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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3443

Corresponding Measures:

De.2. Measure Title: All-cause emergency department utilization rate for Medicaid beneficiaries with complex care needs and high costs (BCNs)

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services, Centers for Medicaid & CHIP Services

De.3. Brief Description of Measure: All-cause emergency department (ED) utilization rate for adult Medicaid beneficiaries who meet BCN population eligibility criteria. The measure is calculated as the number of ED visits per 1,000 beneficiary months and is intended to be reported at the state level.

For the purpose of this measure, the BCN population is defined as Medicaid beneficiaries who are age 18 to 64 during the lookback year (the 12 months prior to the measurement year) and the measurement year and have at least one inpatient admission and at least two chronic conditions, as defined by the Chronic Conditions Data Warehouse (CCW), during the lookback year. Beneficiaries dually enrolled in Medicaid and Medicare and beneficiaries who had fewer than 10 months of Medicaid eligibility in the lookback year are not included in the analytic sample because we did not have enough utilization data to include them in testing. We further limited the analytic file to beneficiaries that met the BCN definition criteria described above.

1b.1. Developer Rationale: The BCN population is heterogeneous, with varying medical, behavioral, and psychosocial care needs. Consistent across the BCN population, however, is a pattern of health care consumption characterized by a disproportionately high use of inpatient and ED services, often coupled with underutilization of preventive and other types of outpatient care. We also see significant variation in the BCN-1 measure by beneficiary subgroups (see 1b.4). Frequent ED utilization among BCNs--especially for ambulatory care sensitive conditions--are very costly and may signal poor access to primary care, suboptimal care coordination, and/or lack of supportive services across transitions in care settings (Billings and Raven 2013; Capp et al. 2013; Doupe et al. 2012; Pukurdpol et al. 2014). Therefore, measuring and subsequently reducing ED utilization through improved care coordination and access to community-based care represents an opportunity to improve both quality and cost of care for BCNs (Durand et al. 2011; Utah Office of Health Care Statistics 2004). This ED utilization measure should be paired with the all-cause inpatient admission measure (BCN-2) under development by CMS/Mathematica to assess overall hospital-based care. Both measures are intended to be used for state-level monitoring and quality improvement activities.

References:

Billings, J., and M.C. Raven. “Dispelling an Urban Legend: Frequent Emergency Department Users Have Substantial Burden of Disease.” Health Affairs, vol. 32, no. 12, 2013, pp. 2099–2108.

Capp, R., M.S. Rosenthal, M.M. Desai, L. Kelley, C. Borgstrom, D.L. Cobbs-Lomax, P. Simonette, and E.S. Spatz “Characteristics of Medicaid Enrollees with Frequent ED Use.” American Journal of Emergency Medicine, vol. 31, no. 9, 2013, pp. 1333–1337.

Doupe, M.B., W. Palatnick, S. Day, D. Chateau, R.A. Soodeen, C. Burchil, and S. Derksen. “Frequent Users of Emergency Departments: Developing Standard Definitions and Defining Prominent Risk Factors.” Annals of Emergency Medicine, vol. 60, no. 1, 2012, pp. 24–32.

Durand, A.C., S. Gentile, B. Devictor, S. Palazzolo, P. Vignally, P. Gerbeaux, and R. Sambuc. “ED Patients: How Nonurgent Are They? Systematic Review of the Emergency Medicine Literature.” American Journal of Emergency Medicine, vol. 29, no. 3, 2011, pp. 333–345.

Pukurdpol, P., J.L. Wiler, R.Y. Hsia, and A.A. Ginde. “Association of Medicare and Medicaid Insurance with Increasing Primary Care-Treatable Emergency Department Visits in the United States.” Academic Emergency Medicine, vol. 21, no.10, 2014, pp. 1135–1142.

Utah Office of Health Care Statistics. “Primary Care Sensitive Emergency Department Visits in Utah, 2001.” Salt Lake City: Utah Department of Health, 2004. Available at http://utah.ptfs.com/awweb/guest.jsp?smd=1&cl=all_lib&lb_document_id=12114. Accessed July 26, 2016.

S.4. Numerator Statement: The number of ED visits in the measurement year among adult Medicaid beneficiaries who meet BCN population eligibility criteria.

S.6. Denominator Statement: Number of Medicaid-eligible months ("beneficiary months") among adult Medicaid beneficiaries who meet BCN population eligibility criteria.

S.8. Denominator Exclusions: N/A

De.1. Measure Type: Outcome

S.17. Data Source: Claims

S.20. Level of Analysis: Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

3444: All-cause hospital utilization for Medicaid beneficiaries with complex care needs and high costs (BCNs)

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? BCN-1 is part of a paired measure set titled “All-cause hospital utilization for Medicaid BCNs.” The other measure in the pair is all-cause inpatient admission rate for BCNs, referred to as “BCN-2.” The BCN-1 and BCN-2 measures are intended for voluntary use by states to monitor and improve the quality of care provided to the Medicaid BCN population.

Monitoring BCN-1 in conjunction with the paired indicator for inpatient admissions (BCN-2) can identify meaningful differences in overall hospital utilization and help ensure that reductions in inpatient admissions accurately reflect true increases in quality care. Examining either measure in isolation has the potential to produce inaccurate inferences about states’ performance due to the potential for the substitution effect. For example, if an accountable entity decreases ED utilization among the BCN population by keeping patients in the hospital overnight, thus shifting the utilization to an inpatient admission, the ED measure concept alone would understate the accountable entity’s overall impact on hospital-based care. Conversely, a successful BCN intervention may decrease inpatient utilization (or the length of an inpatient admission) but increase ED utilization, appearing to have worsened outcomes despite reducing overall rates of hospital-based care. Indeed, several Medicaid BCN initiatives have successfully reduced inpatient admissions but increased ED utilization. For example, Washington State’s Chronic Care Management program resulted in significant

reductions in inpatient admissions and spending, and increases in ED utilization, though not at a statistically significant rate (Xing et al. 2015). If measured solely on ED utilization, Washington State might appear to have worsened outcomes, despite reducing overall rates of hospital-based care.

Reference: Xing, J., C. Goehring, and D. Mancuso. "Care Coordination Program for Washington State Medicaid Enrollees Reduced Inpatient Hospital Costs." Health Affairs, vol. 34, no. 4, 2015, pp. 653-661.

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary or **Summary of prior review in [year]**

- This measure of emergency department utilization for Medicaid beneficiaries with complex care needs and high costs (BCNs) assesses a heterogeneous population with disproportionately high use of inpatient and ED use.
- The developer notes that improvement on this outcome may involve strengthening beneficiaries' relationships with health care providers in the community, improved care coordination, and chronic disease management.

Updates:

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *If derived from patient report, does the target population value the measured outcome and finds it meaningful?*

Updates:

Exception to evidence

n/a

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

OR

If the developer provided updated evidence for this measure:

- The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?
- Questions specific to the measure information provided on evidence
- For possible exception to the evidence criterion:

- Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?
- Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?
- Does the SC agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?

Guidance from the Evidence Algorithm

Box 1: The measure assesses a healthcare outcome →Box 2: The developer has provided empirical data that there is a relationship between the measured outcome and at least one healthcare outcome → Pass

The highest possible rating is pass.

Preliminary rating for evidence: ☒ Pass ☐ No Pass

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer demonstrates an adjusted performance range of 109.5 admissions per 1,000 beneficiary months to 322.0 admissions per 1,000 beneficiary months.

Disparities

The developer examined Medicaid claims data across 10 states and presented a breakdown of performance by a number of subpopulations.

Table of Risk-adjusted measure rate distribution by subgroup

Subgroup	# of Bene-ficiaries	Weighted Mean	Minimum	25th percentile	50th percentile	75th percentile	Maximum
Aged 18 to 24	17,781	246.9	117.6	187.2	248.2	262.6	284.2
Aged 25 to 44	58,949	251.6	114.4	219.9	255.6	285.1	437.6
Aged 45 to 64	65,463	214.9	101.7	187.2	218.7	258.6	303
Male	49,798	231.1	113.3	204.6	229.9	274.2	535.9
Female	92,395	235.6	104.6	179.5	234.3	262.3	277.3
White	82,536	218.2	93.6	189.7	216.9	244.9	265.1
Black	41,627	273.7	0.0	245.2	281.1	303.9	408.8
Hispanic	10,647	247.4	0.0	171.0	204.4	230.3	284.3
Other/unknown	7,383	192.0	113.4	133.3	164.5	213.6	373.8
Aged/Blind/ Disabled	80,569	251.7	113.8	222.4	260.0	312.5	345.7
Adult	58,450	208.9	0.0	176.9	206.7	229.6	272.6
Child	3,174	225.1	120.9	191.3	219.1	236.5	309.3
1 or more behavioral health conditions	113,051	255.7	117.2	220.8	256.1	298.3	317.4
No behavioral health conditions	29,142	149.5	85.19	124.5	149.6	162.6	289.8

Source: Mathematica analysis of 2013 and 2014 MAX PS, LT, OT, and IP files.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission?For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- Appears appropriate
- OK
- Various programs have been shown to impact healthcare utilization, specifically ED use, by BCNs, although some more than others.
- Yes. appropriate evidence exists re: this healthcare outcome

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- Developer examined Medicaid claims across 10 states
- Yes
- Developers presented variability in performance and also some subgroup information.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

2b. Validity: Testing**;** Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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. For maintenance measures – less emphasis if no new testing data provided.

Validity

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. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators:

- Larry Glance
- Karen Joynt Maddox
- Marybeth Farquhar
- Eugene Nuccio
- Christie Teigland
- Steve Horner

Evaluation of Reliability and Validity (and composite construction, if applicable):

Summary of Methods Panel Review:

- In their preliminary analyses, subgroup members did not reach consensus on the validity of the measure. During the Methods Panel measure evaluation call, subgroup members expressed concern regarding the risk-adjustment approach, including inclusion of risk factors that are not statistically significant or are "protective" in nature, and also noted that risk of over-fitting was not assessed. Members also were concerned that comparability between state Medicaid populations and the ease of enrollment in Medicaid are potential threats to validity. Ultimately, the subgroup did not reach consensus on the validity criterion. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure in the Fall 2018 cycle.

Standing Committee Action Item(s):

- The Standing Committee will need to discuss validity (particularly the concerns articulated by the Scientific Methods Panel). It is important to note that the appropriateness of inclusion or exclusion of social risk factors was not within scope for the Scientific Methods Panel ratings.
- The Standing Committee can discuss reliability or accept the ratings of the Scientific Methods Panel.

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

- The Scientific Methods Panel did not reach consensus on validity for the measure. Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?

Preliminary rating for reliability:	☒ **High**	☐ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☐ **Moderate**	☐ **Low**	☐ **Insufficient**

Scientific Acceptability

Measure Number: 3443

Measure Title: All-cause emergency department utilization rate for Medicaid beneficiaries with complex care needs and high costs (BCNs)

Type of measure:

☐ Process ☐ Process: Appropriate Use ☐ Structure ☐ Efficiency ☐ Cost/Resource Use

☒ Outcome ☐ Outcome: PRO-PM ☐ Outcome: Intermediate Clinical Outcome ☐ Composite

Data Source:

☒ Claims ☐ Electronic Health Data ☐ Electronic Health Records ☐ Management Data

☐ Assessment Data ☐ Paper Medical Records ☐ Instrument-Based Data ☐ Registry

Data

☐ Enrollment Data ☐ Other

Level of Analysis:

☐ Clinician: Group/Practice ☐ Clinician: Individual ☐ Facility ☐ Health Plan

☐ Population: Community, County or City ☒ Population: Regional and State

☐ Integrated Delivery System ☐ Other

Measure is:

☒ New ☐ Previously endorsed (NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? ☒Yes ☐ No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

   ***NOTE**: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   - Some Methods Panel members expressed concerns about the specifications being precise, unambiguous, and complete.
     - Please note that to be eligible for a moderate rating per NQF's reliability algorithm submitted specifications must be specifications precise, unambiguous, and complete so that they can be consistently implemented.
   - The developer provided several clarifying comments on the measure specifications to address the Methods Panel members concerns.
     - The developer noted that the exclusions in the numerator and denominator description are not measure exclusions. Rather, they reflect the way the analytic file was constructed given that the developer did not have access to Medicare data.
     - The developer noted that when it comes to the testing population, Medicaid beneficiaries who are dual eligible during the lookback or measurement year are excluded, regardless of months of eligibility. Non-dual Medicaid beneficiaries were excluded from the testing population if they had fewer than 10 months of eligibility during the lookback year.

- The developer also clarified that when it comes to implementation and how states will calculate the measure, dual eligible beneficiaries are included in the measure if they have at least 10 months of Medicaid eligibility during the lookback year. This approach aligns with the Medicaid Core Set.
- A summary of the Methods Panel members feedback is provided in item 2 below. Please note that this summary is intended to inform Standing Committee discussion.

2. **Briefly summarize any concerns about the measure specifications.**

   PANEL MEMBER 1: BCN population restrictions are noted in the brief description of the measure. This is a utilization measure that may be an effective way to monitor the impact of systemic/programmatic changes on utilization—especially when compared to its partner inpatient admissions measure.

   PANEL MEMBER 2: None.

   PANEL MEMBER 3: Incomplete. The developer notes several exclusions in their description of the numerator and denominator of the measure, however, the developer cites "N/A" for exclusions.

   PANEL MEMBER 4: Unclear how ED visits are identified in patients when they are not covered by Medicaid

RELIABILITY: TESTING

Submission document: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Neither**
   - Please note that NQF does not require data element reliability testing if data element validity has been demonstrated.
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**
   ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

   ☐ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   - The developer conducted signal-to-noise (SNR) reliability testing for BCN-1 using MAX data from 10 states.
   - Testing was not precisely conducted for the measure as specified. Specifically, dual-eligible beneficiaries were not included in the testing due to data unavailability, but would be included in the measure if implemented.
   - Panel members, in a future submission, would like to see analyses demonstrating the reliability of the data elements used in the measure. This is important because of the probable differences in the quality of Medicaid data across states.
     - The developer noted that the states selected for the testing sample had the highest quality data and were most likely to provide generalizable testing results. The developer also clarified that the the BCN population definition is intended to be applied uniformly across states in an effort to minimize differences in state-specific populations.
     - Additionally, the developer conducted a qualitative missingness analysis of the measure's key data elements using MAX anomaly tables (see section 2b6.2. in the Testing Attachment). Given the relatively small amount of missing information (key fields were missing less than 5% of the time) used to identify the BCN population and calculate the BCN measures, the developer does not believe there is any systematic bias in their testing. The developer also notes that states implementing the measure with their own data will likely have even fewer missing fields

because they are better equipped to account for state-specific codes when identifying the BCN population and constructing the measure.

- A summary of the Methods Panel members feedback is provided below. This feedback is intended to inform Standing Committee discussion.
    - **PANEL MEMBER 1:** Signal-to-Noise ratio (SNR) used; this is appropriate; sample of 10 states contains 2 small states and 8 larger states
    - **PANEL MEMBER 2:** Developers conducted signal to noise testing at the entity (state) level scores.
    - **PANEL MEMBER 3:** Developer used signal-to-noise ratio to determine reliability.
    - **PANEL MEMBER 4:** Measure signal to noise ratio
    - **PANEL MEMBER 5:** Signal to noise

7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   - Average signal-to-noise reliability estimate = 0.92 (ranging between 0.59 to 0.99 across the ten states in the sample).
       - Note: there was previously a small typo regarding the average - stated as 0.99 in text but shown as 0.92 in table. This has been updated since the Methods Panel review. The two states with very low sample sizes had the lowest signal-to-noise reliability estimates (i.e., 0.59 and 0.66)
   - A summary of the Methods Panel members feedback is provided below. This feedback is intended to inform Standing Committee discussion.
       - **PANEL MEMBER 1:** SNR values are very strong for larger states, and not surprisingly lower for small states. Overall, very good SNR results
       - **PANEL MEMBER 2:** The risk-adjusted measure is shown to be highly reliable, with an overall (mean) signal-to-noise measure reliability of 0.92 The SNR ranged from 0.59 to 0.99 across the ten states in the sample. The overall reliability of 0.92 indicates the measure is highly reliable and can discern performance differences between states.
       - **PANEL MEMBER 3:** Developer indicates 0.99 for SNR for BCN 1 however, Table 2 reflects a different measure (BCN 2) rather than BCN 1 with a different SNR. Interpretation of the results is for BCN 2.
       - **PANEL MEMBER 4:** State-level signal-to-noise ratio mean was 0.92, consistent with excellent reliability
       - **PANEL MEMBER 5:** Highly reliable at the state level. While all calculations meet reasonable thresholds, I'm curious as to what was different about States B and G that led theirs to be markedly lower (e.g were these the managed care states – suggesting data might differ – or another factor)?

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒**Yes**

   ☐**No**

   ☐**Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2

   ☐**Yes**

   ☐**No**

   ☒**Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results): Ultimately, the Methods Panel gave this measure an overall rating of reliability of moderate. Individual Methods Panel reviews responses ranged from insufficient to high. No members selected low however. A summary of the reviewers' rationales can be found in item 11 below.

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

    PANEL MEMBER 1: Very clearly described methodology and excellent SNR results.

    PANEL MEMBER 2: While the measure scores appear to have high reliability, the developers did not discuss reliability of the individual data elements in the measure and risk adjustment models. This seems particularly important for this measure given the probable differences in quality of Medicaid data across states.

    PANEL MEMBER 3: See #7 above for the rationale for the rating.

    PANEL MEMBER 4: State-level signal-to-noise ratio mean was 0.92, consistent with excellent reliability

    PANEL MEMBER 5: Highly reliable, likely because of the large sample size and straightforward claims-based outcome.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

    Submission document: Testing attachment, section 2b2.

