagnostic reference levels. The US House of Representatives, Energy and Commerce Committee, Subcommittee on Health has sponsored hearings specifically focused on radiation associated with medical imaging, with discussion of possible legislative oversight. The Joint Commission has issued a radiation sentinel event and has incorporated assessment of the radiation dose as part of its hospital accreditation metrics beginning in 2015.

The measure as specified could enhance all of the efforts by promoting a simple measurement tool and standard.

Of note, two of the measurements that are specified, CTDIvol, and DLP have been extensively tested and validated. CTDIvol can be most easily understood as the average machine output within a short scan area. The DLP multiples the CTDIvol by the scan length to get an estimate of the total irradiation output. SSDE is an adjusted measure that tries to account for the appropriateness of dose with respect to patient size. A larger dose is needed for a larger patient. SSDE essentially scales the CTDIvol by patient size. Theoretically, SSDE will show greater stability across different patient size groups if the relative dosing amount is appropriately scaled for size. I say theoretically as we have not found it to be more stable across different weight categories. It was developed by a collaboration of medical physicists to scale doses in the abdomen, but has not been validated as a facility-level measure of dose. However, including SSDE has enhanced enthusiasm for this measure and thus this application for renewal the metric has been added.

This metric was created by physicists and the American Association of Physicists in Medicine (AAPM (American Association of Physicists in Medicine). AAPM Report No. 204 - Size-specific dose estimates (SSDE) in pediatric and adult body CT examinations (American Association of Physicists in Medicine;2011).

While it has not undergone rigorous testing, there is widespread interest in this measure, particularly in children, diagnostic reference ranges have been generated in children using this metric (Goske MJ, Strauss KJ, Coombs LP, et al. Diagnostic reference ranges for pediatric abdominal CT. Radiology. Jul 2013;268(1):208-218).

We have found it yields similar results to the other metrics (Keegan et al, JACR 2014), and importantly, this metric has broad stakeholder support.

However, sites can choose whatever metric works best for them: CTDIvol, DLP, SSDE, or other measures endorsed by the AAPM.

1c.4. Citations for data demonstrating high priority provided in 1a.3

Board of Radiation Effects Research Division on Earth and Life Sciences "Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2 Washington, D.C.: The National Academies Press, 2006.

Bogdanich, Walt. At hearing on radiation, calls for better oversight. NY Times. February 26, 2010
Article talking about the trials requiring more oversight regarding radiation safety.

Calvert C, Strauss KJ, Mooney DP. Variation in computed tomography radiation dose in community hospitals. Journal of pediatric surgery. Jun 2012;47(6):1167-1169.
“Radiation exposure is a concern among those who evaluate injured children…This study identified a thirty-times range of radiation dosage for CT scans performed across 40 different hospitals.” – Conclusion statement from Abstract

Caoili, E. M., R. H. Cohan, et al. (2009). Medical decision making regarding computed tomographic radiation dose and associated risk: the patient´s perspective. Arch Intern Med 169(11): 1069-71.
“ The study group’s overall knowledge of radiation risk was poor, but we did not find significant differences between Hispanic vs. non-Hispanic patients.” – Conclusion statement from Abstract

Demb J, manuscript under preparation. CT Radiation Dose Standardization Across the University of California Medical Centers Using Audits to Optimize Dose. 2015.
Following an in-person meeting regarding CT radiation dose, radiologists, technologists and medical physicists from University of California medical centers strategized how to best optimize dosing practices at their sites, which were then analyzed for effectiveness and success after implementation.

Dorfman AL, Fazel R, Einstein AJ, et al. Use of Medical Imaging Procedures With Ionizing Radiation in Children: A Population-Based Study. Arch Pediatr Adolesc Med. Jan 3 2011.

Duncan J, Street M, Strother M, et al. Optimizing Radiation Use During Fluoroscopic Procedures: A Quality and Safety Improvement Project. J Am Coll Radiol. 2013;10:847-853
“A systematic approach to improving radiation use during procedures led to a substantial and sustained reduction in risk with no reduction in benefits. Data were readily captured by both manual and automated processes.” – Conclusion statement from Abstract

Einstein AJ, Henzlova MJ, et al. Estimating risk of cancer associated with radiation exposure from 64-slice computed tomography coronary angiography. JAMA. Jul 18 2007;298(3):317-323.

"… estimates derived from our simulation models suggest that use of 64-slice CTCA is associated with a nonnegligible LAR (lifetime attributable risk) of cancer. This risk varies markedly and is considerably greater for women, younger patients..."– Conclusion statement from Abstract

Fletcher JG, Kofler JM, Coburn JA, Bruining DH, McCollough CH. Perspective on radiation risk in CT imaging. Abdom Imaging. Feb 2013;38(1):22-31.
"The benefits and risks of CT are also highly individualized, and require consideration of many factors by patients, clinicians, and radiologists." – Conclusion statement from Abstract

Food and Drug Administration. FDA Makes Interim Recommendations to Address Concern of Excess Radiation Exposure during CT Perfusion Imaging. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm193190.htm. 2009
Recommendations made by the FDA to prevent excess radiation in patients.

Goske MJ, Strauss KJ, Coombs LP, et al. Diagnostic reference ranges for pediatric abdominal CT. Radiology. Jul 2013;268(1):208-218.
"Calculation of reference doses as a function of BW (body weight) for an individual practice provides a tool to help develop site-specific CT protocols that help manage pediatric patient radiation doses." – Conclusion statement from Abstract

Greenwood T, Lopez-Costa R, Rhoades P, et al. CT Dose Optimization in Pediatric Radiology: A Multiyear Effort to Preserve the Benefits of Imaging While Reducing the Risks. RadioGraphics. Jan 2015;35(5):1539-1554
"This systematic approach involving education, streamlining access to magnetic resonance imaging and ultrasonography, auditing with comparison with benchmarks, applying modern CT technology, and revising CT protocols has led to a more than twofold reduction in CT radiation exposure between 2005 and 2012…" – Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014
Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Kumar K, manuscript under preparation. Radiation Dose Benchmarks in Children.
This paper will describe dose metrics among 29,000 children within age strata <1, 1-4 years, 5-9 years, 10-14 years, and 15-19 years. 2015.

Mathews J, Forsythe A, Brady Z, et al. Cancer risk in 680,000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians. BMJ. 2013;346 doi: http://dx.doi.org/10.1136/bmj.f2360
"Future CT scans should be limited to situations where there is a definite clinical indication, with every scan optimised to provide a diagnostic CT image at the lowest possible radiation dose." – Conclusion statement from Abstract

McBride, D. Radiation may be unnecessary for children with leukemia. ONS Connect 24(10): 29. 2009
"This study suggests that a better result can be attained with lesser long term effects on ALL patients who did not undergo prophylactic cranial irradiation treatment." – Conclusion statement from Abstract

McBride J, Paxton BE, et al. American Roentgen Ray Society, Annual Meeting in Boston, MA, April 26-30. CT Scans: Most Doctors Lack Knowledge of Radiation Exposure Risks. 2009

McCollough C, Branham T, Herlihy V, et al. Diagnostic reference levels from the ACR CT Accreditation Program. Journal of the American College of Radiology : JACR. Nov 2011;8(11):795-803.
"Effective January 1, 2008, the ACR program implemented United States-specific diagnostic reference levels of 75, 25, and 20 mGy, respectively, for the CTDI(vol) of routine adult head, adult abdominal, and pediatric abdominal CT scans." – Conclusion statement from Abstract

Medical Radiation: An Overview of the Issues: US House of Representatives, Energy an Commerce Committee Hearing - Subcommittee on Health, Friday, 26 Februrary 2010. Testimony of witnesses and discussion can be found on web site: http://energycommerce.house.gov/index.php?option=com_content&view=article&id=1910:medical-radiation-an-overview-of-the-issues&catid=132:subcommittee-on-health&Itemid=72

Medicine ABoI. U.S. Physician Groups Identify Commonly Used Tests or Procedures They Say are Often Not Necessary. 2012;

http://www.abimfoundation.org/News/ABIM-Foundation-News/2012/Choosing-Wisely.aspx. Accessed Last accessed on: 5/15/2013, 2012.
A list of commonly used procedures that may not be necessary for patient safety.

Mettler, FA Jr. "Overview of Medical Usage Patterns Radiation Exposures from Imaging and Image Guided Interventions." Eighth Annual Gilbert W. Beebe Symposium. Wednesday, December 9, 2009, The National Academies Washington, D.C. 2009

Mettler, FA Jr Thomadsen BR, et al. Medical radiation exposure in the U.S. in 2006: preliminary results. Health Phys; 95(5):502-507. Nov 2008

Mettler FA Jr., Huda W, et al. "Effective doses in radiology and diagnostic nuclear medicine: a catalog. Radiology. Jul 2008;248(1):254-263. ."

Mettler, F. A., Jr., B. R. Thomadsen, et al. (2008). "Medical radiation exposure in the U.S. in 2006: preliminary results." Health Phys 95(5): 502-7.

Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA pediatrics. Aug 1 2013;167(8):700-707.

Miglioretti DL, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. Journal of the American College of Radiology: JACR 2014.
"Personalized audit feedback and education can change technologists´ attitudes about, and awareness of, radiation and can lower patient radiation exposure from CT imaging." – Conclusion statement from Abstract

National Council on Radiation Protection and Measurements "NCRP Report No 160, Ionizing Radiation Exposure of the Population of the United States.

Ozasa K, Shimizu Y, Suyama A, et al. Studies of the mortality of atomic bomb survivors, Report 14, 1950-2003: an overview of cancer and noncancer diseases. Radiat Res. Mar 2012;177(3):229-243.

Parker M, Shah S, Hall M, et al. Computed Tomography and Shifts to Alternate Imaging Modalities in Hospitalized Children. Pediatrics. 2015-0995.
"For the 10 most common All-Patient Refined Diagnosis Related Groups (APR-DRGs) for which children received CT in 2004, a decrease in CT utilization was found in 2012. Alternative imaging modalities for 8 of the diagnoses were used." – Conclusion statement from Abstract

Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet;380(9840):499-505. Aug 4 2012

Schindera ST, Odedra D, Raza SA, et al. Iterative Reconstruction Algorithm for CT: Can Radiation Dose Be Decreased while Low-Contrast Detectability Is Preserved? Radiology. Jun 20 2013.

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med 2009;169:2078-86.
" Radiation doses from commonly performed diagnostic CT examinations are higher and more variable than generally quoted, highlighting the need for greater standardization across institutions." – Conclusion statement from Abstract

Smith-Bindman, R., D. L. Miglioretti, et al. (2008). "Rising use of diagnostic medical imaging in a large integrated health system." Health Aff (Millwood) 27(6): 1491-502.

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology 2015:277: 134–141
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Steenhuysen, J. (Februrary 26, 2010). "US experts seek more overgsight of medical radiation: Equipment makes need national dose standards." Retrieved March 31, 2010, from http://www.reuters.com/article/idUSN2610991020100226.

Wilson N. CT Radiation Dose Standardization Across the Five University of California Medical Centers. ARRS: Annual Toronto Meeting presentation. April 19-24, 2015
Understanding the reasons for variation in commonly performed CT procedures, and figuring out how to standardize them.

Zablotska LB, Bazyka D, Lubin JH, et al. Radiation and the risk of chronic lymphocytic and other leukemias among Chernobyl cleanup workers. Environ Health Perspect. Jan 2013;121(1):59-65.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cancer, Prevention

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse, Prevention, Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
N/A

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
For a different but the very similar measure #0739: Aug 15, 2011 Most Recent Endorsement Date: Aug 15, 2011

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Radiation Dose metrics among consecutive patients, who have undergone CT of the head, chest, abdomen/pelvis, or chest/abdomen/pelvis. The metrics are 1) mean dose as measured using DLP, CTDIvol, and SSDE: within age strata. And 2) the proportion of exams with doses greater than the 75th percentile of the benchmark you are comparing with for the same anatomic area strata (Kumar, 2015; Smith-Bindman, Radiology, 2015; Goske, Radiology, 2013)

The CTDIvol and DLP are directly reported by the scanner using an "industry wide" standardized dose report (DICOM Radiation Dose Structured Report). The data should be assembled for the entire CT examination. If there are several series, the CTDIvol values should be averaged, and the DLP values should be added.

SSDE can be calculated using any dose monitoring software product, or using published multiplier coefficients which are highly valid.

These different metrics are highly correlated, but nonetheless reveal important differences regarding radiology practice and performance and are thus complimentary. However, if a practice only assesses data from a single metric, there is substantial opportunity for data-driven improvement.

CTDIvol reflects the average dose per small scan length. Modern CT scanners directly generate this.