- Methods Panel members raised a number of concerns with the measure exclusions. The measure developer provided a number of clarifications to address these concerns.
    - The developer noted that the exclusions in the numerator and denominator description are not measure exclusions. Rather, they reflect the way the analytic file was constructed given that the developer did not have access to Medicare data.
    - The developer noted that when it comes to the testing population, Medicaid beneficiaries who are dual eligible during the lookback or measurement year are excluded, regardless of months of eligibility. Non-dual Medicaid beneficiaries were excluded from the testing population if they had fewer than 10 months of eligibility during the lookback year.
    - The developer also clarified that when it comes to implementation and how states will calculate the measure, dual eligible beneficiaries are included in the measure if they have at least 10 months of Medicaid eligibility during the lookback year. This approach aligns with the Medicaid Core Set.
    - The developer noted that state-specific coding and other idiosyncrasies can still affect the interpretability of certain analyses of MAX data. The developer clarified that they required states in the testing sample to be free of data anomalies that could limit the ability to interpret testing results. The developer also required that states align with national benchmarks to confirm that the FFS population in each state was broadly representative and that state-specific coding conventions accurately captured hospital-based utilization in claims. The developer noted that if they included states that did not meet these criteria, they would not be able to tell whether differences in measure performance between states were meaningful or driven by the states' population composition, data anomalies, or inability to identify utilization due to proprietary coding. The states selected for the testing sample have the highest quality data and are most likely to provide generalizable testing results.
    - The developer clarified that ED measures not covered by Medicaid are not intended to be included in this measure.

- A summary of Methods Panel members feedback is provided below. This summary is intended to inform the Standing Committee's discussion.
    - **PANEL MEMBER 1:** BCN population is already restrictive compared with the entire Medicaid population.
    - **PANEL MEMBER 2:** The exclusions appear to be reasonable.
    - **PANEL MEMBER 3:** No exclusions cited by developer, although the developer cites a number of exclusions in the numerator/denominator details e.g., "beneficiaries dually enrolled in Medicaid and Medicare and beneficiaries who had fewer than 10 months of Medicaid eligibility in the previous 12 months", "ED visits contribute to the monthly count only if they do not result in an inpatient admission or observation stay", "Observation stays are not included in the sum of inpatient admission used to identify the denominator population", etc.
    - **PANEL MEMBER 4:** 34 of 50 states were excluded due to data issues – either lack of sufficient data or outcomes that did not align with national benchmarks. This significantly limits the overall generalizability of this measure. Not clear why states that are outliers should be excluded from measure development.Unclear how ED visits are identified in patients when they are not covered by Medicaid. This is a major threat to validity.d
    - **PANEL MEMBER 5:** An analysis of the individuals excluded for less than a year of Medicaid enrollment would be helpful just to see how they differ, though agree that since endpoints can't be ascertained they can't be included. Concern is that states with different ease of enrollment have differential churn, meaning a differential population going on and off Medicaid in any given year. Further, since churn may be linked to any number of health or social issues, these individuals may have higher event rates and thus states with higher churn would artificially look better. Would look at these folks state by state and for major demographics.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    Submission document: Testing attachment, section 2b4.

    - Methods Panel members noted the following concerns regarding the measure's ability to identify meaningful differences in performance.
        - **PANEL MEMBER 1:** Generalizability of the measure results may be problematic given the limited data across all 50 states. Only 15 states had MAX data and 5 of these states were dropped from the analyses due to failing one or more of their data quality criteria. The authors should be applauded for their specification of these data quality criteria.See other discussion which describes concerns in several areas. What constitutes "meaningful differences" vs. simply a difference across the states is not explicit in the document.
        - **PANEL MEMBER 3:** Analyzed the distribution of the measure rate at the state level and among demographic and clinical subgroups. Compared performance across state-level inpatient admission rate to understand variation. Used z-distribution with Benjamini-Hochberg procedure (adjusting p-values) to correct for false positives. Table 8 is confusing as there are a few subgroups identified but are not discussed in the description of the methods to determine if meaningful differences exist. It would be helpful if we had information e.g., data on the performance across state-level inpatient admission rates for comparison to the BCN population.
        - **PANEL MEMBER 5:** Only exclusions as above.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    Submission document: Testing attachment, section 2b5.

- One Methods Panel member noted a concern regarding comparability of results.
    - **PANEL MEMBER 5:** Information on comparability of populations across states would be helpful, since data are presumably coming in differentially for FFS versus managed care.

15. **Please describe any concerns you have regarding missing data.**

Submission document: Testing attachment, section 2b6.

- Methods Panel members noted the following concerns regaring missing data:
    - **PANEL MEMBER 1:** See previous statement about generalizability (and reportability) of results given the large number of states without data to analyze.
    - **PANEL MEMBER 3:** No analysis of the population with missing data or how it will be handled when calculating the measure
    - **PANEL MEMBER 4:** Unclear how ED visits are identified in patients when they are not covered by Medicaid. The amount for the outcome variable cannot determined using material presented by developer.

16. **Risk Adjustment**

16a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☐ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☒ Yes ☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☐ No

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care?

- Methods Panel memebers provided disagreeing repsonses to this question.

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion?

- The developer did not present a rationale for inclusion.

16d.3 Is the risk adjustment approach appropriately developed and assessed?

- Methods Panel memebers provided disagreeing repsonses to this question.

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration)
☒ Yes ☐ No

16d.5.Appropriate risk-adjustment strategy included in the measure?

- Methods Panel memebers provided disagreeing repsonses to this question.

16e. Assess the risk-adjustment approach

- The risk-adjustment approach was developed using data from 10 states. The risk-adjustment model included 69 risk factors. The Panel's concerns about the risk-adjustment approach included:
    - Several risk factors are included that are neither statistically nor clinically significant. However, the risk of over-fitting was not assessed.
    - The risk-adjustment model includes a factor noted as "child'. This is confusing given the measure is limited to individuals ages 18-64.
    - Poly-pharmacy was not included as a risk-factor.
    - The developer states they did not include social risk factors due to the findings from a recent NQF report on admissions/readmissions. This is an erroneous interpretation of that report.

    - Concern with excluding prior hospital-based care utilization as a risk factor.
    - Concern around using the chronic conditions data warehouse (CCW) fields to identify comorbidities used in the risk adjustment model. However, there was no supporting literature cited to support this decision and no validation of those variables (e.g., re-abstracting chart data) was conducted. Because this is a new measure, data element validation is not required. If endorsed, developers should consider presenting this type of analysis when the measure comes back for re-evaluation.
- The developer noted that they included all predictors that were theoretically associated with the measure, including those that were not statistically significant or "protective" in nature. The developed stated that in general, the risk factors associated with a *lower* adjusted risk of ED utilization reflected more serious conditions (e.g. colorectal cancer). This lower risk likely reflects higher substitution away from ED care towards inpatient care. Because BCN-1 will ultimately be paired with a measure of inpatient care, the developer believed it was important to include the "protective" risk factors in the BCN-1 risk adjustment model.
- The developers provided information in the testing attachment about the risk of overfitting. Specifcally, the developer noted that after the exploratory data analysis, they split the analytic sample into two randomly selected half samples. One served as the development sample supporting their model building and exploration work, and the other served as the validation sample against which we assessed the final model's performance. This approach is standard practice to avoid "overfitting" a risk adjustment model, which takes place when a model fits both the true underlying relationships between variables as well as idiosyncratic data fluctuations specific to the particular sample. Finding that their model performs well on the validation sample provides assurance that the model will generalize well to other samples, and is not primarily driven by such idiosyncratic fluctuations.The risk factors exhibiting coefficient instability (for example, Alzheimer's disease and traumatic brain injury) typically exhibited very low sample prevalence and imprecisely estimated null associations with the outcome. As a result, it is reasonable to conclude that coefficient differences were driven by statistical noise as opposed to potentially more worrisome overfitting bias.
- A summary of Methods Panel Feedback is provided below. This summary is intended to inform Standing Committee discussion.
    - **PANEL MEMBER 1:** The methodology used to develop and test the list of 69 risk factors was described in excellent detail. The negative binomial stratified by decile level showed very strong results.
    - **PANEL MEMBER 2:** Developers used bivariate exploratory analyses to identify potential risk factors. They split the sample into 2 randomly selected half samples and used one as development sample and other as validation sample. They also plotted the relationship between the number of emergency department visits and the number of months enrolled to determine whether to include an offset term in our final model. Based on the results, they chose the negative binomial regression with an offset term reflecting the number of enrolled months per beneficiary.The final risk adjustment model included 69 risk factors and an intercept term. The factors included sociodemographic indicators (mean-centered age and its square as it had quadratic relationship, sex, and Medicaid eligibility category), along with the entire set of CCW condition indicators in the risk adjustment model. Other variables indicated whether a beneficiary had physical health conditions only (reference), behavioral health conditions only, or both types of conditions (an interaction of the previous two).The developers excluded area-level socioeconomic status (SES). They indicate that this decision was motivated by the findings from the recent two-year National Quality Forum (NQF) effort, which indicated that the inclusion of area-level SES indicators did not improve the predictive capacity of risk adjustment algorithms of hospital-based care measures developed for Medicare beneficiaries. I do not believe that was the intent of the NQF conclusion, and that in this case, state level variables may have indeed proven significant given the disparities in

median income, poverty, education and other SES across states. We do not know which states are included here, but I do not believe we can say these results are generalizable to other states. Some N's were very small for some states used in the testing. Polypharmacy was also excluded and could potentially have a large impact on the results especially given the current opioid crisis and other issues around medication adherence and high use of medications in the Medicaid population, many of whom are disabled and/or have a large number of chronic conditions (more than the 2 used for cutoff here). I do not necessarily agree with the rationale for excluding prior hospital-based care utilization as a risk factor. A prior hospital stay or ER visit is a well -documented predictor of future stay. While the developers argue that inclusion may set inappropriate care incentives by rewarding poorly performing entities with a lower bar for expected performance in future years. The problem described is that it increases the predicted inpatient admission rate for entities whose beneficiaries previously had more inpatient admissions or ED visits. Thus, entities with previous high inpatient admission rates are expected to have higher inpatient admission rates in the measurement year compared with entities whose patients did not, effectively setting a lower standard for entities with previous poor performance. Therefore, they excluded prior hospital-based care utilization from consideration. An alternative theory is that prior hospital visits indicate the patient has higher level of severity or risk factors that are not captured in the model, such as additional social risk factors that are not measureable in claims data.

- **PANEL MEMBER 3:** Why include "child" in the model specs (Table 5) if that population is not in the measure? Is this population included in the calculation of 2014 centered age?
- **PANEL MEMBER 4:**
  - Use of non-hierarchical negative binomial model is a reasonable approach, although logistic regression modelling is more often used when the outcome is binary (instead of using count model). Model fit is more difficult to assess in negative binomial model because McFadden Rsq is not as easily interpretable compared to C statistic in logistic regression model.
  - Unclear whether CCW comorbidity time-stamp was used to only include risk factors present before ED visit
  - The inclusion of risk factors that lower the risk of the outcome is not a standard practice in risk adj models
  - There is very large number of risk factors, many of which do not achieve statistical significance – and, at the same time, do not have a clinically important effect size. The risk of over-fitting was not assessed.
- **PANEL MEMBER 5:** Extensive list of comorbidities and characteristics. Would have been nice to see some SES testing using area-level variables.

VALIDITY: TESTING

17. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**
18. **Method of establishing validity of the measure score:**

    ☒ **Face validity**

    ☐ **Empirical validity testing of the measure score**

    ☐ **N/A (score-level testing not conducted)**

    - NQF accepts face validity testing for new measure submissions. Empirical validity testing is required at the time of maintenance review.

19. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

    - A face validity assessment was conducted and meets NQF requirements.

- A summary of the Methods Panel members assessment of the methods for establishing validity is presented below. Please note this is intended to inform Standing Committee discussion.
    - **PANEL MEMBER 1:** Only 2/3rds of the 17 TEP members rendered an opinion about the numerator and denominator statement. No information if patients were represented in the TEP
    - **PANEL MEMBER 2:** Developers conducted an online survey of 17 TEP members to obtain their assessment of the measure's components and the extent to which the measure's state-level performance scores distinguish good quality from poor quality of care. Only 11 responded. Most of the respondents (82%) either agreed or strongly agreed that the denominator is appropriate. Among the two respondents who disagreed the developers indicated that one misunderstood the BCN population definition (which could point to a need to further clarify the numerator definition) and one believed that the definition of an eligible month should be clearer (again pointing to further need for clarification). All respondents either agreed or strongly agreed that the numerator is appropriate. Finally, the majority of respondents (70%) either agreed or strongly agreed that measure would distinguish between good and poor performance among states.
    - **PANEL MEMBER 3:** Face validity was conducted among a technical expert panel whose members had "guided the development of the measure"…An independent TEP would have been a better choice for validation purposes which could have resulted in stronger evidence of validity.
    - **PANEL MEMBER 4:** face validity using TEP, predictive validity by assessing predictive performance of risk adj model
    - **PANEL MEMBER 5:** Small sample survey

20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    - 11 of 17 TEP members responded to the relevant question asked as part of the face validity assessment. Seven of the 11 respondents agreed/strongly agreed that the measure can differentiate between good vs poor quality. Among those who disagreed, one misunderstood the BCN population definition and two did not give a reason.
    - A summary of the Methods Panel members feedback is provided below:
        - **PANEL MEMBER 1:** Of those providing an opinion, the majority supported the validity of the measure elements, and that the results could distinguish between good and poor performance.
        - **PANEL MEMBER 2:** I would assess the results above as moderate face validity given the small number of responses and the fact that this was the only face validity testing that was completed.
        - **PANEL MEMBER 3:** Level of agreement suggest that the measure has good face validity for monitoring quality improvement. Minor disagreement noted.
        - **PANEL MEMBER 4:** Only 65% of TEP responded to survey. This is a very low response rate for a TEP and calls into question the amount of confidence that the responses of the responders represents the overall opinion of the entire TEP. Calibration, as assessed using calibration graphs and OE ratios across different sub-groups is very good, providing strong evidence to support the predictive validity of the risk adj model.
        - **PANEL MEMBER 5:** Reasonable for face validity

21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

- Methods Panel members provided varying responses to this question. One Panel member provided the feedback below.
    - **PANEL MEMBER 4:** Construct/convergent validity – the agreement between a new measure and validated measures – is difficult to assess in many cases. It was not assessed here.

22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

    Submission document*: Testing attachment, section 2b1.*

    ☐**Yes**

    ☐**No**

    ☒**Not applicable** (data element testing was not performed)

    PANEL MEMBER 4: Although validity of data elements was not tested, there are concerns around using CCW fields to identify comorbidities for use in risk adjustment model. The validity of using CCW comorbidities to describe comorbidities was not described. No supporting literature was cited. No attempt was made to compare CCW comorbidities to an authoritative source (i.e. re-abstracting chart data).

23. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**
    - The Methods Panel did not come to consensus on an overall rating of validity for this measure. Methods Panel members ratings ranged from low to moderate.
    - A summary of Methods Panel members rationales for their rating is provided below.

24. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

    PANEL MEMBER 1: Because this is a new measure, only face validity is required. Empirical testing of both reliability and validity is warranted in the near future.

    PANEL MEMBER 2: No empirical testing of data element validity or score level validity. Only face validity conducted with TEP of 11 persons and 3 questions. Results indicated moderate agreement. I feel further reliability testing is needed to rate this higher. Also see discussion in 16e above which outlines some concerns regarding the approach to development of the risk adjustment model. In general the results make great sense for the variables that are included. I have some concerns about some that were excluded but there is no empirical comparisons for those to make a determination.

    PANEL MEMBER 3: While a TEP is used frequently to assess face validity, I have concerns that the TEP used to develop the measure is the same one to test the validity of the measure.

    PANEL MEMBER 4: Many ED visits may be missing since measure is based on Medicaid data set, and some or many patients may lack Medicaid coverage during full 12 months. This is the most important limitation of the measure.

    - Risk adj model includes
    - many risk factors that are both not statistically and clinically significant
    - many medical conditions that would lead to a decreased risk of ED visits. Use of risk factors that have a "protective" effect is non-standard and seems counter-intuitive. The explanation that TEP decided to retain these risk factors because they may be indicative of interaction effects should have been explored empirically. The risk of over-fitting was not assessed.
    - Use of CCW comorbidities to identify comorbidities may be appropriate, but no evidence is presented that these are valid data elements, or have been used previously in risk-adjusted outcome measures.

PANEL MEMBER 5: Concerns regarding exclusions and lack of SES testing (though population is more SES-homogeneous than many other measures) – I think the exclusions information should be provided if possible before moving this forward.

ADDITIONAL RECOMMENDATIONS

25. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

    PANEL MEMBER 5: Exclusions; Lack of SES testing

Committee Pre-evaluation Comments:
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- Appears appropriate
- Ok
- The calculation of beneficiary months is fairly complex. Given that MC beneficiaries can churn in/out in any given year this may also be a threat to validity.

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- NO
- Reliability testing used SNR statistic as the proportion of variation between scores relative to random variation; all SNs were high.
- SNR seems reasonable

2b1. Validity -Testing: Do you have any concerns with the testing results?

- Consensus was not reached
- Ultimately, the subgroup did not reach consensus on the validity criterion.
- Face validity was measured via TEP, with the majority of members in agreement with the metric.
- Who are the people who have less than 10 months' participation? Are they functionally different than those that were included in the measure?

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data): 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- Issue with missing data analysis
- Including non-significant factors in adjustment
- Is what we're testing here less about quality/performance differences than the interstate variability in MC programs/enrollment/disenrollment? Maybe they should adjust for rates of same year disenrollment? Do they have more hospital utilization?

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure?2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- Rationale for inclusion and risk adjustment

- Including non-significant factors in adjustment
- See above

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

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.

- This measure is calculated from claims data.
- All data elements are in defined fields in electronic claims.
- There are no fees associated with the use of this measure.

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?
- Is the data collection strategy ready to be put into operational use?

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- Measure calculated from claims data
- This is feasible
- Feasibility was acceptable for the use of state data; the authors conducted a power analysis to determine minimal sample sizes required to detect significant changes in rates.
- Claims based, should not be an issue

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- CMS plans to use the measure for internal quality improvement (state-level monitoring and quality improvement activities), but there are no specific details yet. This measure is intended for voluntary use by states to monitor and improve the quality of care provided for the Medicaid BCN population. States may choose to begin implementing the measures based on their programmatic needs. This measure is intended to be paired with BCN-2, an all-cause inpatient admission measure that was also developed specifically for the Medicaid BCN population.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

The measure is not yet in use. The developer states that there are no formal process to share draft results with measured entities, but they invited feedback from a 19-member TEP, a 7-member risk-adjustment work group, and the public (via a public comment process).