DLP reflects the CTDIvol x scan length, and is directly generated by modern CT scanners.

SSDE is a modified measure of CTDIvol that takes into account the size of the patient scanned and is useful for scaling dose to patient size. Several current radiation tracking software tools directly report SSDE.

Cited in this section

Goske MJ, Strauss KJ, Coombs LP, et al. Diagnostic reference ranges for pediatric abdominal CT. Radiology. Jul 2013;268(1):208-218.
"Calculation of reference doses as a function of BW (body weight) for an individual practice provides a tool to help develop site-specific CT protocols that help manage pediatric patient radiation doses." – Conclusion statement from Abstract

Kumar K, manuscript under preparation. Radiation Dose Benchmarks in Children.
This paper will describe dose metrics among 29,000 children within age strata <1, 1-4 years, 5-9 years, 10-14 years, and 15-19 years. 2015.

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology 2015:277: 134–141
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

Smith-Bindman R, Miglioretti DL. CTDIvol, DLP, and Effective Dose are excellent measures for use in CT quality improvement. Radiology. Dec 2011;261(3):999; author reply 999-1000.
An explanation as to why these radiation dose metrics are useful in calculating a patient's absorbed doses.

Huda W, Ogden KM, Khorasani MR. Converting dose-length product to effective dose at CT. Radiology. Sep 2008;248(3):995-1003.
"This article describes a method of providing CT users with a practical and reliable estimate of adult patient EDs by using the DLP displayed on the CT console at the end of any given examination." – Conclusion statement from Abstract

S.5. Time Period for Data *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The metric is based on cross sectional analyses, and the numerator and denominator have the same time period. The length of time needed to accrue a sufficient number of CT scans to generate sufficient precision will vary by the size of the facility, but for average sized practices, will include review of data from several months. The sample size to generate sufficient precision in each category is 25 CTs within each anatomic and age stratum. More than this number can be included for example if data are automatically generated, they can be generated for a fixed time interval (Keegan JACR 2014, Miglioretti JACR 2014). The sample sizes can be lower if the facilities do not evaluate sufficient children within a year to meet this minimum per strata. Of note, facilities do not need to collate data in all categories, only ones relevant to their practice. Further, if facilities scan children infrequently, they can combine across all age groups and use as their comparison benchmarks that have been published across all age categories [Smith-Bindman, Radiology, 2015 attached in Evidence Document]

Table 2 from Smith-Bindman, Radiology, 2015 shows different the dose benchmarks in children, and where the hospitals within that paper fell, within those percentiles. As can be seen, far more fall into the highest 75th percentile of dosing and CT procedures performed on children, showing the necessity of a measure such as this to be place into effect as soon as possible. Far too many scans, with far too high of a dose, are being performed on children.

All of the data are stored with the CT images and stored electronic data (within DICOM headers or as computer readable structured dose reports) and the dose data can be collected retrospectively for all patients at one time by reviewing existing records. Thus all of the data can be abstracted in a single time period of review.

Cited in this section:

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014
Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Miglioretti DL, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. Journal of the American College of Radiology: JACR 2014.
"Personalized audit feedback and education can change technologists´ attitudes about, and awareness of, radiation and can lower patient radiation exposure from CT imaging." – Conclusion statement from Abstract

Smith-Bindman R, Moghadassi M, Wilson N, et al. Radiation Doses in Consecutive CT Examinations from Five University of California Centers. Radiology 2015:277: 134–141
"These summary dose data provide a starting point for institutional evaluation of CT radiation doses." – Conclusion statement from Abstract

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Radiation dose distribution for the three metrics (CTDIvol, DLP, and SSDE) need to be recorded for a consecutive sample of CT examinations within anatomic area and age stratum. The mean, median, and the percent of examinations above the published 75% percentile needs to be generated.

These data can be extracted from the CT examinations in several ways. These numbers can written down directly from the CT scanner itself at the time of the examination; they can be written down from the PACS (computer terminal where images are reviewed and stored); or can be written down from the medical record if the facility stores these data as part of the medical record (all facilities in California due this based on statutory requirements.) The CT manufacturers have agreed (through MITA, Medical Imaging and Technology Alliance, the professional trade association of imaging manufacturers) to make these data electronically available through export from the CT machines to a local server), and these data can also be collected electronically. A growing number of companies are leveraging the standardized data format to systematically collect dose metrics directly from a facilities imaging infrastructure. This not only improves the accuracy of the data but also markedly reduces the costs of data collection. From the PACS, Radiology Information System, EPIC program if the data are exported there, or using any number of dose monitoring software programs allowing the collection and reporting of these dose data. The easiest way to collect these data is through one of the 6 or so commercial software programs developed for dose tracking, and several free-ware programs that enable directly extracting CT dose information from the PACS. We have published (Keegan, JACR 2014) several examples of techniques for dose extraction that can be completed even by a small facility.

The strata for this measure include:

Anatomic area strata: head, chest, abdomen/pelvis, Chest/abdomen/pelvis

Age strata: infant (<1); small child (1-5); medium child (>5 - 10); large child (>10-15) and adult (>15)

NOTE: The SSDE was developed as a metric for adjusting for size. However, it does not completely adjust for size and analysis within age strata are still needed among children to account for the different doses that are used and should be used for infants to obese children.

Cited in this section:

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014
Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Consecutive sample of CTs conducted in the head, chest, abdomen/pelvis and chest/abdomen/pelvis. No examinations should be excluded

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Consecutive sample of CTs conducted in the head, chest, abdomen/pelvis, chest/abdomen/pelvis

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
CT examinations conducted in anatomic areas not included above (such as CTs of the extremities or lumbar spine) or that combine several areas (head and chest) should not be included. In children, these four included categories will reflect approximately 80% of CT scans.

Examinations performed as part of diagnostic procedures – such as biopsy procedures – should not be included. CT examinations performed as part of surgical planning or radiation therapy should not be included.

Examinations that are considered "limited abdomen" or "limited pelvis" studies should be included in the abdomen and pelvis category. Any examinations that include any parts of the abdomen and or pelvis should count in the abdomen/pelvis category.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Most abdominal/pelvis CT scans in adult patients include scanning of the abdomen and pelvis as one contiguous area. If examinations are conducted limited to one region, these should also be included, as it is difficult/impossible to define what areas would be considered limited.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Anatomic area strata: head, chest, abdomen/pelvis, chest/abdomen/pelvis

These were chosen based on being the most common CT examination types conducted in the US, comprising >80% of all CT scans, and because dose varies by these groups.