Additional Feedback:

N/A

Questions for the Committee:

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

RATIONALE:

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

The measure is not yet in use so there are no trend data.

4b2. Benefits vs. harms. 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).

Unexpected findings (positive or negative) during implementation

This measure has not been implemented yet. However, the developer noted were no unexpected findings identified during testing of this measure.

Potential harms

This measure has not been implemented yet. However, the developer noted were no unexpected findings identified during testing of this measure.

Additional Feedback:

N/A

***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?

Preliminary rating for Usability and use: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided?4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- Measure is not yet in use
- OK
- Not currently being used

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations?4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- Not yet in use
- OK
- Threats to usability include the number of states that do not have adequate data to calculate the rates. A comment was made about the utilization of ED by the BCNs that don't result in Medicaid payment; this should be examined to determine the impact.
- Unclear if data will be used to inform quality, although reporting on hospitalizations for other populations has been associated with decreased utilization, though now with some increased concern about unintended consequences.

Criterion 5: Related and Competing Measures

Related or competing measures

- NQF did not identify competing measures.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- Developer provided justification for differences in specificiations

Public and Member Comments

NQF received no public or member comments on this measure as of January 25, 2019.

Brief Measure Information

NQF #: 3443

Corresponding Measures:

De.2. Measure Title: All-cause emergency department utilization rate for Medicaid beneficiaries with complex care needs and high costs (BCNs)

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services, Centers for Medicaid & CHIP Services

De.3. Brief Description of Measure: All-cause emergency department (ED) utilization rate for adult Medicaid beneficiaries who meet BCN population eligibility criteria. The measure is calculated as the number of ED visits per 1,000 beneficiary months and is intended to be reported at the state level.

For the purpose of this measure, the BCN population is defined as Medicaid beneficiaries who are age 18 to 64 during the lookback year (the 12 months prior to the measurement year) and the measurement year and have at least one inpatient admission and at least two chronic conditions, as defined by the Chronic Conditions Data Warehouse (CCW). Beneficiaries dually enrolled in Medicaid and Medicare and beneficiaries who had fewer than 10 months of Medicaid eligibility in the lookback year are not included in the analytic sample because we did not have enough utilization data to include them in testing. We further limited the analytic file to beneficiaries that met BCN definition criteria.

1b.1. Developer Rationale: The BCN population is heterogeneous, with varying medical, behavioral, and psychosocial care needs. Consistent across the BCN population, however, is a pattern of health care consumption characterized by a disproportionately high use of inpatient and ED services, often coupled with underutilization of preventive and other types of outpatient care. We also see significant variation in the BCN-1 measure by beneficiary subgroups (see 1b.4). Frequent ED utilization among BCNs--especially for ambulatory care sensitive conditions--are very costly and may signal poor access to primary care, suboptimal care coordination, and/or lack of supportive services across transitions in care settings (Billings and Raven 2013; Capp et al. 2013; Doupe et al. 2012; Pukurdpol et al. 2014). Therefore, measuring and subsequently reducing ED utilization through improved care coordination and access to community-based care represents an opportunity to improve both quality and cost of care for BCNs (Durand et al. 2011; Utah Office of Health Care Statistics 2004). This ED utilization measure should be paired with the all-cause inpatient admission measure (BCN-2) under development by CMS/Mathematica to assess overall hospital-based care. Both measures are intended to be used for state-level monitoring and quality improvement activities.

References:

Billings, J., and M.C. Raven. “Dispelling an Urban Legend: Frequent Emergency Department Users Have Substantial Burden of Disease.” Health Affairs, vol. 32, no. 12, 2013, pp. 2099–2108.

Capp, R., M.S. Rosenthal, M.M. Desai, L. Kelley, C. Borgstrom, D.L. Cobbs-Lomax, P. Simonette, and E.S. Spatz “Characteristics of Medicaid Enrollees with Frequent ED Use.” American Journal of Emergency Medicine, vol. 31, no. 9, 2013, pp. 1333–1337.

Doupe, M.B., W. Palatnick, S. Day, D. Chateau, R.A. Soodeen, C. Burchil, and S. Derksen. “Frequent Users of Emergency Departments: Developing Standard Definitions and Defining Prominent Risk Factors.” Annals of Emergency Medicine, vol. 60, no. 1, 2012, pp. 24–32.

Durand, A.C., S. Gentile, B. Devictor, S. Palazzolo, P. Vignally, P. Gerbeaux, and R. Sambuc. “ED Patients: How Nonurgent Are They? Systematic Review of the Emergency Medicine Literature.” American Journal of Emergency Medicine, vol. 29, no. 3, 2011, pp. 333–345.

Pukurdpol, P., J.L. Wiler, R.Y. Hsia, and A.A. Ginde. “Association of Medicare and Medicaid Insurance with Increasing Primary Care-Treatable Emergency Department Visits in the United States.” Academic Emergency Medicine, vol. 21, no.10, 2014, pp. 1135–1142.

Utah Office of Health Care Statistics. “Primary Care Sensitive Emergency Department Visits in Utah, 2001.” Salt Lake City: Utah Department of Health, 2004. Available at http://utah.ptfs.com/awweb/guest.jsp?smd=1&cl=all_lib&lb_document_id=12114. Accessed July 26, 2016.

S.4. Numerator Statement: The number of ED visits in the measurement year among adult Medicaid beneficiaries who meet BCN population eligibility criteria.

S.6. Denominator Statement: Number of Medicaid-eligible months ("beneficiary months") among adult Medicaid beneficiaries who meet BCN population eligibility criteria.

S.8. Denominator Exclusions: N/A

De.1. Measure Type: Outcome

S.17. Data Source: Claims

S.20. Level of Analysis: Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

3444:All-cause hospital utilization for Medicaid beneficiaries with complex care needs and high costs (BCNs)

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? BCN-1 is part of a paired measure set titled “All-cause hospital utilization for Medicaid BCNs.” The other measure in the pair is all-cause inpatient admission rate for BCNs, referred to as “BCN-2.” The BCN-1 and BCN-2 measures are intended for voluntary use by states to monitor and improve the quality of care provided to the Medicaid BCN population.

Monitoring BCN-1 in conjunction with the paired indicator for inpatient admissions (BCN-2) can identify meaningful differences in overall hospital utilization and help ensure reductions in inpatient admissions accurately reflect true increases in quality care. Examining either measure in isolation has the potential to produce inaccurate inferences about states’ performance due to the potential for the substitution effect. For example, if an accountable entity decreases ED utilization among the BCN population by keeping patients in the hospital overnight, thus shifting the utilization to an inpatient visit, the ED measure concept alone would understate the accountable entity’s overall impact on hospital-based care. Conversely, a successful BCN intervention may decrease inpatient utilization (or the length of an inpatient admission) but increase ED utilization, appearing to have worsened outcomes despite reducing overall rates of hospital-based care. Indeed, several Medicaid BCN initiatives have successfully reduced inpatient admissions but increased ED utilization. For example, Washington State’s Chronic Care Management program resulted in significant reductions in inpatient admissions and spending, and increases in ED utilization, though not at a statistically significant rate (Xing et al. 2015). If measured solely on ED utilization, Washington State might appear to have worsened outcomes, despite reducing overall rates of hospital-based care.

Reference: Xing, J., C. Goehring, and D. Mancuso. “Care Coordination Program for Washington State Medicaid Enrollees Reduced Inpatient Hospital Costs.” Health Affairs, vol. 34, no. 4, 2015, pp. 653-661.

1. Evidence and Performance Gap – 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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

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

BCN-1_Evidence_Attachment_FINAL_BCN_team_09.19.18.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:**

Measure Title: All-cause emergency department utilization rate for Medicaid beneficiaries with complex care needs and high costs (BCNs)

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 9/19/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *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.*
- All information needed to demonstrate meeting the evidence sub-criterion (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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing 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:

- Outcome: 3 Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- 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.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

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 Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

☒ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

☐ Structure:

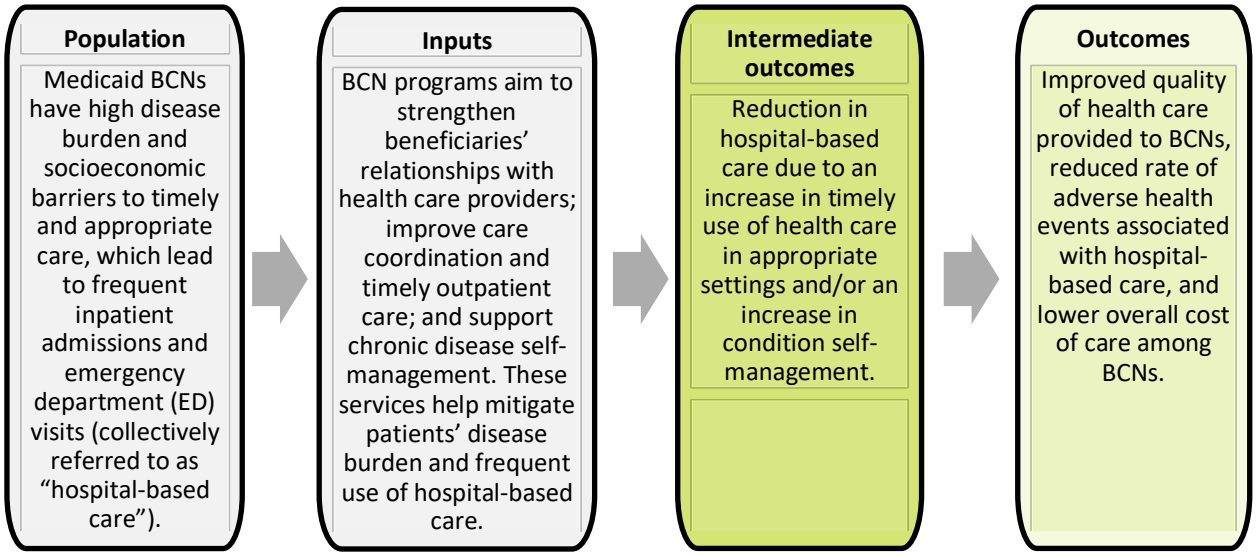
☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Although frequent inpatient admissions and ED visits (collectively referred to as "hospital-based care") are often warranted for high-risk beneficiaries, they could also signal uncoordinated care or a lack of access to care in appropriate settings, such as primary and specialty care practices (Harris et al. 2016; Xing et al. 2015; Agency for Healthcare Research and Quality 2015; Gao et al. 2014; Peikes et al. 2012; Lewis et al. 2012; Misky et al. 2010; Schrag et al. 2006).

To reduce hospital-based utilization, most BCN programs aim to strengthen beneficiaries' relationships with health care providers in the community, improve care coordination across providers (including timely outpatient follow-up care), and support chronic disease self-management. These types of services help mitigate patients' disease burden and frequent need for hospital-based care (Dowd et al. 2014; Tsilimingras et al. 2015; Forster et al. 2003; Grinberg et al. 2016; Friedberg et al. 2010).

Thus, a decrease in the all-cause emergency department utilization rate for Medicaid BCNs may represent an increase in the quality of care for BCNs, including access to appropriate health services, provision of effective care coordination, and improved health-related quality of life outcomes. It may also decrease overall health care costs among BCNs.



1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

Not applicable

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

Evidence suggests that high-risk patients (those with social needs, complex medical problems, multiple chronic conditions, and multiple inpatient admissions) may experience greater benefits (such as improved care, better outcomes, and reduced costs) from care management than the average beneficiary (Harris et al. 2016; Xing et al. 2015; Peikes et al. 2012; Lewis et al. 2012; Hempstead et al. 2014; Wiener et al. 2017). Some evidence suggests that programs designed to reduce cost by decreasing hospital-based care can generate savings among Medicaid beneficiaries (California Medicaid Research Institute 2013; Smulowitz et al. 2013). Reductions in hospital-based care may be able to offset cost increases in other areas, such as primary and specialty care services.

The following results from BCN care management initiatives illustrate the range of potential reductions in hospital-based care and costs. Each example has a unique sample population; although none are identical to the BCN population definition used for this measure, the example populations all overlap this measure's BCN definition to some degree.

Program	Population	Impact on inpatient service utilization	Impact on emergency department service utilization	Impact on costs
Washington State's Chronic Care Management program (Xing et al. 2015)	Clinically complex Medicaid beneficiaries who were identified via risk criteria established through a predictive modeling algorithm (N=907)	Reduced inpatient utilization by 9.64 admissions per 1,000 member months	Increased ED utilization of 10.81 visits per 1,000 member months (not statistically significant)	$318 per member per month reduction in inpatient hospital admission costs
Pilot care coordination intervention at a public hospital in New York City (Bellevue Hospital Center) (Raven et al. 2011)	A small number (N=19) of Medicaid beneficiaries who were considered high risk for hospitalization and experienced homelessness or substance use	Decreased inpatient admissions by 37.5 percent	Decreased the annual number of ED visits by a mean of 0.7 visits	Decreased annual hospital Medicaid reimbursements by 38 percent. The decrease in hospital costs was accompanied by a small increase in outpatient clinic visits.
Pilot care management program operated in Camden, New Jersey, by the Camden Coalition of Healthcare Providers (Green et al. 2010)	Patients with high hospital-based utilization and high costs (N=36), as identified through hospital discharge data. Many patients had complex medical needs, low or no income, and lack of stable shelter	Reduced the total number of monthly visits to hospitals and EDs by roughly 40 percent per month		Decreased overall costs of care of a small sample of patients by 56 percent (Hong et al. 2014)
Coordinated Care Clinic at Hennepin County Medical Center in Minnesota (Mann 2014)	Patients with complex health problems	25 percent decrease in hospitalizations during the first year of the program	38 percent decrease in ED visits during the first year of the program	23 percent reduction in total charges for medical care during the first year of the program
The California Initiative: 6 pilot sites across California (Frequent Users of Health Services Initiative 2008)	Patients had at least 4 ED visits in 12 months, and were also often afflicted with mental illness, homelessness, and/or substance abuse	Inpatient admissions decreased by 14 percent	ED visits decreased by 30 percent	Inpatient spending decreased by 8 percent, and total cost of ED services decreased by 17 percent

Program	Population	Impact on inpatient service utilization	Impact on emergency department service utilization	Impact on costs
Transitional care program operated by Community Care of North Carolina, a large medical home system (DuBard and Jackson 2015)	Non-dual Medicaid recipients with multiple chronic or catastrophic conditions (as defined by 3M Health Information Systems Clinical Risk Group methodology)	Decreased inpatient admissions by 10 percent between 2008 and 2014	Not available.	Decreased the total cost of care by 3 percent between 2009 and 2012 (Hong et al. 2014)

Initiatives to reduce emergency department utilization typically guide patients toward more effective forms of care, such as primary and specialty care providers, while also promoting disease self-management. These avenues provide higher-quality, lower-cost care than tertiary settings, in which care often is fragmented, particularly for people with complex chronic conditions and socioeconomic issues. Specifically, people with a usual source of care (typically, a primary care physician) are more likely than others to receive preventive services, experience greater satisfaction overall with their health care, and have lower rates of both inpatient admissions and ED use for non-urgent conditions (Friedberg et al. 2010). Although this shift in service use could increase outpatient care, particularly primary and specialty care, reliable estimates for the impact of this potential service use shift are scarce. To ensure reductions in ED visits reflect increases in quality care, the ED utilization rate should be monitored alongside indicators for inpatient admissions (also developed by Mathematica/CMS for the Medicaid BCN population), outpatient care, and patients' ability to manage their chronic conditions. Furthermore, analysis of ED utilization should take into account that service use for the identified population might decline over time for purely statistical reasons; this is known as "regression to the mean" and does not signify improvements in health care delivery or outcomes.

References

Agency for Healthcare Research and Quality. "Measures of Care Coordination: Potentially Avoidable Hospitalizations." Chartbook on Care Coordination. Rockville, MD: AHRQ, 2015. Available at https://www.ahrq.gov/research/findings/nhqrdr/2014chartbooks/carecoordination/carecoord-measures3.html. Accessed July 28, 2016.

Billings, J., and T. Mijanovich. "Improving the Management of Care for High-Cost Medicaid Patients." Health Affairs, vol. 26, no. 6, 2007, pp. 1643–1654.

California Medicaid Research Institute. "Emergency Department Visit Reduction Programs: Executive Summary." Prepared for the Medicaid and CHIP Payment and Access Commission. San Francisco: University of California, San Francisco, December 2013.

Dowd, B., M. Karmarker, T. Swenson, S. Parashuram, R. Kane, R. Coulam, and M. Moore Jeffery. "Emergency Department Utilization as a Measure of Physician Performance." American Journal of Medical Quality, vol. 29, no. 2, 2014, pp. 135–143.

DuBard, C.A., and C. Jackson. "Hospitalization Trends in North Carolina Medicaid: Patients with Multiple Chronic Conditions, 2008-2014." Community Care of North Carolina Data Brief, no. 2, 2015.

Friedberg, M., P. Hussey, and E. Schneider. "Primary Care: A Critical Review of the Evidence on Quality and Costs of Health Care." Health Affairs, vol. 29, no. 5, 2010, pp. 766–772.

Frequent Users of Health Services Initiative 2008 Frequent Users of Health Services Initiative. "Summary Report of Evaluation Findings: A Dollars and Sense Strategy to Reducing Frequent Use of Hospital Services." Oakland, CA: California Endowment and California Healthcare Foundation, 2008.

Forster, A.J., H.J. Murff, J.F. Peterson, T.K. Gandhi, and D.W. Bates. "The Incidence and Severity of Adverse Events Affecting Patients after Discharge from the Hospital." Annals of Internal Medicine, vol. 138, no. 3, pp. 161–167.