Age strata: infant (<1); small child (1-5); medium child (>5 - 10); large child (>10-15) and adult (>15)

These patient age groups were chosen based on the variation of CT settings and resulting radiation dose based on patient size (and age is frequently used as a surrogate for size.) The ICRU (International Commission on Radiation Units and Measurements) uses these child size categories, they correspond to available phantoms, and they are the ones found to be most reliable

Geographic location where studies were done (zip code or state), to facilitate using the data to create geographically specific benchmarks

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**

If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
N/A

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
For child categories, 25 patients within each strata will provide adequate sample size. (One year of data should be extracted if the minimum cannot be met within a shorter time interval).

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
The dose metrics are occasionally not available for a particular scan. These can be deleted from the numerator and denominator, although should be exceedingly rare (< 1/1000 examinations).

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data sources will include electronic CT images [captured from the CT console at the time of scanning or harvested from the

PACS (Picture Archiving Communication System) - the computerized systems for reviewing and storing imaging data], Radiology Information System, EPIC, printed CT images, or information stored in the medical record. Numerous other software products are now available for capturing these data (Bayer, GE, etc.) and several free ware programs are also available. Of note, the 2012 California law now requires the reporting of several of the dose metrics outlined in this measure in the patient medical record, and as a results, many software companies have provided techniques for collating these data.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Health Plan, Integrated Delivery System

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Ambulatory Surgery Center (ASC), Ambulatory Care : Clinician Office/Clinic, Ambulatory Care : Outpatient Rehabilitation, Ambulatory Care : Urgent Care, Hospital/Acute Care Facility, Imaging Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_testing_attachment_2015_10_14.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Pediatric CT Safety Measure
Date of Submission: 9/28/2015
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☑Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for <u>all</u> the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐☑clinical database/registry	☐ ☑clinical database/registry
☐ ☑abstracted from electronic health record	☐ ☑abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Several datasets have been used for testing of the measure including data from Individual single institutions, collaborations of institutions, integrated health care systems, electronic medical records, data extracted from stored CT images (captured from the CT console at the time of scanning or harvested from the PACS (Picture Archiving Communication System - the computerized systems for reviewing and storing imaging data), printed CT images, or information stored in the medical record. *Smith-Bindman (JAMA Internal Medicine 2009, JAMA 2012), Miglioretti (JAMA Pediatrics 2013; JACR 2014) and Keegan (JACR 2014)* use various methods of data abstraction. Two manuscripts abstracting and summarizing dose using the NQF endorsed metric are included.

1.3. What are the dates of the data used in testing? January 1, 2008 – December 31 , 2013

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.26*)**	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☑ hospital/facility/agency	☑hospital/facility/agency
☑health plan	☑health plan
☑other: integrated delivery system	☑other: integrated delivery system

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The measure has been tested in several settings: Group Health Research Institute, a large integrated Health System in the Pacific Northwest. CT examinations on over 10,000 examinations have been assembled and included in several publications *(Miglioretti, JACR 2014; Miglioretti JAMA Pediatrics 2013)*

The measure was tested in a consortium of integrated health care systems (n=6) and data were assembled for over 5000 CT examinations, and were published *(Smith-Bindman JAMA 2012)*

The measure was tested across the five University of California Medical Center, including over 100,000 CT examinations. The data has in part been published *(Keegan, JACR 2014)* and additional manuscripts were presented at national meetings (RSNA 2012) and are in preparation. A manuscript is in press describing the results of assessment of dose using this measure using data from across the University of California (In press, Radiology) and a second paper is under preparation demonstrating a 10-30% reduction in dose using a before and after design using assessment as specified in this measure.

For all, analyses were done using consecutive sample of CT examinations within anatomic area, age and machine type strata as specified in this measure, or using a randomly selected subset of examinations and analyzed per measure specifications.

A quality improvement activity assembling data per the NQF specifications was approved by the Board of the American College of Radiology for PQRS credit.

Miglioretti D, Johnson E, Vanneman N, Smith-Bindman R, al e. Use of Computed Tomography and Associated Radiation Exposure and Leukemia Risk in Children and Young Adults across Seven Integrated Healthcare Systems from 1994 – 2010. JAMA Pediatrics Published online June 10, 2013 joli:101001/jamapediatrics2013311 2013.

Miglioretti, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. In press Journal of the American College of Radiology 2014.

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med 2009;169:2078-86.

Smith-Bindman R. Is computed tomography safe? N Engl J Med 2010;363:1-4.

Smith-Bindman R. Environmental causes of breast cancer and radiation from medical imaging: findings from the Institute of Medicine report. Arch Intern Med 2012;172:1023-7.

Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012;307:2400-9.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
The summary of CT dose has been done in children and adults, and using consecutive scans without exclusion (ie scans were not excluded on any individuals) and analyzed within strata. Because the measure is specified at the institutional level, there is no reason to exclude any individuals. While there are individual patients who will and should have doses above averages, the measure calls for assessment of institutional data, and individual patients will have a small impact, if any, on overall calculations.

The strata for this measure include:

- Anatomic area strata: head, chest, abdomen/pelvis. These anatomic areas reflect approximately 85% of CT examination types in adults, and approximately 75% of CT examination types in children

- Age strata: infant (<1); small child (1-5); medium child (>5 - 10); large child (>10-15) and adult (>15)
- CT machine (manufacturer, type)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

There are no differences in the data or sample used for different aspects of testing.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

These were not available, nor tested

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☑ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The Proposed CT Dose measure calls for the collection of several metrics reflecting CT dose indices including DLP, CTDIvol, and SSDE. CTDI and DLP are calculated automatically by all current CT scanners, without variability. When these data are manually extracted, there is possibility of errors of writing down the values and has been found to be in the ballpark of a 5-10% error related to transcription, not calculation (Keegan JACR). SSDE is a calculated variable, and while dose monitoring programs automatically calculate this variable, sites that choose to calculate this manually will likely introduce errors, although this has not been quantified.

CTDIvol and DLP measures have been widely used for over a decade in several other countries, are used for in a bill that is in effect in California and I have personal and recent experience collecting these dose Indices across 12 large institutions reflecting dozens of machines and thousands of patients Reliability of CT radiation dose metric abstraction (DLP and CTDIvol) was tested by our group in several ways. First, manual data abstraction of data recorded from the PACS system was repeated in two large samples (one at Group Health, and one at UCSF) where the data was abstracted by a single observer, yielding highly reliable measures between abstractions (i.e. the measures were concordant, nearly perfect Kappa statistics, with a 5% variation in our analysis.) Second, data were extracted via commercial software product using two different tools for extracting the data from the stored CT files in PACS, and these were reviewed by a medical physicist to ensure the data were correct. This was performed at five separate institutions, and found the electronically captured data was identical to the manual review, perfect Kappa statistics.