Gao, J., E. Moran, Y.F. Li, and P.L. Almenoff. "Predicting Potentially Avoidable Hospitalizations." Medical Care, vol. 52, no. 2, 2014, pp. 164–171.

Green, S.R., V. Singh, and W. O'Byrne. "Hope for New Jersey's City Hospitals: The Camden Initiative." Perspectives in Health Information Management, vol. 7, spring 2010.

Grinberg, C., M. Hawthorne, M. LaNoue, J. Brenner, and D. Mautner. "The Core of Care Management: The Role of Authentic Relationships in Caring for Patients with Frequent Hospitalizations." Population Health Management, vol. 19, no. 4, 2016, pp. 248–256.

Harris, L.J., I. Graetz, P.S. Podila, J. Wan, T.M. Waters, and J.E. Bailey. "Characteristics of Hospital and Emergency Care Super-Utilizers with Multiple Chronic Conditions." Journal of Emergency Medicine, vol. 50, no. 4, 2016, pp. 203–214.

Hempstead, K., D. Delia, J.C. Cantor, T. Nguyen, and J. Brenner. "The Fragmentation of Hospital Use Among a Cohort of High Utilizers: Implications for Emerging Care Coordination Strategies for Patients with Multiple Chronic Conditions." Medical Care, vol. 52, suppl. 3, 2014, pp. S67–S74.

Hong, C.S., A. Siegel, and T. Ferris. "Caring for High-Need, High-Cost Patients: What Makes for a Successful Care Management Program?" Commonwealth Fund, vol. 19, no. 1764, 2014.

Lewis, V.A., B.K. Larson, A.B. McClurg, R.G. Boswell, and E.S. Fisher. "The Promise and Peril of Accountable Care for Vulnerable Populations: A Framework for Overcoming Obstacles." Health Affairs, vol. 31, no. 8, 2012.

Mann, C. "CMCS Informational Bulletin: Reducing Nonurgent Use of Emergency Departments and Improving Appropriate Care in Appropriate Settings." Baltimore: Center for Medicaid & CHIP Services, January 16, 2014. Available at https://www.medicaid.gov/federal-policy-guidance/downloads/cib-01-16-14.pdf. Accessed August 28, 2018.

Misky, G.J., H.L. Wald, and E.A. Coleman. "Post-Hospitalization Transitions: Examining the Effects of Timing of Primary Care Provider Follow-Up." Journal of Hospital Medicine, vol. 5, no. 7, 2010, pp. 392–397.

Peikes, D., A. Zutshi, J. Genevro, K. Smith, M. Parchman, and D. Meyers. "Early Evidence on the Patient-Centered Medical Home." AHRQ Publication no. 12-0020-EF. Rockville, MD: Agency for Healthcare Research and Quality, February 2012.

Raven, M.C., K.M. Doran, S. Kostrowski, C. C. Gillespie, and B.D. Elbel. "An Intervention to Improve Care and Reduce Costs for High-Risk Patients with Frequent Hospital Admissions: A Pilot Study." BMC Health Services Research, vol. 11, no. 270, 2011.

Schrag, D., F. Xu, M. Hanger, E. Elkin, N. Bickell, and P. Bach. "Fragmentation of Care for Frequently Hospitalized Urban Residents." Medical Care, vol. 44, no. 6, 2006, pp. 560–567.

Soril, L., L. Leggett, D. Lorenzetti, T. Noseworthy, and F. Clement. "Reducing Frequent Visits to the Emergency Department: A Systematic Review of Interventions." PLoS One, vol. 10, no. 4, 2015.

Smulowitz, P.B., L. Honigman, and B.E. Landon. "A Novel Approach to Identifying Targets for Cost Reduction in the Emergency Department." Annals of Emergency Medicine, vol. 61, no. 3, 2013, pp. 293–300.

Tsilimingras, D., J. Schnipper, A. Duke, J. Agens, S. Quintero, G. Bellamy, J. Janisse, L. Helmkamp, and D.W. Bates. "Post-Discharge Adverse Events among Urban and Rural Patients of an Urban Community Hospital:

A Prospective Cohort Study." Journal of General Internal Medicine, vol. 30, no. 8, August 2015, pp. 1164–1171.

Wiener, J.M., M. Romaire, N. Thach, A. Collins, K. Kim, H. Pan, G. Chiri, A. Sommers, S. Haber, M. Musumeci, and J. Paradise. "Characteristics of Hospital Stays for Nonelderly Medicaid Super-Utilizers, 2012." Kaiser Family Foundation, Issue Brief June 2017. Available at https://www.kff.org/medicaid/issue-brief/strategies-to-reduce-medicaid-spending-findings-from-a-literature-review.

Xing, J., C. Goehring, and D. Mancuso. "Care Coordination Program for Washington State Medicaid Enrollees Reduced Inpatient Hospital Costs." Health Affairs, vol. 34, no. 4, 2015, pp. 653–661.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses **explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)**

Not applicable.

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 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.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

Not applicable.

1a.4.2 What process was used to identify the evidence?

Not applicable.

1a.4.3. Provide the citation(s) for the evidence.

Not applicable.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

The BCN population is heterogeneous, with varying medical, behavioral, and psychosocial care needs. Consistent across the BCN population, however, is a pattern of health care consumption characterized by a disproportionately high use of inpatient and ED services, often coupled with underutilization of preventive and other types of outpatient care. We also see significant variation in the BCN-1 measure by beneficiary subgroups (see 1b.4). Frequent ED utilization among BCNs--especially for ambulatory care sensitive conditions--are very costly and may signal poor access to primary care, suboptimal care coordination, and/or lack of supportive services across transitions in care settings (Billings and Raven 2013; Capp et al. 2013; Doupe et al. 2012; Pukurdpol et al. 2014). Therefore, measuring and subsequently reducing ED utilization through improved care coordination and access to community-based care represents an opportunity to improve both quality and cost of care for BCNs (Durand et al. 2011; Utah Office of Health Care Statistics 2004). This ED utilization measure should be paired with the all-cause inpatient admission measure (BCN-2) under development by CMS/Mathematica to assess overall hospital-based care. Both measures are intended to be used for state-level monitoring and quality improvement activities.

References:

Billings, J., and M.C. Raven. "Dispelling an Urban Legend: Frequent Emergency Department Users Have Substantial Burden of Disease." Health Affairs, vol. 32, no. 12, 2013, pp. 2099–2108.

Capp, R., M.S. Rosenthal, M.M. Desai, L. Kelley, C. Borgstrom, D.L. Cobbs-Lomax, P. Simonette, and E.S. Spatz "Characteristics of Medicaid Enrollees with Frequent ED Use." American Journal of Emergency Medicine, vol. 31, no. 9, 2013, pp. 1333–1337.

Doupe, M.B., W. Palatnick, S. Day, D. Chateau, R.A. Soodeen, C. Burchil, and S. Derksen. "Frequent Users of Emergency Departments: Developing Standard Definitions and Defining Prominent Risk Factors." Annals of Emergency Medicine, vol. 60, no. 1, 2012, pp. 24–32.

Durand, A.C., S. Gentile, B. Devictor, S. Palazzolo, P. Vignally, P. Gerbeaux, and R. Sambuc. "ED Patients: How Nonurgent Are They? Systematic Review of the Emergency Medicine Literature." American Journal of Emergency Medicine, vol. 29, no. 3, 2011, pp. 333–345.

Pukurdpol, P., J.L. Wiler, R.Y. Hsia, and A.A. Ginde. "Association of Medicare and Medicaid Insurance with Increasing Primary Care-Treatable Emergency Department Visits in the United States." Academic Emergency Medicine, vol. 21, no.10, 2014, pp. 1135–1142.

Utah Office of Health Care Statistics. "Primary Care Sensitive Emergency Department Visits in Utah, 2001." Salt Lake City: Utah Department of Health, 2004. Available at http://utah.ptfs.com/awweb/guest.jsp?smd=1&cl=all_lib&lb_document_id=12114. Accessed July 26, 2016.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

We included data from 10 states in measure testing. These states had the most current Medicaid Analytic eXtract (MAX) data available at the time of measure testing and met data quality standards (see Testing Attachment Appendix). We included data on FFS beneficiaries from seven states and managed-care beneficiaries from three states. In this document, state names are redacted and referred to as State A through State J. The measurement year is 2014. Performance rates are reported as number of inpatient admissions per 1,000 beneficiary months. Measure testing details are included in the testing attachment.

Table 1. Unadjusted and risk-adjusted measure rates by state

State	Unadjusted BCN-1 rate	Adjusted BCN-1 rate
State A 263.5	270.2*	
State B	282.5	322.0*
State C	244.7	231.4
State D	186.0	194.2^
State E	243.1	275.1*
State F	167.9	180.5^
State G	88.5	109.5^
State H	239.6	233.6
State I	263.8	241.3
State J	271.1	280.7*
Overall	234.2	234.0

Source: Mathematica analysis of 2014 MAX PS, LT, OT, and IP files.

Sample: Full sample (N = 142,193)

^ Measure rate is significantly lower than the overall rate (95% CI)

* Measure rate is significantly higher than the overall rate (95% CI)

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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

We included data from 10 states in measure testing. These states had the most current MAX data available at the time of measure testing and met data quality standards (see Testing Attachment Appendix). We included data on FFS beneficiaries from seven states and managed-care beneficiaries from three states. In this document, state names are redacted and referred to as State A through State J. The measurement year is 2014. Performance rates are reported as number of inpatient admissions per 1,000 beneficiary months. Measure testing details are included in the testing attachment.

Table 2. Risk-adjusted measure rate distribution by population subgroup

Subgroup	Mean	25th pctl	50th pctl	75th pctl
Aged 18 to 24	246.9	187.2	248.2	262.6
Aged 25 to 44	251.6	219.9	255.6	285.1
Aged 45 to 64	214.9	187.2	218.7	258.6
Male	231.1	204.6	229.9	274.2
Female	235.6	179.5	234.3	262.3
White	218.2	189.7	216.9	244.9
Black	273.7	245.2	281.1	303.9
Hispanic	247.4	171.0	204.4	230.3
Other/unknown	192.0	133.3	164.5	213.6
Aged/Blind/ Disabled	251.7	222.4	260.0	312.5
Adult	208.9	176.9	206.7	229.6
Child	225.1	191.3	219.1	236.5
1+ behavioral health conditions	255.7	220.8	256.1	298.3
No behavioral health conditions	149.5	124.5	149.6	162.6

Source: Mathematica analysis of 2013 and 2014 MAX PS, LT, OT, and IP files.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

Not applicable

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 sub criteria 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. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

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.)*

The measure does not yet have published specifications. Therefore no link exists.

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:** BCN-1_Value_Set_Attachment_09.19.18-636735630830856632.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

No

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update 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) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The number of ED visits in the measurement year among adult Medicaid beneficiaries who meet BCN population eligibility criteria.

S.5. 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, time period for data collection, 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 (S.14).

The numerator is calculated as the sum of ED visits in the measurement year. ED visits contribute to the monthly count only if they do not result in an inpatient admission or observation stay (of any length). ED visits and observation stays are identified using Healthcare Effectiveness Data and Information Set (HEDIS) value

sets, which are trademarked by the National Committee for Quality Assurance. Specifically, ED visits are identified using any of the following claim type, revenue code, and procedure code combinations:

1. Outpatient claims with revenue codes in the ED Value Set (HEDIS 2015)
2. Professional claims with CPT codes in the ED Value Set (HEDIS 2015)
3. Professional claims with Place of Service code in the ED POS Value Set (HEDIS 2015) and CPT codes in the ED Procedure Code Value Set (HEDIS 2015)

If an ED visit's dates of service overlap with or are within one calendar day of an inpatient admission date, we do not include it in the numerator count.

Inpatient admissions are identified by using institutional claims for inpatient hospital services.

Observation stays are identified in two ways:

1. Procedure codes in the Observation Value Set (HEDIS 2015)
2. Revenue and procedure codes created by the Centers for Medicare & Medicaid Services (CMS) to identify observation stays. We identify observation stays of any length.

Claims are deduplicated to ensure there is no more than one ED visit per beneficiary per day.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Number of Medicaid-eligible months ("beneficiary months") among adult Medicaid beneficiaries who meet BCN population eligibility criteria.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The denominator is calculated as the number of Medicaid-eligible months during the measurement year among adult beneficiaries who meet BCN population eligibility criteria. The measurement period is 12 months. An additional 12 months of lookback data is needed to identify the BCN population, for a total of 24 months of data.

BCNs are defined as Medicaid beneficiaries who are age 18 to 64 during the lookback and measurement years and who have at least one inpatient admission and at least two chronic conditions (as defined by the CCW) in the lookback year. Inpatient admissions are identified by using institutional claims where type of service = 01 (for "inpatient"). Observation stays are not included in the sum of inpatient admissions used to identify the denominator population.

Beneficiaries dually enrolled in Medicaid and Medicare and beneficiaries who had fewer than 10 months of Medicaid eligibility in the lookback year are not in the analytic sample because we did not have enough utilization data to include them in testing.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

N/A

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and*

the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

UNADJUSTED RATE

Step 1: Determine eligible denominator population and number of eligible beneficiary months among BCNs.

BCNs are defined as Medicaid beneficiaries who are age 18 to 64 during the lookback and measurement years and who have at least one inpatient admission and at least two chronic conditions (defined by the CCW), in the lookback year.

Inpatient admissions are identified by using institutional claims where type of service = "01" (for "inpatient"). Observation stays are not included in the sum of inpatient admissions used to identify the denominator population. CCW condition algorithms are publicly available at https://www.ccwdata.org; they are also available in the BCN-1 Value Set Attachment.

An eligible beneficiary month is one in which the beneficiary is enrolled in Medicaid fee-for-service (FFS) or managed care. Sum eligible months across all beneficiaries.

Step 2: Determine the number of ED visits among BCNs.

The numerator is calculated as the sum of ED visits that did not result in an inpatient admission or observation stay in the measurement year.

ED visits are identified using any of the following claim type, revenue code, and procedure code combinations:

1. Outpatient claims with revenue codes in the ED Value Set (HEDIS 2015)
2. Professional claims with CPT codes in the ED Value Set (HEDIS 2015)
3. Professional claims with Place of Service code in the ED POS Value Set (HEDIS 2015) and CPT codes in the ED Procedure Code Value Set (HEDIS 2015)

If an ED visit's dates of service overlap with or are within one calendar day of an inpatient admission date, we do not include it in the numerator count.

Inpatient admissions are identified by using institutional claims where type of service = 01 (for "inpatient").

Observation stays are identified in two ways:

(1) Procedure codes in the Observation Value Set (HEDIS 2015)

(2) Revenue and procedure codes created by CMS to identify observation stays.

Deduplicate ED visits to ensure there is no more than one ED visit counted per beneficiary per day. Sum unique ED visits across all beneficiaries.

Step 3: Calculate the ED visit rate among BCNs.

Divide the number of ED visits (from Step 2) by the number of enrollee months (from Step 1), and multiply the resulting ratio by 1,000, as follows:

(Number of ED visits/Number of enrollee months) x 1,000 = Unadjusted ED utilization rate

RISK-ADJUSTED RATE

Step 1: Calculate the observed number of ED visits.

Calculate the observed number of ED visits (as described in Unadjusted Rate Step 2) for each beneficiary identified as a BCN. Sum the observed number of inpatient admissions across beneficiaries. This "observed" value will be used as the numerator in the observed-to-expected (O/E) calculation in Step 3.

Step 2: Calculate the expected number of ED visits.

For each beneficiary identified as a BCN,

a) Determine the value of each of the 69 risk factors as described in the BCN-1 Value Set attached to question S.2b.

NOTE: The weights were based on the BCN population used during testing: adult, non-dual Medicaid beneficiaries with FFS or managed care claims data from 10 states in 2013 and 2014 (i.e., the development population). These coefficients will be revised using updated Medicaid claims data at the time of NQF endorsement maintenance review.

b) Multiply each nonzero risk factor value by the weight provided in the BCN-1 Value Set Attachment.

NOTE: The reference category for each factor has a value of zero. For example, beneficiaries who are female (the reference category for sex) would have a beneficiary value of zero for sex when computing the sum of coefficient estimates. Beneficiaries who are male (the included category for sex) have a beneficiary value of one for sex.

c) Sum the products that resulted from multiplying risk factor values and coefficients.

d) Exponentiate the resulting sum (from Step 2.c) and multiply it by the number of enrolled months as follows:

[# months] x e^([sum from Step 2.c]) = # of expected ED visits

Sum the expected number of ED visits (from Step 2.d) across all beneficiaries in the population. This "expected" value will be used as the denominator in the O/E ratio calculation in Step 3.

Step 3: Calculate the state's O/E ratio.

Divide the state's observed number of ED visits (from Step 1) by the state's expected number of ED visits (from Step 2) to obtain this state's O/E ratio.

NOTE: The O/E ratio calculation does not require dividing the observed (numerator) or expected (denominator) counts by the total number of enrolled months because these terms would cancel out in the division.

Step 4: Calculate the state's risk-adjusted ED utilization rate.

Multiply the state's O/E ratio by the national benchmark rate of 234.2 ED visits per 1,000 beneficiary months:

(O/E for state) x (national benchmark rate) = Risk-adjusted ED utilization rate for state

NOTE: The national benchmark BCN-1 rate is equivalent to the unadjusted national BCN-1 rate. This value will change over time and as the BCN population characteristics change.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

N/A

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

N/A

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Medicaid claims data: person-summary (PS), inpatient (IP), other services (OT), and long-term care (LTC) files

S.19. 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.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Population : Regional and State

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Emergency Department and Services

If other:

S.22. 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

2. Validity – See attached Measure Testing Submission Form

BCN-1_Testing_Attachment_FINAL_07.30.18.docx,BCN-1_Testing_Attachment_Supplemental_Material_SUBMITTED_09.26.18-636736585378121891.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** NA

Measure Title: All-cause emergency department utilization rate for Medicaid beneficiaries with complex care needs and high costs (BCNs)

Note to NQF staff and reviewers: This measure is referred to as "BCN-1." BCN-1 is part of a paired measure set titled "All-cause hospital utilization for Medicaid BCNs." The other measure in the pair is an all-cause inpatient admission rate for BCNs, referred to as "BCN-2." The BCN-1 and BCN-2 measures are intended for voluntary use by states to monitor and improve the quality of care provided to the Medicaid BCN population.

Date of Submission: 8/1/2018

Type of Measure:

☒ Outcome (*including PRO-PM*)	☐ Composite – ***STOP – use composite testing form***
☐ Intermediate Clinical Outcome	☐ Cost/resource
☐ Process *(including Appropriate Use)*	☐ 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, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 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 social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Standing 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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. 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 **instrument-**

based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; 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

2b3. 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 social risk 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.

2b4. 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.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. 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. The degree of consensus and any areas of disagreement must be provided/discussed.

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.17*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

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 Analytic eXtract (MAX) claims to obtain the data elements needed for testing. Specifically, we used the MAX person summary (PS), inpatient (IP), long-term care (LT), and other therapy (OT) files. The PS, IP, LT, and OT files served as the primary source of information for the measure denominator, and the IP and OT files enabled us to identify the numerator events. The PS file contained additional demographic and enrollment information, such as age, sex, and race or ethnicity.

1.3. What are the dates of the data used in testing?

We analyzed MAX data from 2013 (the *lookback year*) and 2014 (the *measurement year*). We used data in the *measurement year* to calculate eligible emergency department (ED) visits and eligible enrollment months, and data in the *lookback year* to define the BCN population to be used in the measurement year. The years of data used for testing were based on the most current MAX data available at the time that testing began.

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.20*)	**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	☒ other: state

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 data from 10 states in measure testing. These states had the most current MAX data available at the time of measure testing and met data quality standards (see Testing Attachment Appendix 1). We included data on fee-for-service beneficiaries from seven states and managed-care beneficiaries from three states. In this document, state names are redacted and referred to as State A through State J.

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 measure is intended to be reported among adult Medicaid beneficiaries at the state level. Beneficiaries dually enrolled in Medicaid and Medicare and beneficiaries who had fewer than 10 months of Medicaid eligibility in the previous 12 months are not included in the analytic sample because we did not have enough utilization data to include them in testing. We further limited the analytic file to beneficiaries that met BCN definition criteria. For the purpose of this measure, BCNs are defined as Medicaid beneficiaries who have at least one inpatient admission and at least two chronic conditions, as defined by the Chronic Conditions Data Warehouse (CCW)[1], in the past 12 months (the lookback year). The final analytic sample comprises 142,193 beneficiaries classified as BCNs across 10 states (Table 1).

Table 1. Analytic sample: Geographical distribution and sociodemographic composition

Characteristic	Number of BCNs (N = 142,193)	Percentage of BCN population
State		
State A	19,862	14.0
State B	201	0.1
State C	40,214	28.3
State D	19,054	13.4
State E	15,271	10.7
State F	11,443	8.0
State G	120	0.1
State H	27,228	19.1
State I	8,025	5.6
State J	775	0.5
Gender		
Male	49,798	35.0
Female	92,395	65.0
Age		
18-24	17,781	12.5
25-44	58,949	41.5
45-64	65,463	46.0

[1] The CCW contains 44 physical health conditions and 18 behavioral health conditions. The CCW algorithm calls for lookback periods of up to three years for some conditions. However, we used a consistent, one-year reference period across all conditions to avoid excluding enrollees on the basis of a more stringent, longer continuous-enrollment requirement.

Characteristic	Number of BCNs (N = 142,193)	Percentage of BCN population
Race/ethnicity		
White, non-Hispanic origin	82,536	58.0
Black, non-Hispanic origin	41,627	29.3
Hispanic	10,647	7.5
Other/unknown	7,383	5.2
Eligibility category		
Aged/blind/disabled	80,569	56.7
Adult	58,450	41.1
Child	3,174	2.2
Plan type		
Fee-for-service	59,480	41.8
Managed care	82,713	58.2

Source: Mathematica analysis of 2013 and 2014 MAX PS, LT, OT, and IP files.

BCN = Beneficiaries with complex care needs and high costs

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 social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

As shown in section 1.6, we collected information on the following variables from MAX 2013 and 2014 files: Medicaid eligibility category, age, sex, and race/ethnicity. We included most of these factors in risk adjustment (see section 2b3) and assessed disparities in performance rate for key subgroups (see section 2b4).

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*)

We conducted signal-to-noise (SNR) reliability for BCN-1. The SNR statistic, *R* (ranging from 0 to 1), summarizes the proportion of the variation between entity scores that is due to real differences in underlying quality of care as opposed to random variation (for example, due to measurement or sampling error). If *R*=0, there is no variation on the measure across entities, and all observed variation is due to sampling variation. In this case, the measure is not useful to distinguish between entities with respect to healthcare quality. Conversely, if *R*=1,

all entity scores are free of sampling error, and all variation represents real differences between entities in the measure result.

To calculate the SNR reliability for the risk-adjusted BCN-1 measure, we first estimated the "noise" (within-state variability) by calculating the variance of the $\frac{\sum O_i}{\sum E_i} \cdot \bar{Y}$ within each state, where the randomness is contributed by the observed inpatient stays of each beneficiary within the state. We next estimated the "signal" (between-state variance) iteratively, using a maximum likelihood estimation approach by Morris.[2] We computed the SNR statistic, *R*, as the ratio of the signal variance (which is common across all entities) to the sum of the signal variance and the noise variance (which varies by entity): $R = \frac{\sigma^2_{between}}{\sigma^2_{between}+\sigma^2_{within}}$

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 risk-adjusted BCN-1 measure is shown to be highly reliable, with an overall (mean) signal-to-noise measure reliability of 0.99. The SNR ranged from 0.59 to 0.99 across the ten states in the sample.

Table 2. Performance and signal-to-noise reliability of the risk-adjusted BCN-2 measure, by state

State	Risk-adjusted BCN-1 performance (inpatient admissions per 1,000 member-months)	Risk-adjusted BCN-2 signal-to-noise reliability
State A	270.2	0.99
State B	322.0	0.59
State C	231.4	0.99
State D	194.2	0.99
State E	275.1	0.99
State F	180.5	0.99
State G	109.5	0.66
State H	233.6	0.99
State I	241.3	0.99
State J	280.7	0.93
Overall (mean)[a]	**234.0**	**0.92**

Source: Mathematica analysis of 2013 and 2014 MAX PS, LT, OT, and IP files.

Note: The SNR coefficients for States A, C, D, E, F, H, and I were truncated to 0.99 rather than rounded to 1.00 to reflect the uncertainty in the estimates.

[a] The overall signal-to-noise measure reliability is estimated by first calculating the mean noise across all states and using that as the overall noise to derive the final SNR reliability. The results are close to what is calculated using the mean value of the state-specific reliability across states, but less sensitive to outliers.

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 risk-adjusted BCN-2 measure is highly reliable overall and for each state in the sample. The overall (mean) signal-to-noise measure reliability of 0.92 is higher than the reliability threshold of 0.9 in the literature, of which one can discern the performance differences between individual reporting entities.[3]

[2] Morris, C. N. (1983) Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association, 78:381, 47-55

[3] Adams, J.L. 2009. The Reliability of Provider Profiling. A Tutorial. http://www.rand.org/pubs/technical_reports/TR653.html (accessed February 23, 2017)

2b1. VALIDITY TESTING

2b1.1. 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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

We were unable to conduct empirical validity testing (such as the convergent validity test conducted for the paired measure, BCN-2) because an appropriate external comparison measure could not be found.[4] Instead, we conducted a systematic face validity analysis among technical expert panel (TEP) members who had guided the development of the measure and were therefore familiar with its specifications and rationale.

2b1.2. 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)*

BCN-1 is intended for voluntary use by states to monitor and improve the quality of care provided to the Medicaid BCN population. To assess face validity of BCN-1, we conducted an online survey of TEP members to obtain their assessment of the measure's components and the extent to which the measure's state-level performance scores distinguish good quality from poor quality of care. Specifically, we asked TEP members the extent to which they agreed or disagreed with the following three statements:

1. "BCN-1 denominator is appropriate given the intent of the measure."

2. "BCN-1 numerator is appropriate given the intent of the measure."

3. "In the future, performance scores on BCN-1 will distinguish between good and poor performance.

The four response options included strongly agree, agree, disagree, and strongly disagree. A comment box was also available for TEP members to describe reasons for their answers.

Of the 17 TEP members who received the face validity survey, 11 responded to the survey and 6 did not respond, resulting in a response rate of 65%. Two additional TEP members were not reachable during the fielding period and therefore did not respond.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

We received 11 responses from TEP members regarding the ability of BCN-1 to distinguish good from poor performance. Most of the respondents (82%) either agreed or strongly agreed that the BCN-1 denominator is appropriate. Among the two respondents who disagreed with the statement, one misunderstood the BCN population definition and one believed that the definition of an eligible month should be clearer. All respondents either agreed or strongly agreed that the BCN-1 numerator is appropriate. Finally, the majority of respondents (70%)[5] either agreed or strongly agreed that BCN-1 would distinguish between good and poor

[4] Reasons that no appropriate external measures could be found for convergent validity testing include incompatible level of measurement (e.g., plan instead of state), incompatible population (e.g., beneficiaries in Medicare home health programs are substantially different from the BCN population), incompatible data source (e.g., measures derived from survey data), and/or lack of validation (e.g., concepts related to all-cause ED visits, such as all-cause primary care utilization, are not yet validated).

[5] Only ten respondents replied to the third question.

performance. Among those who disagreed, one misunderstood the BCN population definition and two did not give a reason.

2b1.4. 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 level of accord in support of BCN-1 suggests that the measure has good face validity for monitoring quality improvement. In addition, the sources of disagreement were minor and could be resolved through clearer instructions.

2b2. EXCLUSIONS ANALYSIS

☒ **no exclusions — *skip to section*** 2b3

2b2.1. 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*)

Not applicable.

2b2.2. 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.

2b2.3. 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.

2b3. 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 2b4***.***

2b3.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with** 69 **risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

We began with a graphical inspection of the outcome measure and its bivariate relationships with potential risk factors. We found that the BCN-1 outcome exhibited a high prevalence of zero values (specifically, 37.3 percent of beneficiaries in the sample did not have an ED visit in the measurement period); the relationship between age and ED visit was quadratic; and the number of ED visits was appreciably higher for beneficiaries with both physical and behavioral health conditions. We accounted for each of these findings in the model development work.

After the exploratory data analysis, we split the analytic sample into two randomly selected half samples. One served as the development sample supporting our model building and exploration work, and the other served

as the validation sample against which we assessed the final model's performance.[6] Finding that our model performs well on the validation sample provides assurance that the model will generalize well to other samples, and is not primarily driven by such idiosyncratic fluctuations.

We also plotted the relationship between the number of emergency department visits and the number of months enrolled to determine whether to include an offset term in our final model. The plot indicated an approximately constant rate, providing reassurance regarding the appropriateness of the offset assumption. After this assessment, we chose the negative binomial regression with an offset term reflecting the number of enrolled months per beneficiary. This method fit the data well and exhibited fewer potential computational challenges for implementers relative to alternative methods. Specifically, the observed number of inpatient admissions for a beneficiary i, denoted by O_i, follows a negative binomial distribution: $O_i \sim NB\left(\frac{1}{k}, \frac{m_i}{k^{-1}+m_i}\right)$ where k is called the dispersion parameter and m_i is the expected number of inpatient admissions for beneficiaries with the same risk factor values as beneficiary i: $m_i = E(O_i|t_i, \beta) = t_i \cdot \exp\{\beta X_i\}$. t_i is the number of eligible BCN months during the measure period, X_i is the vector of the risk factor values for beneficiary i, and β is the vector of coefficients for the risk factors.

The final BCN-1 risk adjustment model included 69 risk factors and an intercept term (Table 4). We included sociodemographic indicators (mean-centered age and its square, sex, and Medicaid eligibility category), along with the entire set of CCW condition indicators in the risk adjustment model. We also included variables indicating whether a beneficiary had physical health conditions only (reference), behavioral health conditions only, or both types of conditions (an interaction of the previous two).

Table 4. Final model specification: Risk factor weights (raw coefficients)

Risk factor	Beta
Intercept	-2.474
2014 centered age	-0.021
2014 centered age squared	0.000*
Aged/blind/disabled	0.141
Child	0.034
Male	-0.052
Acquired hypothyroidism	-0.071
Acute myocardial infarction	-0.088
ADHD, conduct disorders, hyperkinetic syndrome	0.047
Alcohol use disorder	0.304
Alzheimer's disease	-0.205
Alzheimer's disease and related disorders	-0.096
Anemia	0.091
Anxiety disorders	0.229
Asthma	0.284
Atrial fibrillation	0.041
Autism spectrum disorders	-0.274
Benign prostatic hyperplasia	0.223
Bipolar disorder	0.131
Cataract	-0.039

[6] This approach is standard practice to avoid "overfitting" a risk adjustment model, which takes place when a model fits both the true underlying relationships between variables as well as idiosyncratic data fluctuations specific to the particular sample.

Risk factor	Beta
Cerebral palsy	-0.175
Chronic kidney disease	0.125
Colorectal cancer	-0.042
COPD and bronchiectasis	0.166
Cystic fibrosis, other metabolic disorders	-0.018
Depression	0.089
Depressive disorders	0.052
Diabetes	0.171
Drug use disorder	0.214
Endometrial cancer	0.218
Epilepsy	0.236
Female/male breast cancer	-0.091
Fibromyalgia, chronic pain, and fatigue	0.488
Glaucoma	-0.028
Heart failure	0.050
Hip/pelvic fracture	-0.133
HIV/AIDS	0.050
Hyperlipidemia	-0.064
Hypertension	0.146
Intellectual disabilities and related conditions	-0.036
Ischemic heart disease	0.173
Learning disabilities	-0.152
Leukemias and lymphomas	0.003
Liver disease, cirrhosis, other liver conditions	0.202
Lung cancer	0.034
Migraine and chronic headache	0.460
Mobility impairments	-0.065
Multiple sclerosis and transverse myelitis	-0.044
Muscular dystrophy	-0.175
Obesity	0.037
Osteoporosis	-0.014
Other developmental delays	0.033
Peripheral vascular disease	-0.073
Personality disorders	0.160
Posttraumatic stress disorder	-0.028
Pressure and chronic ulcers	0.008
Prostate cancer	0.024
Rheumatoid arthritis, osteoarthritis	0.081
Schizophrenia	-0.083
Schizophrenia, other psychotic disorders	0.187
Sensory—blindness and visual impairment	0.074
Sensory—deafness and hearing impairment	-0.008
Spina bifida, other congenital anomalies	-0.037

Risk factor	Beta
Spinal cord injury	0.137
Stroke/transient ischemic attack	0.019
Tobacco use	0.158
Traumatic brain Injury	0.115
Viral hepatitis	0.134
Has only behavior condition, no physical condition	-0.082
Has both physical and behavior condition	0.058

Sources: Mathematica analysis of 2013 and 2014 MAX PS, LT, OT, and IP files. Full sample (N = 142,193).

Note: The values in the beta column represent the raw regression coefficients generated by the risk adjustment model. These values are often referred to as "risk adjustment weights," and may be used to compute beneficiary-level risk scores.

* Denotes rounded value.

2b3.2. If an outcome or resource use component 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.

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

The choice of predictors was guided by Andersen's Behavioral Model of Health Services Use, which frames the determinants of health care utilization into the following: factors such as demographic characteristics like age and sex that predispose individuals to use care; factors such as income and distance to a clinic that enable individuals to seek care; and factors such as the presence of chronic conditions and/or functional limitations that drive individuals to need care.[7] Andersen's model also treats health care utilization as a measure of "realized access" to care, signifying that individuals were able to overcome any perceived barriers to care receipt.

We selected risk factors for inclusion in the risk adjustment model based on the four criteria: (1) likely predictive importance, (2) feasibility, (3) appropriate care incentives, and (4) ability to construct the factor using claims and encounter data in MAX files.

Likely predictive importance was assessed by reviewing previous risk adjustment algorithms and the related health services literature. Using this criterion, we excluded area-level socioeconomic status (SES). This decision was motivated by the findings from a recent two-year National Quality Forum (NQF) effort, which indicated that the inclusion of area-level SES indicators did not improve the predictive capacity of risk adjustment algorithms of hospital-based care measures developed for Medicare beneficiaries.[8] Feasibility reflects a project team assessment regarding whether constructing the measure would entail a reasonable scope of work, given time and resource constraints. Polypharmacy was excluded under this criterion, as extracting and cleaning the requisite MAX pharmacy data would require significant additional resources and time. Finally, we

[7] Andersen, R.M. "Revisiting the Behavioral Model and Access to Medical Care: Does It Matter?" Journal of Health and Social Behavior, vol. 36, no. 1, March 1995, pp. 1–10.

[8] National Quality Forum. "Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement." Washington, DC: National Quality Forum, August 2017. Available at http://www.qualityforum.org/Show_Content.aspx?id=322. Accessed February 6, 2018.

assessed whether the inclusion of a measure was consistent with the aim of setting appropriate care incentives. For example, including prior inpatient admissions or ED visits would likely increase the predictive capacity of the risk adjustment model, but at the cost of rewarding poorly performing entities with a lower bar for expected performance in future years. Specifically, if beneficiaries with prior inpatient admissions or ED visits are more likely to have an inpatient admission during the measurement year, including prior inpatient admissions or ED visits in the model improves its ability to predict inpatient admissions in the measurement year. The problem is that it increases the predicted inpatient admission rate for entities whose beneficiaries previously had more inpatient admissions or ED visits. Thus, entities with previous high inpatient admission rates are expected to have higher inpatient admission rates in the measurement year compared with entities whose patients did not, effectively setting a lower standard for entities with previous poor performance. Therefore, we excluded prior hospital-based care utilization from consideration.