SSDE is a relatively new metric that tries to take into account patient size. In our published work *(Keegan JACR 2014)* it tracks in parallel to the other metrics, but has not undergone formal reliability testing in large cohorts of patients.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Highly reliable, Kappas > 95%

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Highly reliable, Kappas > 95%

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☑**Empirical validity testing**

☐ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The three dose parameters (metrics) included in this measure reflects slightly different aspects of dose, and each was included because it provides a unique reflection of dose and can be used to improve quality and safety.

These dose parameters specified in this measure primarily reflect the dose that comes out of the machine and the dose that the patient is exposed to and *dictate the absorbed organ doses to the patient*. Absorbed doses (these are the doses a patient actually receives) will vary by sex and weight, but are *primarily* determined by the doses that come out of the machine. These dose metrics are highly correlated with the doses patients receive; higher DLPs, CTDIs, and SSDE are associated with higher absorbed dose to the patient's organs and higher patient detriment (harm). If these doses were lowered patients would be exposed to lower doses of radiation, have correspondingly lower absorbed organ doses and would be expected to have less detriment from these exposures to radiation. While patient absorbed doses are important, they are difficult to quantify.

However, the dose parameters themselves are vitally important as they 1) closely reflect organ doses and 2) are precisely those measurements that the technologist and physician can influence to lower doses. That is why these metrics were chosen for this measure. Estimating absorbed organ doses might be a more precise way to compare doses between two examinations on two patients. However, this is simply not practical. It is much more complicated to estimate these parameters, there are over 30 different organs where these doses can be compared and it does not make sense to measure because the technologist cannot directly influence these measures, and there would be practical way to compare facilities as there ware so many organ doses to compare. Using organ dose might add a very small amount more precision for an estimate of an individual

patient, but it's not clear that it's relevant or possible to measure and compare at the facility level. Thus organ dose was not proposed as a practical or useful metric for patient safety assessment.

The output of radiation from the machine is far simpler to measure and in fact is the important variable, as this is what the radiologist and the technologist can influence. The measures are primarily proposed to reflect the average CT dosing at the institutional level and small variations in patient size will average out across institutions.

We have conducted comparison of each of the dose metrics with measures of absorbed dose among a sample of 10,000 CT examinations and the correlations are high (> 90%). Further, the correlation within the metrics is also high. Details of this comparison were provided at the time of consideration of this measure when it was first endorsed. The organ doses were calculated by Dr. Choonsik Lee, PhD an Investigator in the Radiation Epidemiology Branch, in the Division of the Cancer Epidemiology and Genetics at the National Cancer Institute. His research includes the development of dosimetry databases and Monte Carlo dose calculations using human models that permit estimating absorbed radiation dose that takes into account patient weight. His method for estimating organ doses has been validated against direct measurement.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The different metrics are highly correlated *(see Keegan JACR 2014, Attachment of UCSF CT DOSE Report).* The metrics are highly correlated with absorbed doses.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The metrics are valid and meaningful and will reflect a facilities average CT doses

Of note, these are not patient level metrics and, for an individual patient, do not provide information about whether the dose that was used was appropriate.

2b3. EXCLUSIONS ANALYSIS

NA ☑no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☑**Stratification by three anatomic areas and five pediatric age groups risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The radiation doses used for CT vary by anatomic area (head, chest, abdomen and pelvis) and age (adult versus various child age groups.) *Smith-Bindman (JAMA Int Med 2009; JAMA 2012); Miglioretti JAMA Pediatrics 2013, JACR 2014).* There are further the categories that have been used in radiation safety programs for data collection In Europe and the UK. While there are other factors that influence radiation doses, these are relatively minor in comparison to these groups, and it is not feasible to collect or report data into smaller stratifications. Nor does the measure lose validity by not stratifying by clinical indication or patient size (see below.)

Why it is not important to stratify for clinical indication or protocol

The way CT scans are conducted should vary by why the patient is being scanned. For example, a search for occult malignancy may require very different parameters than an assessment for bleeding for trauma. Unfortunately there is currently no standardization for either categorizing the indications for imaging, nor for deciding the best way to image given a patients suspected problem, nor for categorizing the protocols that are used. I am currently leading a project to standardize the protocols we use for imaging across the five University of California Medical Centers and for each indication, the different institutions have adopted dozens of different ways to image patients with similar clinical questions. Smith-Bindman (*JAMA Int Med 2009*) highlighted the issue of differing ways to imaging a very standard problem and the resulting radiation dose. For example, while some institutions chose to image patients with suspected stroke using a standard 2 mSv CT, others routinely use a high dose CT where the doses were on average 20, and some facilities used 58 mSv. There are almost no standards for defining how to image different clinical questions and the profound variation reflects physician preferences, and the promotion of certain protocols by the manufacturers, rather than evidence that the higher dose protocols are more accurate or diagnostic or truly needed based on evidence. As an example, in a NEJM article (*Smith-Bindman 2010*) images were included from a patient who underwent two chest CT examinations for the same clinical indication at the same institution one year apart. The patient had a 1.5 mSv dose study on one occasion and a 15.9 mSv study on the second occasion. Both studies were done for exactly the same reason of the surveillance of a pulmonary nodule, and both were done within a single institution and within the setting of a clinical trial where what was done should be standardized. Thus there is profound variation in how studies are conducted, even in the few situations where the reason for imaging is known and guidelines exist. Further, there are often financial incentives that drive the decision to image using repeated imaging protocols versus single imaging protocols (even though the former could lead to doses that are twice as high as the latter). For example, there were recent reports that some facilities use double imaging protocols (with and without contrast) for conducting Chest CT, thereby double billing and double radiating the patient, in a setting where doing two scans is considered rarely necessary. Thus while some facilities were using double scanning in 1% of patients, others were using this in 80% of patients, and CMS has concluded that this reflected overuse of CT(see http://www.nytimes.com/2011/06/18/health/18radiation.html?_r=1)

Anatomic area, rather than specific indication or protocol, will actually provide the patient with the information they want to know – i.e. if I go to a facility, how high or low will my dose be. It will also allow

facilities to identify where they need to explore their doses in greater detail to assess why they are outside the normative range – is it that the are using too high doses within a protocol or using high dose protocols too often. The way the measure is currently written the choice of protocol will be reflected within the facilities metrics, whereas if dose were reviewed only within protocol, the facility that chooses to use high dose studies and repeated studies on most of its patients would appear fine.

Why it is not important to adjust for patient size

Weight will contribute to the variation in dose used for CT, and if individual patients were compared, it would be extremely important to assess weight when deciding about optimum ways to set up CT scans. Differences in weight may account for a 1-3 fold difference in the radiation used. Dr. Huda has published several relevant recent papers showing that doses vary up to 2 fold based on patient weight . "Radiation related cancer risks in a clinical patient population undergoing cardiac CT" AJR 2011 and "Estimating cancer risks to adults undergoing body CT examinations" Radiation Protection Dosimetry 2011. However, its important to point out that it is in no way established exactly how to increase doses for larger patients – i.e. there is no clear standard. A recent and interesting article found that machines that automatically adjust for patient weight seem to be giving too much radiation so that the organ doses increase even more so than does the weight (Israel, G. M., Cicchiello, L., Brink, J. and Huda, W. Patient size and radiation exposure in thoracic, pelvic, and abdominal CT examinations performed with automatic exposure control. Am. J. Roentgenol. 195, 1342–1346 (2010).

We have assessed the association between weight and the doses used, and presented at the initial submission of this metric, with an explanation of why it is not important to adjust for weight. When we compared the radiation dose used among patients in the top quartile of weight, to the radiation dose used in the bottom quartile of weight, the average doses increased by a factor of less than 2. For example among adult patients age 25 and older in the lowest quartile of weight (i.e. those under 152 lbs) the mean DLP among patients who underwent an abdominal and pelvic CT was of 781 . Among patients in the largest quartile of weight (ie those between 220 and 425 lbs, reflecting a mean weight twice as high), the mean dose was 1282 DLP or around 60% higher. However, within each of the weight groups, there was much more dramatic variation within group, then between groups. For example, among the smallest patients (those <25%) the range in dose between the 1st and 99th distribution was 54 – 1890 (40 fold variation between the highest and lowest group), and in variation in the highest quartile of weight was 352 – 2885 (8 fold variation). Thus the variation in dose based on weight was small in comparison to weight based on other factors (such as physician and facility preferences).

We have a paper in press in Radiology that assesses, among a large sample of 800,000 CT scans factors that influence dose, site variation contributes far more variation to the model than even patient size.
These weight differences are not relevant at the facility level, as while patient size may influence dose by 2 fold (between the smallest and largest patients) other factors, can influence the dose by up to 100 fold (based on our data), and these factors, rather than individual patient weight, will drive the facility level dose indices measures. Even if a facility had ALL patients of a size <25%, versus all patients over the 75% the influence would be very modest.

However, while I do not believe including weight would influence a facility's measures, there have been several recent publications which provide simple ways to account for size when reporting radiation dose, and including one of these metrics in the measure may allow greater adoption of the measure by various stakeholders. These measures essentially have determined for a fixed amount of machine dose, how the absorbed dose to the patient varies by their size; larger patient will tend to have a lower adjusted dose (because the same dose is spread out in their larger body) whereas a smaller patient will have a correspondingly larger dose (because

the same dose is distributed in a small volume of tissue.) Using these adjustment factors, it is possible to get a more precise estimate of the dose absorbed by the patient based on the machine output and a conversion factor based on the patient's size. The SSDE measurement (AAPM Report 204, Size Specific Dose Estimates in Pediatric and Adult Body CT Examinations) is now included in this measure and is a measure that accounts for patient size.

In our work across the University of California Medical Centers (as part of the UCDOSE, PI Smith-Bindman), among > 100,000 CT scans, there was no difference in facility level conclusions about performance when any of the metrics were used (i.e. SSDE, CTDIvol, DLP and ED) all characterized facilities the same.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Comparing institutional performance to benchmarks permits identification of outlying performance. Because the metric is based on summarizing dose for a large number of individuals (> 100 within each strata) and comparison to benchmarks, the comparisons are stable at identifying outlying performance. In the attached document (UCSF CTDOSE Report), we illustrate the result of comparing institutions (using t-tests and quantile regression) using the NQF measure format. Basically, facilities can be identified and compared with benchmarks, and stable estimates of facilities with outlying performance can be identified. *See Miglioretti 2014 JACR, Keegan JACR 2014*

While the generation of averages will permit the comparison of facilities to benchmarks, the measure does not specify cutoffs or how the comparisons would be judged. These can be set based on the clinical or quality improvement needs of a facility, organization, etc.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

Outlying institutions can be easily identified

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean*

and what are the norms for the test conducted)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The measure calls for collecting consecutive scans so that participants cannot choose their best or most optimum dose metrics to quantify. The data will be available, or can be calculated from essentially all (>95%) of CT scans

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Other

If other: The two of the specified metrics (CTDIvol and DLP) are generated as part of clinical CT examinations. The two additional metrics can be easily calculated from these two primary metrics and these calculations are done within existing software products or can be done manually, or using various additional approaches.

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Two of the dose indices that are specified (DLP and CTDIvol) are available on nearly all (>95%) of CT scans conducted in the US. The FDA collects dose data on a sample of imaging examinations every year as part of a collaborative effort with states called the NEXT survey. The last year data were collected on CT exams was in 2005. These data are collected based on phantom studies (ie CTs conducted on sophisticated plastic phantoms rather than patients, thus providing data different from, although complimentary to, the proposed metric). However, as part of that survey the FDA documented that he vast majority of CT machines in operation will document DLP and CTDIvol. (Unpublished, information provided by Dave Spelic, FDA). Given the adoption of uniform standards described above, this number should be higher today.

SSDE can be calculated manually, and is currently calculated by many vendors who developed software to extract radiation dose metrics from CT machines or PACS. Thus this metric is almost as available as the other metrics.

Thus nearly all facilities that perform CT examinations can collect the specified indices outlined in this measure. There could be a

small number of facilities that have only very old CT scanners that do not routinely record this information, yet even for these, there are simple excel based programs – such as IMPACT CT, or CT EXPO - that allow the input of technical parameters to generate these values.

There may be a small number of CT scans where these data are simply missing (probably < 1/1000 examinations) but their exclusion from both the numerator and denominator will have no significant bearing on the overall distribution of the dose indices.