The subsequent model development and testing was limited to the following variables that were rated "high" across all three categories: sex, age, eligibility category, and the presence of chronic conditions. We did not include race in the model for two reasons: the data are of poor quality for many states[9] and, as described in "A Blueprint for the CMS Measures Management System" (subsequently referred to as "the Blueprint"), the inclusion of race can potentially mask important disparities across racial/ethnic groups.[10]

We also examined the model fit for specifications accounting for comorbidities across physical and behavioral health conditions, as earlier beta-testing activities indicated that beneficiaries with comorbidities across physical and behavioral domains exhibit uniquely high levels of hospital-based care utilization and health care costs.[11] The model fit improved when we accounted for interactions between physical and behavioral health comorbidities.

We convened a risk adjustment expert workgroup to review the risk factors, model development, and approach, and the workgroup agreed with the final set of risk factors and risk adjustment model.

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

☒ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

We estimated the model separately for the development and the validation half samples, in addition to the pooled sample (Table 5). For ease of interpretation, we present model coefficients as incident rate ratios (IRR); IRRs less than one indicate that a risk factor is associated with a lower risk of the outcome, and IRRs greater

[9] Ruttner, L., R. Borck, J. Nysenbaum, and S. Williams. "Guide to MAX Data." Medicaid Policy Brief No. 21. Prepared for the Centers for Medicare & Medicaid Services. Mathematica Policy Research, August 2015. Available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/Downloads/MAX_IB21_MAX_Data_Guide.pdf. Accessed November 30, 2016.

[10] Centers for Medicare & Medicaid Services. "Blueprint for the CMS Measures Management System, Version 13.0." Baltimore, MD: Centers for Medicare & Medicaid Services, May 2017. Available at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/Downloads/Blueprint-130.pdf. Accessed September 4, 2017.

[11] Hula, L., C. Stepanczuk, L. Leininger, B. Marder, L. Hughey, A. Collins, C. Bobst, E. Geil, M. Head, A. Keshaviah, X. Lin, L. Lu, K. Miller, S. Nelson, D. Poznyak, M. Smith, K. Sredl, A. B. Tehrani, F. Xing, and B. Yu. "All-Cause Emergency Department Utilization Rate for Medicaid Beneficiaries with Complex Care Needs and High Costs (BCN-1) Measure Testing Report." Princeton, NJ: Mathematica Policy Research, 2017.

than one indicate that a risk factor is associated with a greater risk of the outcome. The coefficient magnitudes are comparable across the development and validation samples. The risk factors exhibiting coefficient instability (for example, colorectal cancer and glaucoma) typically exhibited very low sample prevalence and imprecisely estimated null associations with the outcome. As a result, it is reasonable to conclude that coefficient differences were driven by statistical noise as opposed to potentially more worrisome overfitting bias.

Fourteen conditions were associated with statistically significant decreases in ED utilization. In general, these risk factors (including several cancers) are accompanied by very serious levels of morbidity; as discussed above, beneficiaries with these conditions may require care in an inpatient versus an ED setting.

Table 5. Final model specification: Risk factor prevalence and incident rate ratios

	Development sample n = 71,096			Validation sample n = 71,097			Full sample N = 142,193		
Risk factor	% of bene-ficiaries	IRR	IRR (95% CI)	% of bene-ficiaries	IRR	IRR (95% CI)	% of bene-ficiaries	IRR	IRR (95% CI)
2014 centered age		0.98	(0.98, 0.98)*		0.98	(0.98, 0.98)*		0.98	(0.98, 0.98)*
2014 centered age squared		1.00	(1.00, 1.00)*		1.00	(1.00, 1.00)*		1.00	(1.00, 1.00)*
Aged/blind/disabled	56.41	1.13	(1.10, 1.16)	56.91	1.18	(1.15, 1.21)	56.66	1.15	(1.13, 1.17)
Child	2.26	1.04	(0.97, 1.13)	2.20	1.03	(0.95, 1.10)	2.23	1.03	(0.98, 1.09)
Male	34.90	0.95	(0.93, 0.97)	35.14	0.95	(0.92, 0.97)	35.02	0.95	(0.93, 0.97)
Acquired hypothyroidism	9.76	0.92	(0.89, 0.95)	9.83	0.95	(0.91, 0.98)	9.79	0.93	(0.91, 0.95)
Acute myocardial infarction	1.75	0.84	(0.77, 0.92)	1.82	0.99	(0.91, 1.07)	1.78	0.92	(0.86, 0.97)
ADHD, conduct disorders, hyper-kinetic syndrome	4.95	1.06	(1.01, 1.11)	4.94	1.04	(0.99, 1.09)	4.94	1.05	(1.01, 1.08)
Alcohol use disorder	15.40	1.37	(1.33, 1.42)	15.58	1.34	(1.29, 1.38)	15.49	1.35	(1.32, 1.39)
Alzheimer's disease	0.28	0.81	(0.65, 1.01)	0.29	0.83	(0.67, 1.03)	0.29	0.81	(0.70, 0.95)
Alzheimer's disease and related disorders	2.17	0.96	(0.89, 1.04)	2.17	0.85	(0.79, 0.92)	2.17	0.91	(0.86, 0.96)
Anemia	30.37	1.09	(1.07, 1.12)	30.23	1.10	(1.07, 1.12)	30.30	1.10	(1.08, 1.11)
Anxiety disorders	29.94	1.25	(1.22, 1.28)	30.33	1.27	(1.23, 1.30)	30.14	1.26	(1.23, 1.28)
Asthma	20.80	1.31	(1.28, 1.35)	20.95	1.34	(1.31, 1.37)	20.88	1.33	(1.30, 1.35)
Atrial fibrillation	3.38	1.07	(1.01, 1.13)	3.27	1.02	(0.96, 1.08)	3.32	1.04	(1.00, 1.09)
Autism spectrum disorders	0.77	0.71	(0.63, 0.81)	0.85	0.80	(0.71, 0.90)	0.81	0.76	(0.70, 0.83)
Benign prostatic hyperplasia	0.66	1.34	(1.19, 1.52)	0.69	1.15	(1.02, 1.31)	0.68	1.25	(1.14, 1.36)
Bipolar disorder	21.31	1.14	(1.10, 1.17)	21.43	1.15	(1.11, 1.18)	21.37	1.14	(1.12, 1.16)
Cataract	3.05	0.92	(0.87, 0.98)	3.03	1.00	(0.94, 1.06)	3.04	0.96	(0.92, 1.00)
Cerebral palsy	1.36	0.85	(0.77, 0.94)	1.42	0.83	(0.75, 0.91)	1.39	0.84	(0.78, 0.90)
Chronic kidney disease	16.19	1.13	(1.10, 1.17)	16.18	1.13	(1.10, 1.17)	16.18	1.13	(1.11, 1.16)
Colorectal cancer	0.82	0.94	(0.83, 1.06)	0.87	0.98	(0.87, 1.10)	0.84	0.96	(0.88, 1.04)
COPD and bronchiectasis	22.09	1.18	(1.15, 1.21)	22.17	1.18	(1.15, 1.22)	22.13	1.18	(1.16, 1.20)

	Development sample n = 71,096			Validation sample n = 71,097			Full sample N = 142,193		
Risk factor	% of bene-ficiaries	IRR	IRR (95% CI)	% of bene-ficiaries	IRR	IRR (95% CI)	% of bene-ficiaries	IRR	IRR (95% CI)
Cystic fibrosis, other metabolic disorders	0.90	1.02	(0.92, 1.14)	0.94	0.94	(0.85, 1.04)	0.92	0.98	(0.91, 1.06)
Depression	43.47	1.13	(1.08, 1.18)	43.60	1.06	(1.02, 1.10)	43.54	1.09	(1.06, 1.12)
Depressive disorders	34.89	1.02	(0.98, 1.06)	34.71	1.08	(1.04, 1.13)	34.80	1.05	(1.02, 1.08)
Diabetes	25.76	1.20	(1.17, 1.24)	25.32	1.17	(1.14, 1.20)	25.54	1.19	(1.16, 1.21)
Drug use disorder	25.49	1.25	(1.22, 1.28)	25.92	1.23	(1.20, 1.26)	25.71	1.24	(1.22, 1.26)
Endometrial cancer	0.32	1.23	(1.02, 1.47)	0.29	1.26	(1.04, 1.53)	0.30	1.24	(1.09, 1.42)
Epilepsy	9.03	1.29	(1.25, 1.34)	9.09	1.24	(1.20, 1.28)	9.06	1.27	(1.23, 1.30)
Female/male breast cancer	1.55	0.84	(0.77, 0.92)	1.54	0.99	(0.91, 1.08)	1.55	0.91	(0.86, 0.97)
Fibromyalgia, chronic pain, and fatigue	19.06	1.64	(1.60, 1.68)	18.92	1.62	(1.58, 1.66)	18.99	1.63	(1.60, 1.66)
Glaucoma	2.39	0.94	(0.87, 1.00)	2.45	1.01	(0.95, 1.08)	2.42	0.97	(0.93, 1.02)
Heart failure	11.67	1.08	(1.04, 1.12)	11.74	1.02	(0.98, 1.06)	11.71	1.05	(1.02, 1.08)
Hip/pelvic fracture	0.60	0.96	(0.84, 1.10)	0.63	0.80	(0.70, 0.91)	0.62	0.88	(0.80, 0.96)
HIV/AIDS	2.72	1.04	(0.97, 1.11)	2.74	1.07	(1.00, 1.14)	2.73	1.05	(1.00, 1.10)
Hyperlipidemia	24.52	0.95	(0.92, 0.97)	24.58	0.93	(0.91, 0.95)	24.55	0.94	(0.92, 0.96)
Hypertension	47.00	1.16	(1.13, 1.19)	46.77	1.15	(1.12, 1.18)	46.88	1.16	(1.14, 1.18)
Intellectual disabilities and related conditions	3.95	0.97	(0.91, 1.03)	4.04	0.96	(0.90, 1.02)	4.00	0.97	(0.92, 1.01)
Ischemic heart disease	16.13	1.17	(1.14, 1.21)	16.49	1.20	(1.17, 1.24)	16.31	1.19	(1.16, 1.22)
Learning disabilities	0.26	0.84	(0.68, 1.03)	0.30	0.88	(0.73, 1.07)	0.28	0.86	(0.75, 0.99)
Leukemias and lymphomas	0.87	1.07	(0.96, 1.20)	0.86	0.93	(0.83, 1.04)	0.86	1.00	(0.93, 1.09)
Liver disease, cirrhosis, other liver conditions	7.81	1.21	(1.16, 1.26)	7.93	1.24	(1.19, 1.29)	7.87	1.22	(1.19, 1.26)
Lung cancer	0.93	1.08	(0.96, 1.21)	1.01	1.00	(0.89, 1.11)	0.97	1.03	(0.96, 1.12)
Migraine and chronic headache	7.72	1.62	(1.56, 1.68)	7.39	1.55	(1.49, 1.60)	7.56	1.58	(1.54, 1.63)
Mobility impairments	4.31	0.91	(0.86, 0.97)	4.07	0.96	(0.91, 1.02)	4.19	0.94	(0.90, 0.98)
Multiple sclerosis and transverse myelitis	1.00	0.95	(0.86, 1.06)	1.05	0.97	(0.87, 1.07)	1.03	0.96	(0.89, 1.03)
Muscular dystrophy	0.22	0.93	(0.74, 1.16)	0.21	0.75	(0.59, 0.95)	0.21	0.84	(0.71, 0.99)
Obesity	19.99	1.03	(1.00, 1.06)	19.89	1.05	(1.02, 1.07)	19.94	1.04	(1.02, 1.06)
Osteoporosis	1.45	0.96	(0.88, 1.05)	1.46	1.01	(0.93, 1.10)	1.46	0.99	(0.93, 1.05)
Other developmental delays	0.52	1.15	(1.00, 1.32)	0.59	0.92	(0.81, 1.06)	0.56	1.03	(0.94, 1.14)
Peripheral vascular disease	4.81	0.92	(0.87, 0.97)	5.01	0.94	(0.90, 0.99)	4.91	0.93	(0.90, 0.96)
Personality disorders	5.05	1.21	(1.16, 1.27)	5.07	1.13	(1.08, 1.19)	5.06	1.17	(1.14, 1.21)
Posttraumatic stress disorder	5.31	0.97	(0.92, 1.01)	5.44	0.98	(0.93, 1.03)	5.38	0.97	(0.94, 1.00)

	Development sample n = 71,096			Validation sample n = 71,097			Full sample N = 142,193		
Risk factor	**% of bene-ficiaries**	**IRR**	**IRR (95% CI)**	**% of bene-ficiaries**	**IRR**	**IRR (95% CI)**	**% of bene-ficiaries**	**IRR**	**IRR (95% CI)**
Pressure and chronic ulcers	4.78	1.03	(0.97, 1.08)	4.75	0.99	(0.94, 1.04)	4.77	1.01	(0.97, 1.05)
Prostate cancer	0.36	1.00	(0.83, 1.19)	0.42	1.06	(0.90, 1.24)	0.39	1.02	(0.91, 1.15)
Rheumatoid arthritis/osteoarthritis	15.89	1.11	(1.08, 1.15)	15.88	1.05	(1.02, 1.08)	15.89	1.08	(1.06, 1.11)
Schizophrenia	10.88	0.94	(0.89, 1.00)	10.99	0.90	(0.85, 0.95)	10.93	0.92	(0.89, 0.96)
Schizophrenia, other psychotic disorders	15.96	1.18	(1.13, 1.24)	16.14	1.23	(1.18, 1.29)	16.05	1.21	(1.17, 1.25)
Sensory—blindness and visual impairment	0.47	1.10	(0.95, 1.28)	0.49	1.06	(0.92, 1.22)	0.48	1.08	(0.97, 1.20)
Sensory—deafness and hearing impairment	1.84	0.96	(0.89, 1.03)	1.79	1.03	(0.95, 1.11)	1.82	0.99	(0.94, 1.05)
Spina bifida, other congenital anomalies	0.66	1.02	(0.90, 1.16)	0.67	0.91	(0.80, 1.03)	0.67	0.96	(0.88, 1.05)
Spinal cord injury	0.73	1.12	(0.99, 1.27)	0.73	1.18	(1.04, 1.33)	0.73	1.15	(1.05, 1.25)
Stroke/transient ischemic attack	5.78	1.02	(0.97, 1.07)	5.68	1.02	(0.97, 1.07)	5.73	1.02	(0.99, 1.05)
Tobacco use	37.35	1.18	(1.16, 1.21)	37.56	1.16	(1.13, 1.19)	37.46	1.17	(1.15, 1.19)
Traumatic brain injury	0.80	1.11	(0.99, 1.24)	0.79	1.14	(1.02, 1.27)	0.80	1.12	(1.04, 1.21)
Viral hepatitis	5.39	1.17	(1.11, 1.22)	5.51	1.12	(1.07, 1.17)	5.45	1.14	(1.11, 1.18)
Has only behavior condition, no physical condition	9.42	0.90	(0.86, 0.95)	9.40	0.94	(0.89, 0.99)	9.41	0.92	(0.89, 0.96)
Has both physical and behavior conditions	70.06	1.05	(1.02, 1.09)	70.14	1.06	(1.03, 1.10)	70.10	1.06	(1.03, 1.09)

Source: Mathematica analysis of 2013-2014 MAX PS, LT, OT, and IP files.

ADHD = Attention deficit hyperactivity disorder; CI = Confidence interval; COPD = Chronic obstructive pulmonary disease; IRR = Incidence rate ratio.

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk 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.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

We did not select social risk factors (see 2b3.3a). Our expert workgroup also reached broad agreement that it was appropriate to exclude area-level indicators of SES because of the lack of likely predictive capacity.

2b3.5. 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*)

We used a discrimination statistic (McFadden's adjusted R-squared), risk decile plots, and observed to expected ratios to develop and validate the adequacy of the model (see 2b3.6.).

If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

We used a negative binomial model for risk adjustment. The interpretation of R-squared as the proportion of the variation explained by the model is limited to ordinary linear regression, and the count model analogs to R-squared cannot be appropriately interpreted as such. Therefore, we estimated McFadden's R-squared as an analogue. The McFadden's R-squared is defined as $1 - \frac{logl(model)}{logl(null)}$, where $logl(model)$ is the log likelihood value for the fitted model and $logl(null)$ is the log likelihood for the null model which includes only an intercept as predictor in the risk adjustment model. The McFadden's adjusted R-squared takes on values much closer to zero relative to the traditional R-squared measure.[12] We found that the risk adjustment model had a McFadden's R-squared of 0.0387. As a note, the McFadden's adjusted R-squared takes on values much closer to zero relative to the traditional R-squared measure.[13]

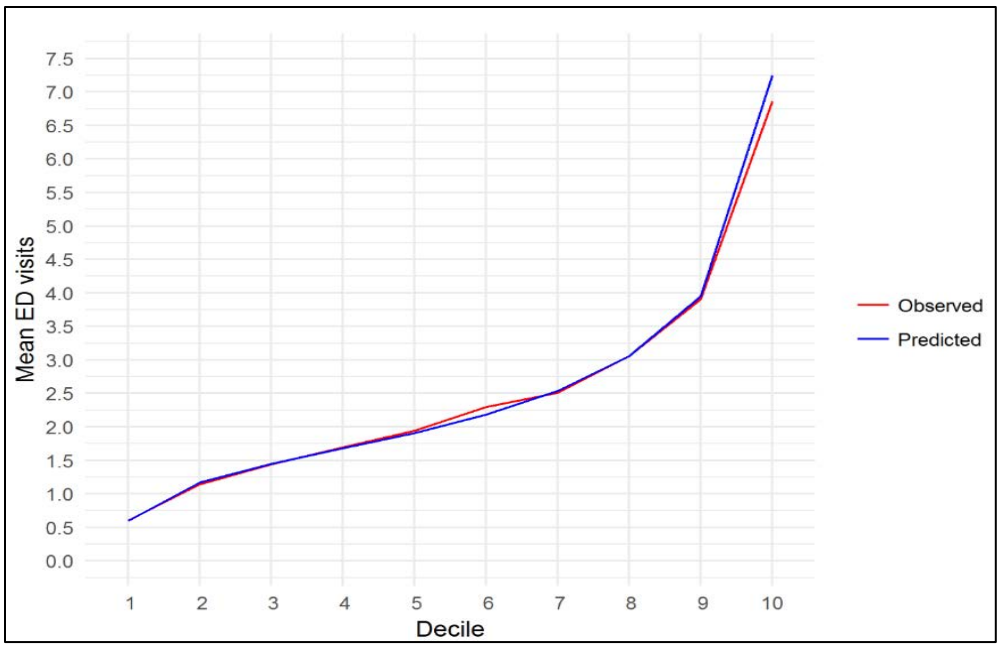
2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

The Hosmer-Lemeshow statistic is not applicable on the BCN-1 risk adjustment model because the outcome is count-valued. Instead, we used risk decile plots to assess model calibration.

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

We used risk decile plots to assess negative binomial model calibration. The decile plot depicted in Figure 1 indicates that the negative binomial generated well-calibrated predictions across the BCN-1 distribution.

Figure 1. Negative binomial decile plot



Source: Mathematica analysis of 2014 MAX PS, LT, OT, and IP files. Model development half sample (n = 71,097).

Note: This figure shows the average predicted and observed rates for each decile of the predicted risk distribution, as estimated using negative binomial regression.

To assess overall calibration of the preferred model specification, we used the validation sample to compute the mean predicted and observed outcome values at each decile in the predicted risk distribution. As seen in Table 6, the predicted and observed outcomes are similar in magnitude at most deciles, scaling

[12] T. Domencich and D.L. McFadden. 1996. "Urban Travel Demand: A Behavioral Analysis." New York, New York: Elsevier. Available at https://eml.berkeley.edu//~mcfadden/travel.html. Accessed April 30, 2017.

[13] T. Domencich and D.L. McFadden. 1996. "Urban Travel Demand: A Behavioral Analysis." New York, New York: Elsevier. Available at https://eml.berkeley.edu//~mcfadden/travel.html. Accessed April 30, 2017.

proportionately as predicted risk thresholds increase. The predicted outcome does not perform as well for the 10th decile, but overall, the similarity across predicted and observed values across the majority of the distribution indicates that the model is well calibrated.

Table 6. Negative binomial decile table

Decile	Number of beneficiaries	Observed mean ED visits	Predicted mean ED visits
1 (lowest)	7,109	0.60	0.60
2	7,110	1.15	1.17
3	7,110	1.44	1.45
4	7,110	1.69	1.68
5	7,109	1.95	1.91
6	7,110	2.29	2.19
7	7,110	2.51	2.54
8	7,110	3.05	3.05
9	7,110	3.90	3.95
10 (highest)	7,109	6.86	7.25

Source: Mathematica analysis of 2014 MAX PS, LT, OT, and IP files. Model validation half sample (n =71,097).

Note: Deciles are classified on the basis of the predicted number of inpatient admissions from the risk adjustment model.

We also calculated a series of observed versus expected (O/E) ratios to assess how well the model performed for important subgroups. A ratio of one indicates that the expected (synonymous with "adjusted" or "predicted") values are approximately equivalent to the observed (synonymous with "unadjusted") values for the subgroup, the desired finding. A ratio greater than one indicates that the observed values are greater than the predicted values, reflecting underprediction of the model. Conversely, a ratio less than one indicates that the observed values are less than the predicted values, reflecting overprediction of the model. Subgroup-specific prediction errors can exert potentially serious unintended consequences. For example, a model that underpredicts events for beneficiaries with multiple comorbidities would inadvertently penalize accountable entities for serving this vulnerable subgroup. Taken as a whole, the results in Table 7 provide reassurance that the model does not suffer from major subgroup-specific prediction errors. The absolute deviation from one for all subgroups fell within 0.04.

Table 7. Predictive performance by key beneficiary characteristics

Characteristic	Observed-to-expected ratio
Sex	
Female	0.99
Male	0.99
Age group	
18–24	1.03
25–44	0.98
45–64	0.98
Eligibility category	
Adult	0.99
Aged/blind/disabled	0.98
Child	0.96
Number of chronic conditions	
2–5	1.04

Characteristic	Observed-to-expected ratio
6–9	0.96
10 or more	0.97
Type of chronic conditions	
Only physical health	1.03
Only behavioral health	1.02
Both physical and behavioral health	0.98

Source: Mathematica analysis of 2014 MAX PS, LT, OT, and IP files. Model validation half sample (n = 71,097).

Note: Expected values are generated from the risk adjustment model. Observed values are the unadjusted, actual measurements.

2b3.9. Results of Risk Stratification Analysis:

Not applicable. We used risk adjustment instead of risk stratification.

2b3.10. 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*)

The preferred specification was well calibrated, with predicted and observed values scaling similarly across predicted risk deciles. Importantly, the model algorithm accurately predicted ED visit rates for important vulnerable subgroups of interest, including beneficiaries covered under the aged/blind/disabled eligibility category and beneficiaries with numerous chronic conditions. This finding provides reassurance that the algorithm appropriately incentivizes accountable entities to serve higher-risk populations.

In summary, the BCN-1 risk adjustment algorithm employs well-established, publicly available risk factors to estimate predicted risk scores at the beneficiary level. These predicted risk scores are statistically sound, as assessed by a series of standard statistical tests. We aligned the risk adjustment approach for BCN-2 to the approach for BCN-1, and the predicted risk scores performed well for both measures. The beneficiary-level scores can be aggregated up to a different level of reporting—for example, the state-level, as detailed in this report—and subsequently translated into performance scores that account for differences across entities in their respective patient sociodemographic and health profiles.

2b3.11. 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*)

Recognizing that small samples may be uniquely vulnerable to the influence of statistical noise, we conducted a power calculation designed to provide guidance to implementers regarding the minimum sample size required for statistically appropriate reporting. Specifically, using a two-sided test, we computed the minimum sample size necessary to detect a 5 percentage point (0.05) difference and a 2.5 percentage point (0.025) difference with 90 percent certainty, noting that the outcome distribution was scaled as a proportion (for example, 0.20 instead of 200 per 1,000) for the purposes of computing statistical power.[14] We chose these thresholds to reflect the standard deviation of the state-level outcome distribution, which was approximately 0.06 (again, scaled as a proportion). As an additional benchmark, we note that the spread of values across the interquartile range of state-level scores was approximately 0.07 (or 70 visits per 1,000 member months).

Using the standard equation for a power calculation for a difference in rates, that is, $n = \frac{2\sigma^2(Z_\beta + Z_{1-\alpha/2})^2}{(p_1-p_2)^2,}$

where $p_1 - p_2$ is the desired minimum detectable difference (either 0.05 or 0.025), and $Z_{1-\alpha/2}$ is the 100(1-$\alpha/2$)th percentile of the standard normal distribution. We set σ^2 at 0.25, the maximum possible variance for

[14] The proportion approximation was appropriate because the outcome distribution had a lower bound of zero and its empirical maximum was considerably smaller than one.

proportions, which provides the most conservative estimate. We chose $\alpha = 0.10$ as the desired significance level and $\beta = 0.8$ as the desired power.

The results indicated that at least 1,232 beneficiary months are required to reliably detect a 0.05 difference and at least 4,920 beneficiary months are required to reliably detect a 0.025 difference. Assuming an average enrollment length of 11 months among beneficiaries—reflecting the analytic sample average—these minimum sample sizes translate into 112 beneficiaries and 447 beneficiaries, respectively.

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. 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 analyzed the distribution of the BCN-1 measure rate at the state level and among demographic and clinical subgroups of interest. We compared performance across state-level inpatient admission rates to understand any variation in performance. We calculated the 95% confidence interval of the ED visit rates for each state using a z-distribution for proportion. We then compared each state's confidence interval to the overall measure rate that uses all beneficiaries across states. State measure rates that are significantly higher than the overall rate indicate an evidence of room for improvement. We conducted similar statistical tests (using a z-test and adjusting all p-values for multiple comparisons using the Benjamini-Hochberg procedure) to understand variation in performance by subgroups such as age, sex, and race.

2b4.2. 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 that the risk-adjusted BCN-1 measure rates across the 10 states cover a wide range with meaningful variation. We found that the risk-adjusted BCN-1 measure rates across the 10 states cover a wide range with meaningful variation. Specifically, the measure rate ranges from 109.4 to 321.8, with a mean of 234 and standard deviation of 60.5. When looking into state specific BCN-1 measure rates, four of the ten, or 40 percent of states, exhibit significantly higher measure rates than the average performer, with their 95% confidence intervals above the overall performance rate (Figure 2). This suggests room for improvement in care coordination for these states. Three states show significantly lower measure rates than the overall performance rate, and there are three states with performance that is statistically indistinguishable from the overall performance rate.

Although evidence is mixed regarding the proportion of hospital-based care that is potentially avoidable,[15] some Medicaid and BCN-specific initiatives have been successful. Specialized care management interventions to reduce inpatient admissions and ED visits among the highest-risk beneficiaries have shown decreases as high as 30 or 40 percent.[16,17] A more plausible estimate for potential reductions in ED visits among BCNs is the

---

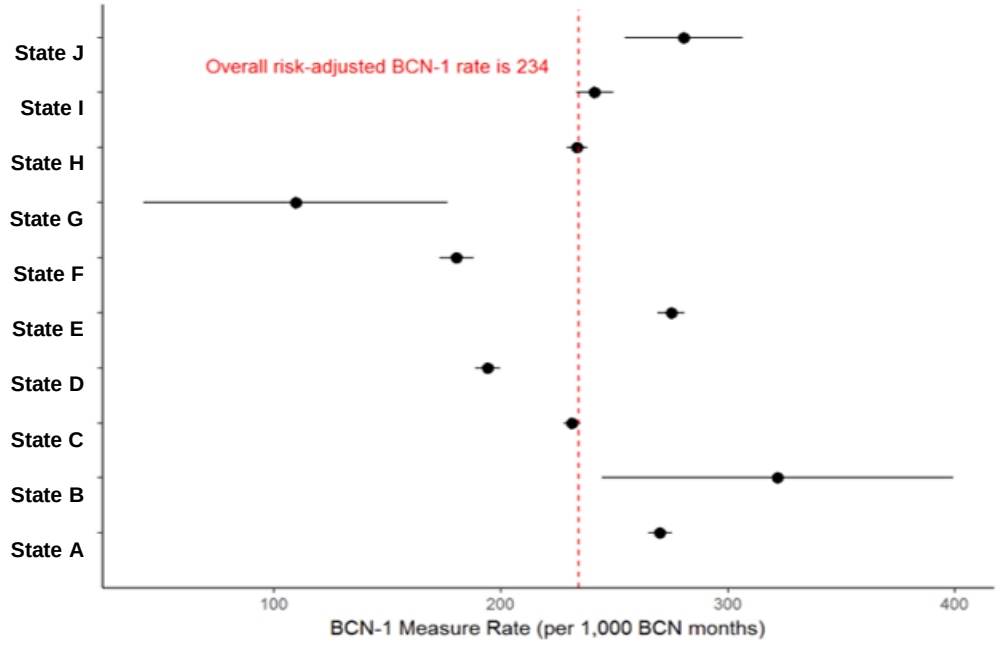
[15] Soril, L., L. Leggett, D. Lorenzetti, T. Noseworthy, and F. Clement. "Reducing Frequent Visits to the Emergency Department: A Systematic Review of Interventions." PLoS One, vol. 10, no. 4, 2015.

[16] Raven, M.C., K.M. Doran, S. Kostrowski, C. C. Gillespie, and B.D. Elbel. "An Intervention to Improve Care and Reduce Costs for High-Risk Patients with Frequent Hospital Admissions: A Pilot Study." BMC Health Services Research, vol. 11, no. 270, 2011.

[17] Green, S.R., V. Singh, and W. O'Byrne. "Hope for New Jersey's City Hospitals: The Camden Initiative." Perspectives in Health Information Management, vol. 7, spring 2010.

15 percent reduction in utilization and/or cost that emerged from a meta-evaluation of patient-centered medical home programs for chronically ill patients.[18]

Figure 2. Risk-adjusted measure rate by state



Source: Mathematica analysis of 2014 MAX PS, LT, OT, and IP files. Full sample (n = 142,193**).**

In general, there is a significant difference in the BCN-1 measure rate between most of the subgroups of age, sex, race, Medicaid eligibility, and comorbidity status. Beneficiaries in the older age group (45 to 64) have a lower BCN-1 rate than the younger age groups (i.e., 18 to 24 and 24 to 44 (both comparisons with p-value < 0.0001). (Table 8). The BCN-1 measure rate is significantly higher (p-value = 0.028) for females than for males. The BCN-1 measure rate is significantly higher for black beneficiaries than other race groups (p-value < 0.001). In terms of Medicaid eligibility, aged/blind/disabled beneficiaries have a significantly higher BCN-1 measure rate as compared with adults and children (p < 0.001). The largest difference in the BCN-1 mean rate across subgroups occurs with comorbidity status: beneficiaries with 1 or more behavioral health conditions have a significantly higher BCN-1 measure rate (255 ED visits per 1,000 member months) than beneficiaries who do not have a behavioral health condition (149.5 ED visits per 1,000 member months; p < 0.001). This finding is consistent with the literature that shows beneficiaries with behavioral health conditions account for a disproportionate level of hospital-based utilization.[19,20,21]

---

[18] Peikes, D., S. Dale, and E. Lundquist. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need?" AHRQ Publication no. 11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

[19] Billings, J., and M.C. Raven. "Dispelling an Urban Legend: Frequent Emergency Department Users Have Substantial Burden of Disease." Health Affairs, vol. 32, no. 12, 2013, pp. 2099–2108.

[20] Prince, J., A. Akincigil, D. Hoover, J. Walkup, S. Bilder, and S. Crystal. "Substance Abuse and Hospitalization for Mood Disorder Among Medicaid Beneficiaries." American Journal of Public Health, vol. 99, no. 1, 2009, pp. 160–167. doi: 10.2105/AJPH.2007.133249.

[21] Boyd, C., B. Leff, C. Weiss, J. Wolff, R. Clark, and T. Richards. "Clarifying multimorbidity to improve targeting and delivery of clinical services for Medicaid populations." Princeton, New Jersey: Center for Health Care Strategies, December 2010. Available at http://www.chcs.org/media/Clarifying_Multimorbidity_for_Medicaid_report-FINAL.pdf. Accessed June 7, 2018.

Table 8. Risk-adjusted measure rate distribution by subgroup

Subgroup	# of Bene-ficiaries	Weighted Mean	Minimum	25th percentile	50th percentile	75th percentile	Maximum
Aged 18 to 24	17,781	246.9	117.6	187.2	248.2	262.6	284.2
Aged 25 to 44	58,949	251.6	114.4	219.9	255.6	285.1	437.6
Aged 45 to 64	65,463	214.9	101.7	187.2	218.7	258.6	303
Male	49,798	231.1	113.3	204.6	229.9	274.2	535.9
Female	92,395	235.6	104.6	179.5	234.3	262.3	277.3
White	82,536	218.2	93.6	189.7	216.9	244.9	265.1
Black	41,627	273.7	0.0	245.2	281.1	303.9	408.8
Hispanic	10,647	247.4	0.0	171.0	204.4	230.3	284.3
Other/unknown	7,383	192.0	113.4	133.3	164.5	213.6	373.8
Aged/Blind/ Disabled	80,569	251.7	113.8	222.4	260.0	312.5	345.7
Adult	58,450	208.9	0.0	176.9	206.7	229.6	272.6
Child	3,174	225.1	120.9	191.3	219.1	236.5	309.3
1 or more behavioral health conditions	113,051	255.7	117.2	220.8	256.1	298.3	317.4
No behavioral health conditions	29,142	149.5	85.19	124.5	149.6	162.6	289.8

Source: Mathematica analysis of 2013 and 2014 MAX PS, LT, OT, and IP files.

2b4.3. 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, the measure indicates both statistically significant and practically meaningful differences in performance across states and among key subgroups, which suggests opportunities for improvement.

In addition, evidence reviewed for BCN-1 suggests that monitoring and improving ED visit rates would improve health outcomes for BCNs. BCNs have disproportionately high levels of ED visits due to their complex medical problems, which are often exacerbated by multiple chronic conditions and social needs. However, evidence suggests at least a portion (around 15%) of hospital-based utilization (including ED visits) may be avoidable. Therefore, we believe all states have an opportunity to reduce the inpatient admission rate among Medicaid BCNs by improving the quality of care for this vulnerable population.

BCN-1 shows meaningful and statistically significant variation across nearly all key subgroups tested, suggesting that the quality of care provided to these subgroups varies. The ability of BCN-1 to detect differences across these subgroups underline its importance in equalizing the care provided to these subpopulations.

2b5. 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 social risk 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 social risk 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.***

2b5.1. 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*)

Not applicable; we have only one set of specifications.

2b5.2. 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*)

Not applicable; we have only one set of specifications.

2b5.3. 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*)

Not applicable; we have only one set of specifications.

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. 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*)

We assessed the extent of missing data using the MAX validation and anomaly tables (citations can be found in the table source notes).

2b6.2. 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*)

The vast majority of the Medicaid eligibility and claims data elements required to both identify the Medicaid BCN population and calculate the BCN-1 measure exhibit negligible missingness in 2013-2014 MAX data for the ten states included in the analytic sample.

We used date of birth in order to calculate age in each year. Almost all states and all years had available dates of birth for enrollees; State F (2014) had the highest missing rate, but still had only 1.3% of enrollees with missing data in that field. Data were not available to assess the missingness for fields used to determine dual status.

We used monthly Medicaid eligibility data from the PS file to identify beneficiaries with at least 10 months of eligibility in the lookback year. Across all states and years, more than 95 percent of beneficiaries in the PS file with claims have Medicaid eligibility information.

We used the type of service data element to identify inpatient stays, which is critical to the utilization component of the BCN definition (at least 1 inpatient stay in the lookback period). Across all states and years,

more than 95 percent of inpatient claims had a type of service of "01" (for "inpatient"), which meets expectations based on historical MAX data for the field.[22]

We used the diagnosis code fields to identify chronic conditions (via the CCW algorithms), which is critical to the chronic condition component of the BCN definition (at least 2 chronic conditions in the lookback period). Across all states and years, more than 95% of IP and LT claims and more than 60% of OT claims had a primary diagnosis code (Table 9).