A busy facility center can abstract data on scans that were conducted over a few days to have sufficient sample size, whereas smaller centers may to compile data from several months to generate sufficient data within each anatomic area/age/machine type category.

On a practical level, these data are readily available and easy to assemble. Specifically, a medical chart abstractor or technologist would need to record the CTDIvol and DLP data from a review of the CT images on a PACS scanner, CT console, or medical record. These data are thus captured from displayed values on the CT operator console or otherwise electronically harvested.

If facilities strive to achieve a high rate of reporting the radiation dose data in the medical records it would be easy in the future to compile the data for this measure using data in the medical records.

Lastly, the CT manufacturers have agreed to uniformly adopt the same standard for reporting the radiology dose data (called the Dose SR [standard report]) and all new machines have had this feature since the end of 2010, providing a method whereby this is available to a proportion of existing scanners. With this feature, generating these metrics will be extremely simple.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
N/A

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose

- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

The assessment of radiation dose for diagnostic imaging is a relatively new concept. Several large and small organizations, hospitals, and hospital associations (particularly those that focus on children) are beginning to assess radiation. The Joint Commission, starting this year (2015), will start asking the facilities they oversee to begin assessing radiation doses as well. But otherwise, the concept is new.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

There are innumerable organizations that would potentially be interested in collecting and using this data for accreditation and benchmarking, if the measure were adopted and endorsed. These organizations range from the Joint Commission, to innumerable hospital organizations, to insurers and to CMS (if the measure were applied to adults). Further, UCSF has a large project entitled Partnership for Dose that Dr. Rebecca Smith-Bindman leads (the title author of this measure), and she would be willing to commit to allowing any organizations who are interested to submit their data to this project, for use in performing benchmarking and certification. The Partnership for Dose currently have 150 hospitals/outpatient facilities that participate in the project, and we have the team and expertise to be able to do this. If the measure is endorsed, Dr. Smith-Bindman will work closely with all of these mentioned organizations to try to move ahead to submit an accountability application.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

We have used the NQF method of assessing facility level and provider level radiation dose metrics and have demonstrated substantial improvements in dose over time (Keegan, 2014; Demb, 2015; Duncan, JACR 2013) and as the result of a randomized trial of an educational intervention and process whereby technologists were shown their performance using the dose summary that was designed to follow the NQF measure (Miglioretti, 2014).

Cited in this section:

Demb J, manuscript under preparation. CT Radiation Dose Standardization Across the University of California Medical Centers Using Audits to Optimize Dose. 2015.

Following an in-person meeting regarding CT radiation dose, radiologists, technologists and medical physicists from University of California medical centers strategized how to best optimize dosing practices at their sites, which were then analyzed for effectiveness and success after implementation.

Duncan J, Street M, Strother M, et al. Optimizing Radiation Use During Fluoroscopic Procedures: A Quality and Safety Improvement Project. J Am Coll Radiol. 2013;10:847-853

"A systematic approach to improving radiation use during procedures led to a substantial and sustained reduction in risk with no reduction in benefits. Data were readily captured by both manual and automated processes." – Conclusion statement from Abstract

Keegan J, Miglioretti DL, Gould R, Donnelly LF, Wilson ND, Smith-Bindman R. Radiation Dose Metrics in CT: Assessing Dose Using the National Quality Forum CT Patient Safety Measure. Journal of the American College of Radiology: JACR; 11(3):309-315. http://download.journals.elsevierhealth.com/pdfs/journals/1546-1440/PIIS1546144013006625.pdf. Mar 2014

Looking at dose metrics as per compliance with the previously endorsed #0739 NQF measure results in reasonably timed acquisition of CT doses, and seeing such doses resulted in 30-50% dose reduction.

Miglioretti DL, YX Zhang, E Johnson, N Vanneman, R Smith-Bindman. Personalized Technologist Dose Audit Feedback for Reducing Patient Radiation Exposure from Computed Tomography. Journal of the American College of Radiology: JACR 2014.
"Personalized audit feedback and education can change technologists´ attitudes about, and awareness of, radiation and can lower patient radiation exposure from CT imaging." – Conclusion statement from Abstract

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.
There are two potential limitations of the proposed measures that need to be described. CT radiation dose will vary by patient size, and the specific protocols used, and yet we are not suggesting the dose indices be collected in separate strata for size (other than for children) nor for different protocols. These two issues will be addressed separately.

PATIENT SIZE

One factor that influences the radiation dose in CT is patient size. In general higher doses are used in large patients in order to maintain the same image quality as can be achieved with lower doses in smaller patients. It simply takes higher doses of radiation to penetrate (get through) larger sized patients. Thus the recorded radiation doses in part will reflect the size of the patients seen.

If a facility sees a very high proportion of obese patients, their doses will be higher than a facility that sees very thin patients. This issue will be important when facilities compare their dose indices to normative data (to the diagnostic reference level data), as they should compare their actual data to data of facilities that assess similar patients. This is the reason that facilities should note the state where their facility is located if they submit their data to a national organization. Diagnostic reference levels should be generated at a local enough level (state, or region of the country) so they are most useful and relevant with respect to the size of patients scanned. Thus diagnostic reference data should reflect geographic differences and be appropriate to the typical patients seen in a given area, as called for in the FDA white paper on radiation safety. Thus if patients tend to be larger in the Northwestern states, the diagnostic reference levels may be higher in that region. As long as a given facility is compared to the correct area, this would have no impact unless a facility differs profoundly from the other facilities in its geographic region. Of note, the differences in patient size will only have a relative small impact on dose (around a two fold difference between the smallest and largest adult patients,) whereas variation in dose by 20-50 fold have been seen unrelated to patient size (Smith-Bindman, JAMA 2012; Smith-Bindman, JAMA IM 2009; Miglioretti, JAMA Pediatrics 2013; and Miglioretti, JACR 2014). Thus, while the current metrics does not perfectly account for size, size is a small contributor to dose, in comparison to much larger, unexplained and unjustified variation. (Smith-Bindman, Radiology, 2015).

Thus the validity of the proposed NQF measure dose not require consideration individual level adjustment of patient size. Facilities (even without consideration of external data) can compare their own data from one year to their data from prior years, and unless there is a profound shift in the weight of their patients, this will have no impact on their data. Facilities should still perform in-depth analysis of patient's who receive high radiation doses (perhaps above the 75% distribution at their own institution) to determine if those doses were appropriate and justified, or if they could have been reduced.

Further, none of the quality control programs in existence and described above (UK, European or American College of Radiology Programs) assess patient weight in conjunction with CT dose measures. It is simply not feasible, and would make it far more difficult for facilities to assemble dose data, as this information is not recorded as part of the radiology medical record, and is typically not available anywhere for most patients seen in outpatient settings. Difference in patient size is only one factor contributing to dose, and likely accounts for only a small amount of the large variation in dose within and between facilities.

The issue of the validity of this measure without consideration of patient size was vetted with a large number of physicists. There was widespread agreement that this measure as specified was highly valuable. Three letters of support originally submitted with this

measure (from the ACR, NCRP and FDA) supporting the measure as specified were included with the initial submission of this NQF measure when it was first approved.

CT PROTOCOLS

The way CT studies are conducted (the "protocols" using the language of CT) leads to the radiation doses patients will receive. These are the specific instructions the radiologist or other physician and technologists program into the CT machine at the time of scanning. The instructions include how large an area to scan, how many times to scan each area and the settings of kVp and mAs to use. If a larger anatomic area is imaged, the dose the patient receives will be higher. If a multiphase study is done (meaning a single anatomic area is imaged many times) the dose will be higher than if a single-phase study is done. If a facility chooses to use multiphase protocols frequently, or to scan large anatomic areas frequently, their doses will be higher than facilities that try to minimize the area imaged or number of scans taken. The type of scans done in Los Angeles California and Huntsville Alabama that led to the extreme radiation dose exposures for CT, were perfusion scans, a type of scan where a small area of the brain is imaged dozens, and sometimes hundreds of times.

The two ways to collect and compare CT dose index information would be first to compare doses WITHIN the specific study type - thus compare doses for routine single phase studies and compare doses for multiphase studies, or second, to compare typical doses for all patients who undergo a CT within a single anatomic area (ignoring considering of the specific protocol used).

The latter method is far more practical. It's a large amount of work to determine the specific protocol, why a study was done, whether it was routine or not, how many phases were used, and it is simply not practical to have a data abstractor or technologist necessarily know how distinguish the study type. However, the latter method is far more valid, reproducible and a reliable measure of quality. This is particularly true as there are no evidenced based guidelines about when particular protocols should be used. In particular the multiphase, higher dose protocols are not clearly indicated in particular clinical situation, studies have not shown they lead to improved diagnoses or quantified the potential harm in their use, and differences reflect practice variation more than any objective criteria of the need for these multiphase, studies. That's not to say that these higher dose protocols don't have any value – but only that decisions about when to use different protocols are more based on physician preferences that patient outcomes, and choosing to frequently use these higher dose protocols should be reflected in the radiation dose quality metrics generated at a facility. Comparing doses within protocol would profoundly mask true differences to patients. In the example provided above relating to renal protocol CT is an example. While most institutions indeed had on their books a low dose protocol, these protcols were infrequently used. Comparing dose within protocol would have masked the actual doses patients receive.

To highlight this issue, a concrete and very realistic example has been provided below of two facilities and their choice regarding imaging patients with head CT. Keep in mind that the question a patient, a referring clinician, a radiologist, a hospital administrator or payer might wonder is what is the dose Ms. Smith will likely receive if she goes to a particular facility for a head CT.

Two facilities (A and B) will have different doses for different exam types and will have a different distribution of how often the different exam types are used.

For the sake of this example, we will estimate that a basic head CT has a dose of 2-3 mSv and a multiphase head CT has a dose of 20-30 mSv (Smith-Bindman, Arch Intern Med, 2009) did you want something else shown from this section?

Routine head CT	2-3 mSv
Multiphase head CT	15-20 mSv

For the sake of this example, we will estimate that facility "A "uses the routine head CT for most of their patient's (95%) and that at facility "A" the dose for the basic head CT is 2.5 mSv, and is 20 mSv for a multiphase head CT .

At facility "B" they use the routine head CT less often (50%) and use the multiphase CT more often (also 50%.). Their dose for the basic head CT 2 mSv (lower at this facility as they use the much higher dose, multiphase study more often, so can get away with lowering the dose on the routine study). They also have a lower dose for the multiphase study, at 15 mSv.

For the sake of this example, we estimate each facility will conduct 100 head CTs over the course of a week.

If the two facilities were compared within protocol study type, facility "B" would appear to be doing a better job at dose reduction, as they have a lower dose for a routine head CT (2 mSv versus 2.5 mSv) and have a lower dose for a multiphase head CT (15 mSv versus 20 mSv). And yet this facility is using the higher dose multiphase protocol far more often which results in higher doses on

average to patients.

Thus if we would compare the average dose per head CT at facility A (which is the clinical quality question a patient and payer would care about), it would be far lower at facility "A." Facility "A" has an average dose of 3.4 mSv (95% low dose studies * 2.5 mSv + 5% high dose studies*20 mSv)/100] whereas facility "B" has an average dose that is substantially higher at 8.5 mSv (50% low dose studies * 2 mSv + 50% high dose studies*15 mSv)/100].

Thus the dose patients receive will be driven by the choice of protocol more than the dose within protocol and doing comparisons only within protocol with mask real and important differences. Thus comparing overall exposure within anatomic area is not only more feasible, it is more appropriate if the goal is to identify facilities where the typical doses are simply too high. The facility with atypical doses could explore why their doses are high.

Cited in this section:
Smith-Bindman 2015, Predictors of Computed Tomography Radiation Dose and Their Impact on Patient Care. In Press, Radiology

Smith-Bindman R, Lipson J, Marcus R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med 2009;169:2078-86.
" Radiation doses from commonly performed diagnostic CT examinations are higher and more variable than generally quoted, highlighting the need for greater standardization across institutions." – Conclusion statement from Abstract

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** NQF_Attachments_2015_10_09.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** University of California, San Francisco
Co.2 **Point of Contact:** Rebecca, Smith-Bindman, Rebecca.smith-bindman@ucsf.edu, 415-353-4946-
Co.3 **Measure Developer if different from Measure Steward:** University of California, San Francisco
Co.4 **Point of Contact:** Rebecca, Smith-Bindman, Rebecca.smith-bindman@ucsf.edu, 415-353-4946-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Karishma Kumar, MPH, assisted in the drafting and development of this measure.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** xx
Ad.7 **Disclaimers:** xx

Ad.8 **Additional Information/Comments:** xxx