Table 9. Percent of total claims with a primary diagnosis code

State	Year	% of IP claims with primary diagnosis code[a]	% of LT claims with primary diagnosis code[b]	% of OT claims with primary diagnosis code[c]
State A[d]	2013	100.00	100.00	88.79
	2014	NA	NA	NA
State B[d]	2013	100.00	100.00	95.74
	2014	100.00	100.00	96.14
State C[e]	2013	100.00	100.00	97.19
	2014	100.00	100.00	96.81
State D[d]	2013	100.00	100.00	97.50
	2014	100.00	100.00	97.71
State E[e]	2013	100.00	100.00	82.63
	2014	100.00	100.00	79.46
State F[d]	2013	100.00	100.00	97.37
	2014	NA	NA	NA
State G[d]	2013	100.00	95.83	84.21
	2014	100.00	95.56	84.41
State H[e]	2013	100.00	100.00	100.00
	2014	98.10	99.99	99.93
State I[d]	2013	100.00	100.00	90.67
	2014	100.00	100.00	96.93
State J[d]	2013	100.00	100.00	84.77
	2014	100.00	100.00	83.50

Source: MAX validation tables. Available at < https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports.html >.

[a] Based on historical MAX data, the expected range for this column is 98-100%.

[b] Based on historical MAX data, the expected range for this column is 95-100%.

[c] Based on historical MAX data, the expected range for this column is >60%.

[d] We assessed fee-for-service non-crossover claims for State A, State B, State D, State F, State G, State I, and State J.

[e] We assessed managed care (encounter) claims for State C, State E, and State H.

NA = Not available

[22] Source: MAX validation tables. Available at < https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports.html >.

Table 10 contains missingness information related to the remaining data elements necessary to identifying those ED visits retained in the BCN-1 numerator, including revenue codes, procedure codes, and place of service codes.

All states and years fall within the expected range (35-70%) for the percentage of IP claims with a procedure code and one state (State F in 2013) exceeded that range. Most states and years also exceed the expected threshold (>95%) for the percentage of OT claims with a procedure code, and all states and years with available data showed that OT claims without a procedure code at least had a revenue code. Only one-third of states and years with available data have a proportion of OT files with place of service codes that exceeds the expected threshold (>95%). However, no state has below 87% of OT files with place of service codes.

Dates of service were also used in the calculation of the BCN-1 measure, to distinguish or condense ED visits, but they are uniformly available because they are required fields for MAX IP and OT claims.

Table 10. Percentage of total claims with data elements necessary to identify Emergency Department visits)

State	Year	% of IP claims with a procedure code[a]	% of OT claims with a procedure code[b]	% of OT claims with a procedure code or UB-92 revenue code	% of OT claims with place of service code[b,c]
State A[d]	2013	58.37	91.33	100.00	92.29
	2014	NA	NA	NA	NA
State B[d]	2013	60.51	96.33	100.00	88.27
	2014	60.74	96.58	100.00	88.31
State C[e]	2013	62.09	97.52	100.00	NA
	2014	62.75	97.71	100.00	NA
State D[d]	2013	42.93	100.00	100.00	93.15
	2014	44.13	100.00	100.00	92.59
State E[e]	2013	59.38	94.43	100.00	NA
	2014	61.66	96.28	100.00	NA
State F[d]	2013	74.83	99.24	100.00	87.64
	2014	NA	NA	NA	NA
State G[d]	2013	56.15	84.71	100.00	95.27
	2014	55.56	85.19	100.00	95.56
State H[e]	2013	62.58	99.88	100.00	NA
	2014	60.61	99.88	100.00	NA
State I[d]	2013	59.69	98.91	100.00	96.37
	2014	60.74	98.72	100.00	98.06
State J[d]	2013	61.02	99.84	100.00	94.13
	2014	61.64	99.87	100.00	93.08

Source: MAX validation tables. Available at < https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports.html >.

[a] Based on historical MAX data, the expected range for this column is 35-70%.

[b] Based on historical MAX data, the expected range for these columns is >95%.

[C] Place of service was not assessed among managed care claims in the MAX validation tables, as reflected in "NA" entries for State C, State E, and State H.

[d] We assessed fee-for-service non-crossover claims for State A, State B, State D, State F, State G, State I, and State J.

[e] We assessed managed care (encounter) claims for State C, State E, and State H.

NA = Not available

2b6.3. 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*)

Given the relatively small amount of missing information used to identify the BCN population and calculate the BCN-1 measure, we do not believe there is any systematic bias in our testing. States implementing the measure with their own data will likely have fewer missing fields than reported above because they are better equipped to account for state-specific codes when identifying the BCN population and constructing the measure.

APPENDIX 1

The states used for testing were based on the most current MAX data available at the time that testing began and data quality checks in which we determined which states had sufficient quality of their FFS or encounter records. Specifically, we reviewed the FFS and managed care population in each of the 15 states with available 2014 MAX data (as of the beginning of testing) for inclusion in the analytic sample. We developed criteria to review each state's Medicaid fee-for-service (FFS) and managed care populations:

Selection of states with FFS data. States had to meet three criteria to be included in the analytic sample:

(1) At least 25 percent of the state's Medicaid population had to be enrolled in FFS in 2014

This requirement ensured that each state would have a sufficient FFS population throughout the testing period. Three states did not meet this requirement.

(2) There had to be no data anomalies that impacted our testing ability

Data anomalies limit our ability to interpret testing results. One state was excluded because the Medicaid population identified in its 2014 MAX enrollment data was significantly smaller than the 2013 and 2012 population. The drop coincided with the state's 1115 wavier demonstration, which provided newly eligible Medicaid beneficiaries premium assistance for private insurance plans on the states' marketplace.

(3) State-level measures of inpatient admissions and ED visits had to align with national benchmarks.

We required state-level outcome measures to align with national benchmarks to confirm that the FFS population in each state was broadly representative and that state-specific coding conventions accurately captured hospital-based utilization in claims. Historically, roughly 9.5 percent of adult Medicaid beneficiaries have at least one inpatient admission per year and 31.5 percent have at least one ED visit per year.[23] [24] Using these figures as benchmarks, we eliminated four states that diverged by more than 60 percent from either hospital-based care benchmark.

Selection of states with managed care data. States had to meet three criteria to be included in the analytic sample:

(1) Enrollment in comprehensive managed care had to be greater or equal to 50 percent of state's Medicaid population in 2014

This requirement ensured that each state would have a sufficient managed care population throughout the testing period. Eight states did not meet this requirement.

[23] Wherry, L. R., M. E. Burns, L. J. Leininger. "Using self-reported health measures to predict high-need cases among Medicaid-eligible adults." Health Services Research, vol. 49, no. 6.1, December 2014, pp. 2147-72.

[24] Garcia, T.C., A.B. Bernstein, and M.A. Bush. "Emergency Department Visitors and Visits: Who Used the Emergency Room in 2007?" National Center for Health Statistics Data Brief, no. 38. Atlanta, GA: Centers for Disease Control and Prevention, 2010.

(2) The state had to have quality assurance checks and policies to promote utilization reporting from all health plans

Although states fairly consistently report enrollment in managed care plans and capitation payments, encounter claims (unlike FFS claims) have been known to under report service utilization.[25] Insufficient utilization information in encounter claims would impede measure accuracy by undercounting the number of ED visits in the measure numerator. Four states were excluded because they did not require all health plans to submit data, impose financial penalties for failing to submit data, and/or report that they are using their data for analyses.

(3) State-level measures of inpatient admissions and ED visits had to align with national benchmarks

We required state-level outcome measures to align with national benchmarks to confirm that the FFS population in each state was broadly representative and that state-specific coding conventions accurately captured hospital-based utilization in claims. Historically, roughly 9.5 percent of adult Medicaid beneficiaries have at least one inpatient admission per year and 31.5 percent have at least one ED visit per year.[26,27] We did not eliminate any state for this criterion.

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*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic claims

---

[25] Byrd, Vivian L. H., and Allison Hedley Dodd. "Assessing the Usability of Encounter Data for Enrollees in Comprehensive Managed Care 2010-2011." Medicaid Policy Brief No. 22. Prepared for the Centers for Medicare & Medicaid Services. Mathematica Policy Research, August 2015. Available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/Downloads/MAX_Encounter_Brief_2010_2011.pdf. Accessed December 8, 2017.

[26] Wherry, L. R., M. E. Burns, L. J. Leininger. "Using self-reported health measures to predict high-need cases among Medicaid-eligible adults." Health Services Research, vol. 49, no. 6.1, December 2014, pp. 2147-72.

[27] Garcia, T.C., A.B. Bernstein, and M.A. Bush. "Emergency Department Visitors and Visits: Who Used the Emergency Room in 2007?" National Center for Health Statistics Data Brief, no. 38. Atlanta, GA: Centers for Disease Control and Prevention, 2010.

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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

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. Please also complete and attach the NQF Feasibility Score Card.

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Not applicable

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.

Specific Plan for Use	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

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

- Level of measurement and setting

Not applicable

4a1.2. 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?*)

CMS is considering implementation plans for this measure. There are no identified barriers to implementation in a public reporting application.

4a1.3. 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.*)

CMS is developing measures to improve the quality of care of Medicaid populations served by CMS's Innovation Accelerator Program, which includes Medicaid BCNs. This measure is intended for use by states to monitor and improve the quality of care provided for the Medicaid BCN population. States may choose to begin implementing the measures based on their programmatic needs.

This measure is intended to be paired with BCN-2, an all-cause inpatient admission measure that was also developed specifically for the Medicaid BCN population.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

This measure has not been implemented yet. Unlike Medicare measures, there is no formal process by which draft results for Medicaid measures are shared with measured entities. However, we invited feedback from a 19-member technical expert panel (TEP), a 7-member risk-adjustment work group, and the public (via a public comment process). The TEP included at least three current or former state Medicaid officials. Refer to Ad.1. for details on the TEP and risk-adjustment work group.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Not applicable.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Not applicable.

4a2.2.2. Summarize the feedback obtained from those being measured.

Not applicable.

4a2.2.3. Summarize the feedback obtained from other users

Not applicable.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

Not applicable.

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

This measure is being considered for initial endorsement and was not in use for performance improvement at the time of initial NQF submission. We have no historical rates for this measure.

Across all states in the analytic sample, the risk-adjusted BCN-1 measure rate is 234 ED visits per 1,000 beneficiary months, with a range of 109.5 in State G to 322.0 in State B. The ED utilization measure may be useful for monitoring the rate of ED visits among BCNs and encourage states to develop interventions to decrease the rates. This is important because evidence suggests that at least a portion of these ED visits may be avoidable, given adequate access to outpatient care, care coordination, and disease self-management skills (Billings and Raven 2013; Capp et al. 2013; Doupe et al. 2012; Pukurdpol et al. 2014).

References:

Billings, J., and M.C. Raven. "Dispelling an Urban Legend: Frequent Emergency Department Users Have Substantial Burden of Disease." Health Affairs, vol. 32, no. 12, 2013, pp. 2099–2108.

Capp, R., M.S. Rosenthal, M.M. Desai, L. Kelley, C. Borgstrom, D.L. Cobbs-Lomax, P. Simonette, and E.S. Spatz "Characteristics of Medicaid Enrollees with Frequent ED Use." American Journal of Emergency Medicine, vol. 31, no. 9, 2013, pp. 1333–1337.

Doupe, M.B., W. Palatnick, S. Day, D. Chateau, R.A. Soodeen, C. Burchil, and S. Derksen. "Frequent Users of Emergency Departments: Developing Standard Definitions and Defining Prominent Risk Factors." Annals of Emergency Medicine, vol. 60, no. 1, 2012, pp. 24–32.

Pukurdpol, P., J.L. Wiler, R.Y. Hsia, and A.A. Ginde. "Association of Medicare and Medicaid Insurance with Increasing Primary Care-Treatable Emergency Department Visits in the United States." Academic Emergency Medicine, vol. 21, no.10, 2014, pp. 1135–1142.

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

BCN-1 has not been implemented yet. There were no unexpected findings identified during testing of this measure.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

BCN-1 has not been implemented yet. There were no unexpected benefits identified during testing of 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)

0173 : Emergency Department Use without Hospitalization During the First 60 Days of Home Health

2505 : Emergency Department Use without Hospital Readmission During the First 30 Days of Home Health

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

Non-NQF endorsed: HEDIS Emergency Department Utilization (EDU), a version of which is also included in the Medicaid Child Core Set: Ambulatory Care–Emergency Department Visits (AMB-CH)

5a. Harmonization of Related Measures

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 harmonized to the extent possible?

No

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

Parts of the specifications for BCN-1 harmonize with the three related measures listed in question 5. Differences between BCN-1 and the other measures, described below, do not impose additional data collection burden to states, because the data elements are available in administrative data and are consistent with some measures states are already likely collecting.

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.)

Not applicable; there are no competing NQF-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.

No appendix Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services, Centers for Medicaid & CHIP Services

Co.2 Point of Contact: Roxanne, Dupert-Frank, Roxanne.Dupert-Frank@cms.hhs.gov, 410-786-9667-

Co.3 Measure Developer if different from Measure Steward: Mathematica Policy Research

Co.4 Point of Contact: Melissa, Azur, mazur@mathematica-mpr.com, 202-250-3518-

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 Technical Expert Panel (TEP) members advised on the development of the initial measure concept and preliminary specifications; reviewed testing results and feedback obtained during public comment; and advised on the refinements of the technical specifications.

TEP members:

Consumers or consumer representatives

Carol McDaid, Capitol Decisions, Inc.

Janice Tufte, Patient-Centered Outcomes Research Institute (PCORI) ambassador

Kayte Thomas, PCORI ambassador

State officials

David Mancuso, Washington State Department of Social and Health Services

Roxanne Kennedy, New Jersey Division of Mental Health and Addiction Services

Health plans

Alonzo White, Aetna Medicaid

Deb Kilstein, Association for Community Affiliated Plans

Jim Thatcher, Mass. Behavioral Health Partnership, Beacon Health Options

Provider organizations

Daniel Bruns, Health Psychology Associates

Aaron Garman, Coal Country (ND) Community Health Center (and American Academy of Family Practice Comm. on Quality & Practice)

Annette DuBard, Aledade, Inc.

Joe Parks, National Council for Behavioral Health (formerly of Missouri HealthNet Division-- Medicaid)

Subject matter experts and researchers

Andrew Bindman, University of California San Francisco School of Medicine

Alex Sox-Harris, Department of Veterans Affairs

Benjamin Miller, Farley Health Policy Center, University of Colorado School of Medicine

Kimberly Hepner, RAND Corporation

Mady Chalk, Treatment Research Institute

Federal agency officials

DEB Potter, Office of the Assistant Secretary for Planning and Evaluation

Laura Jacobus-Kantor, Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality

The Risk Adjustment Work Group members advised on the development of the initial risk-adjustment models, reviewed risk-adjusted testing results, and advised on the refinements of the risk-adjustment model.

Risk Adjustment Work Group members:

Marguerite Burns, PhD (Assistant Professor, University of Wisconsin School of Medicine and Public Health)

Ezra Golberstein, PhD (Associate Professor, University of Minnesota School of Public Health)

Lisa Iezzoni, MD, MSc (Professor, Harvard Medical School)

Joanna Jiang, PhD (Senior Social Scientist, Agency for Healthcare Research and Quality)

Zhenqiu Lin, PhD (Director of Data Management and Analytics, Center for Outcomes Research and Evaluation, Yale University)

Patrick Romano, MD (Professor, University of California, Davis, School of Medicine)

Jonathan Shaw, MD, MS (Clinical Assistant Professor, Stanford University 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: The International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) is published by the World Health Organization (WHO). ICD-10-CM is an official Health Insurance Portability and Accountability Act standard.

The International Classification of Diseases, 10th Revision, Procedure Coding System (ICD-10-PCS) is published by the World Health Organization (WHO). ICD-10-PCS is an official Health Insurance Portability and Accountability Act standard.

Current Procedural Terminology (CPT)® codes copyright 2018 American Medical Association (AMA). All rights reserved. CPT is a trademark of the American Medical Association. No fee schedules, basic units, relative values or related listings are included in CPT. The AMA assumes no liability for the data contained herein. Applicable FARS/DFARS restrictions apply to government use.

Healthcare Common Procedure Coding System (HCPCS) Level II codes and descriptors are approved and maintained jointly by the alpha-numeric editorial panel (consisting of the Centers for Medicare & Medicaid Services, America´s Health Insurance Plans, and Blue Cross and Blue Shield Association).

The American Hospital Association (AHA) holds a copyright to the National Uniform Billing Committee (NUBC) codes contained in the measure specifications. The NUBC codes in the specifications are included with the permission of the AHA. The NUBC codes contained in the specifications may be used by states for the purpose of calculating and reporting Measure results or using Measure results for their internal quality improvement purposes. All other uses of the NUBC codes require a license from the AHA. Anyone desiring to use the NUBC codes in a commercial product to generate measure results, or for any other commercial use, must obtain a commercial use license directly from the AHA. To inquire about licensing, contact ub04@healthforum.com.

Healthcare Effectiveness Data and Information Set (HEDIS) Value Sets

This measure contains HEDIS® Value Sets that were developed, are owned by and are included with the permission of the National Committee for Quality Assurance ("NCQA"). Proprietary coding is contained in the HEDIS Value Sets. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. NCQA disclaims all liability for use or accuracy of any coding contained in the HEDIS Value

Sets. The HEDIS Value Sets are provided "as is" without warranty of any kind. Users shall not have the right to alter, enhance or otherwise modify the HEDIS Value Sets, and shall not disassemble, recompile or reverse engineer the HEDIS Value Sets. All uses of the HEDIS Value Sets outside the measure must be approved by NCQA and are subject to a license at the discretion of NCQA.

Ad.7 Disclaimers: This performance measure is not a clinical guideline and does not establish a standard of medical care, and has not been tested for all potential applications. The measure and specifications are provided without warranty.

Ad.8 Additional Information/Comments: