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 #: 3451

Measure Title: Non-Acute Mental Health Services Utilization for Dual Eligible Beneficiaries

Measure Steward: Centers for Medicare & Medicaid Services

Brief Description of Measure: The percentage of dual eligible beneficiaries with a mental health service need who received a non-acute mental health service in the measurement year.

Developer Rationale: Appropriate access to and use of evidence-based mental health services can reduce the probability that individuals diagnosed with mental health conditions suffer prolonged distress and may help prevent unintended consequences caused by untreated mental illness. Depression, anxiety disorders, schizophrenia/other psychotic disorders, and other bipolar disorders are among the most common behavioral health conditions among dual-eligible beneficiaries (MedPac/MACPAC, 2015).

Ensuring appropriate use of ongoing, non-acute treatment for individuals with mental illnesses could significantly improve population health and quality of life. Measurement of mental health service use for dual eligible beneficiaries with mental health needs provides important information to health plans, consumers and other stakeholders as to how well a system of care helps individuals access the resources necessary to treat their mental illness. The health plan can play a central role in improving access to timely and affordable mental health services through encouraging integration of mental health services into primary care, ensuring an adequate number of mental health professionals in their provider networks, and ensuring accurate information about these professionals is provided to individuals with mental health service needs.

Reference:

MedPAC & MACPAC. (2015). Data Book: Beneficiaries Dually Eligible for Medicare and Medicaid. Retrieved February 10, 2017 from https://www.macpac.gov/wp-content/uploads/2017/01/2015-Dually-Eligible-Beneficiaries-Data-Book.pdf

Numerator Statement: The number of dual eligible beneficiaries receiving at least one non-acute mental health service in the 12-month measurement year. The following services are included as non-acute mental health services:

- Outpatient service with a mental health provider for a mental health diagnosis
- Mental health outpatient encounter
- Mental health condition management in primary care

Denominator Statement: The number of dual eligible beneficiaries age 21 and older with a mental health service need in the 18-month identification window (the 12-month measurement year plus six months prior to the measurement year).

Denominator Exclusions: None

Measure Type: Process

Data Source: Claims

Level of Analysis: Health Plan

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- The developer included several scientific references with brief descriptions to support the use of non-acute mental health services to treat mental health conditions and included categorization of that evidence into 4 distinct modalities of treatment (i.e., psychosocial, pharmacologic, ECT, and behavioral integration into primary care). The references cited included many systematic reviews composed of many randomized trials with brief, alibet sometimes vague mention to effect sizes and their meanings, and with rare mention of study/finding limitations that would result in grading of the evidence. Still the evidence presented is rich and robust, and supportive of their measurement.
- One concern that remains in the developers' presentation is that they do not directly state how they connected the evidence presented to the code lists they generated as operational defintions of the measurement numerator and denonminator (i.e., the treatment and indication).

Exception to evidence

N/A

Questions for the Committee:

- Is the literature review (absent specific grading of the evidence) sufficient to demonstrate that the specific measure is rationally connected to the specified and desirable patient outcomes?
- Is the measure conceived of and cast with the appropriate level of logical sensitivity and specificity to mental illness diagnoses and treatments?
- Are effect sizes and sample sizes sufficiently described and summarized for the Committee to accept this measure as proferred?

Guidance from the Evidence Algorithm

Box 1→ Box 3→ Box 7→ Box 8 → Box 9 (YES) => Moderate evidence, a systematic review that suggests the measure benefits moderately outweight its risks.

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

RATIONALE: The developers present a solid review of many meta-analyses including many randomized trials in each. Weakness in this presentation include: the absence of clearly described comparisons between non-acute versus acute interventions, and limited discourse and citations regaring the efficacy of integration and ECT as treatments for mental illnesses. Details about effect sizes are sometimes missing. Details regarding how the summarized evidence-base was used directly to generate the measurement numerator and denominator are also missing.

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 overall performance rate on this measure across 40 Medicare-Medicaid plans including 77,497 duals with data from 2015-6, was 32.4%. Substantial variation around this mean was also evident (standard deviation= 15.6%).
- Data from Washington State Department of Social Services revealed that in 2017 less than half (43%, n=356,222) of Medicaid enrollees in need of mental health services receive care.

Disparities

- Substantial differences were observed between the elderly (≥65) and all other adults (21-64), 26.8% and 55.1%, respectively.
- Aggregate reporting by the developer also revealed marked differences based on each of the following subject characteristics: gender, race/ethnicity, although the latter effects were mainly evident in duals over the age of 65.

Questions for the Committee:

- Are the age strata presented sufficiently fine-grained for this application? Do they suggest that age-adjustment is necessary for use of the measure? (The developers do not require such stratification in their specifications, though they advocate for it).
- Do the disparities presented demonstrate gaps in care that warrants a national performance measure that is so constructed to broadly assess MH service use?
- Is the measure too broadly constructed to identify evidence-based MH services care (as opposed to any care), and to assess if that care is of proper intensity? (Note that one service in 12 months is all that is necessary to be counted in the numerator).

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

RATIONALE:

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

1a. Evidence
Comments:

**Connecting the evidence, which is strong, with indications and treatment definitions and codes remains unclear.

**The evidence for this measure is strong. Very large sampe size. Targeted to a high need population.

**The systematic review, quality, quantity, and consistency of evidence is adequate. In terms of potential exceptions to the evidence, I'm not sure there is good evidence that seeing a clinician for one mental service is at all correlated with improved clinical outcomes or quality of life. It may be intuitive and we clinicians like to believe it, but it is still far from being proven.

**Meets evidence. No New studies I am aware of.

**Evidence is presented, but this measure seems broad and vague to me. there's no information describing severity of symptoms, 'how is the patient doing,' etc... and it's unclear if the patient does receive a service, does it positively impact patient's outcome... very non-specific.

**The literature review, which lacks grading, is very broad and is not sufficient to demonstrate that this measure is rationally connected to the specified and desirable patient outcomes. The effect size is also lacking and the connection of the evidence to teh code set for the numerator and denominator is missing.

**This measure assesses % of dual eligible beneficiaries with a broadly defined mental health need who receive any broadly defined "non-acute mental health service" which is really any contact with mh care in the past 12 months. This does not align closely with the logic model that assumes access to appropriate care and receipt of needed mh services. Appropriate care is not assessed. Medication and therapy are treated equally. Does not access intensity of service use. Does not differentiate acute and non-acute mh services. There is no capacity to adjust for social risk factors, but the logic model includes as an outcome reduced risk of homelessness,violent episodes, incarceration.

**The developer includes several scientific references that support the use of non-acute mental health services to treat mental health conditions and are inclusive of systematic reviews of many randomized trials. There is no direct linkage of how these findings related to their data specifications.

1b. Performance Gap

Comments:

**Large, significant performance gap that needs to be addressed in both >65 and <65 groups.

**Very strong evidence of gender, racial and ethnic disparities. Gaps in case for mental health service delivery in the dual eligible population are significant.

**Very strong evidence of gender, racial and ethnic disparities. Gaps in case for mental health service delivery in the dual eligible population are significant.

**Yes there is a performance gap.

**No concerns; gap demonstrated that patients may not be receiving adequate services.

**The overall performance rate on this measure was 32.4% Substantial variation around this mean was evident. Fewer than 43% of the overall number of dual eligible beneficiaries with a mental healht need acccessed non-acute mental health services. There were also statistically significant differences in beneficiary level performance by age (55.1% for adutsl vs. 25.8% for older) as well as by sex and race/ethnicity. All of these indicators demonstrate that there are definite disparities across a number of metrics which warrant a national performance measure.

**Yes, performance gap and variation by age strata (21-64 vs. 65+), sex and race/ethnicity are provided and findings generally support disparities in care.

**The developer demonstrated that the overall performance was 32.4% with substantial variation around the mean. Marked differences noted in the elderly as well as by gender and race/ethnicity—particularly in those over the age of 65.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

2c. For composite measures: empirical analysis support composite approach

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.

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

Evaluators: Behavioral Health Project Staff

Review A

Evaluation of Reliability and Validity:

SNR calculated at the health plan level were high and based on substantial sample (n= 59K). The SNR ratios were similar after stratification into age groupings above and below 65 years. Face validity of the measure was confirmed as "strong" via a TEP.

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 staff 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:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- Does the Committee think there is a need to discuss and/or vote on validity?

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

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

Evaluation A: Scientific Acceptability

Measure Number: 3451

Measure Title: Non-Acute Mental Health Services Utilization for Dual Eligible Beneficiaries

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.*

2. **Briefly summarize any concerns about the measure specifications.**
   - Completeness and validity of the value sets put forth in section S.5 is presently not clear, though they look reasonable. (procedure and diagnostics number more than 1,200; drug codes number more than 20,000)

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**
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** **N/A**
6. **Assess the method(s) used for reliability testing**
   - SNR calculations, seemed appropriate.

   Submission document: Testing attachment, section 2a2.2
7. **Assess the results of reliability testing**
   - SNR scores are above an average of 0.9 for both age strata. Ranges drop as low as 0.62, but the 25th percentile is above 0.87 across all plans.

   Submission document: Testing attachment, section 2a2.3
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):

    ☐ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

    ☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**
    - The score level reliability was assessed using SNR and demonstrated to be above 0.87 for most plans. However, this is all dependent on reliability of claims values which were not tested. Exclusions seemed reasonable, but were not fully justified (e.g., why were plans with <30 observations excluded, why were California claims considered bad, but other states not so? Why did all of Rhode Island's plans have such low enrollment?)

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

13. Completeness of value sets, and specificity of those sets. HEDIS 2019 and TEP review by 6 experts who "agree" (4) or "strongly agree" (2) that the measure is valid.**Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**
    - Broadness, which is assumed to be lack of specificity, was expressed by at least one commentator on this measure.

    Submission document: Testing attachment, section 2b4.

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.
    - None

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

    Submission document: Testing attachment, section 2b6.
    - None

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? ☐ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

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? ☐ Yes ☐ No

16e**. Assess the risk-adjustment approach**

Notation: The Developers do briefly justify and report very crude (above and below 65 years) age strata for their measures, but they choose not to identify this as social risk adjustment on their submission, and they also submit aggregate scores. Instead, it seems the developers wish only to suggest that such stratification is useful for implementation of this measure, but not essential. Moreover, they ultimately seem to use the age strata to justify the reliability of the measure to identify meaningful group differences.

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)

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

- Face validity assessed exclusively, though the age differences could have been utilized as indirect evidence because increasing age is known to correlate with decreasing mental disorder symptomology in those who survive past 65 years. Technical evaluation panel (TEP) was convened and surveyed to evaluate face validity.

Submission document: Testing attachment, section 2b2.2

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

- General consensus that the measure had adequate face validity.

Submission document: Testing attachment, section 2b2.3

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

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☐ **No**

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

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)

23. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

☐ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐ **Low** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT.)

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

- This is a very broad measure where face validity is based only on expert consensus, and where claims were not tested against some external "gold" diagnostic standard. Still, the measure presents with reasonable face validity. The developers could argue the age strata analyses support validity given what is known about the epidemiology of mental health disorder morbidity. Sensitivity analyses regarding the exclusions (e.g., of California and of <30 groups) might be informative regarding the consequences of those exclusionary choices applied to the reliability of this measure.

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

2a1. Reliability – Specifications
Comments:

**No concerns.

**The submission contains significant detail on data elements and claims coding. Given that this measure would be required as part of a Medicaid managed care contract, consistency in collection should not be a problem.

**Seems adequate.

**No concerns. Would be reliable.

** It has not been made sufficienlty clear why the data from 2 states (California and Rhode Island) were eliminated from inclusion. The testing for reliabilty used a signal to noise analysis to capture the magnitude of differences in underlying performance between MMPs taking into consideration noise (case mix and randome error). SNR calculated at the health plan level were high and teh calculated algorithms seemed accurate. There are no concerns about teh likelihood that this measure can be consistently implemented.

**Specifications are clearly described and could be likely implemented consistently.

**All clearly defined; appears it can be consistently implemented as based on claims data; however, there were states that had too few plans to be included or data determined not to be reliable (e.g CA).

2a2. Reliability – Testing
Comments:

**No concerns.

**No.

**None.

**No.

**No.

** Reliability testing was based on SNR ratios. The methods are clearly described and the results presented support high SNR. Even when stratified they are all above cutpoint of .70.

** No--SNR calculations appear appropriate with a level > 0.87 for most plans; however as noted above there was missing data from several states.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:

**No concerns.

**I'm not convinced 1 visit in a calendar year truly proves the desired outcome. I woiuld rate overall validity as "low" for this reason.

**Not sure the measure is valid if it is only seeking to identify the patient had 1 session in the year. I do not think that would be adequate to show that the patient received adequate treatment.

**Only face validity was assessed based on expert consensus. Claims were not tested against an external substantiated standard. There was no sensitivity analysis regarding the exclusions all of which raises cncerns about the testing.

**Based solely on face validity: data sources: feedback from expert work group and TEP, 4-item survey of TEP (n=6), and comments from two health plans which included 5 comments total during a 3-week public comment period. No TEP member rated disagree or strongly disagree on the ratings related to face validity, and only one public comment out of 5 total was not supportive.

**Adequate face validity.

**Concern about inclusion of dementia in denominator; on the face of it this makes no sense.

**I do not see any threats to validity in this meausre.

**These sub-measures adequately meet the need.

**As above, the threat to validity stems from data excluded without analysis. The measure was stratified by ageto allow for more appropriate comparison of MMP performance.

**2b4: meaningful differences: does not meet this criteria, not implemented and this measure broadly assesses any contact with mh care not really the quality of care provided.

**Overall appears valid. It is unclear what the sample effect would be re: states that were not included.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:

**One could consider the stratification into >65 and <65 risk adjustment since there are large differences; developer chose not to view the presentation of the measure this way though data are presented.

**The non-elderly dual eligible population with a mental illness diagnosis is a dramatically higher risk of poor health outcomes. Additional measures focused on this population -- particularly encounter data -- are desperately needed.

**The rationale is reasonable and intuitive but not sufficiently proven.

**I see that they allow for a telehealth visit in primary care but tthey do not allow for a telehalth visit in a MH setting. why exclude telehealth in a MH setting. Could the developer discuss why this is different for primary care vs. MH?

**Children are not included. in my opinion, this is a more important group to establish early intervention services to improve outcomes/developmental trajectory.

**Data from 2 states was excluded wihtout sufficient explanatin.

**No capacity to adjust for social risk.

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 completely reliant on the accuracy and availability of claims data, otherwise it is quite feasible to implement and maintain.
- All data elements are in defined fields in electronic claims.

Questions for the Committee:

- Are the value sets sensitive and specific to the measures intent, and are they likely to be used in claims submission?
- Is there any concern about the way the value sets are to be deployed? E.g., are any procedure codes used to determine denominator status? Are remission codes used to determine denominator status?

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

RATIONALE: Please see above questions which mark concerns. Otherwise the feasibility with claims is quite reasonable.

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

** No concerns.

**42 CFR Part 2 routinely limits the sharing addiction treatment encounter data for dual eligibles. In addition, behavioral health providers are behind the rest of health care in use of EHRs.

**It has been shown to be feasible.

** No concerns.

** this is a new measure.

** I have no concerns about how hte data collection strategy can be put into operational use relying on the accuracy and availabilty of claims data.

**Feasiblity is acceptable because uses study variables from existing administrative data.

** Relies on accuracy of claims data and appears feasible.

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

Developer notes the following planned uses:

- CMS' Financial Alignment Initiative (FAI) core measure set for Medicare-Medicaid Plans (MMPs)
- Measure was not tested in California because of unspecified problems with "quality" of data.
- Measure was not tested in Rhode Island because sample size was too small.
- Measure was successfully tested in 8 other states (IL, MA, MI, NY, OH, SC, TX, VA)

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

- Feedback, so far, has come only from their convened TEP which included "representatives from national associations of state health officers, two individuals representing two states out of the eight in the MMP demonstration, and two representatives from health plans, selected based on convenience and prior experience as participating key experts." Notable feedback from the TEP, not previously described, was that dementia cases might warrant exclusion. The inclusion of dementia cases in the denominator of the measure might be an error that would bias the measure's magnitude downward, assuming dementia cases typically do not warrant mental health services compared to cases absent such pathology.

Additional Feedback:

Questions for the Committee:

- How have can the performance results be used to further the goal of high-quality, efficient healthcare? Is it too general to spur a concrete response?
- How has the measure been vetted in real-world settings by those being measured or others? Did the described TEP process seem complete and credible?
- Have unintended consequences of this measure been considered and have the developers clearly demonstrated that such negative consequences are outweighed by benefits of this measure?
- Given that the developer says the measures was successfully testing in several states, but not in CA (because of "quality issues") and RI (because of small samples), does the plan for downstream implementation seem credible?

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

RATIONALE: This is a new measure without empirical support for efficacy or safety of the measure. It has been tested in some states, but data quality and sample size issues have prevented such testing in other states. Finally, a clear plan of expansion/implementation is not articulated in the application. (Note: passing this criteria, is not a "must" for a new measure submitted to NQF; but it is a "must pass" for maintenance measures).

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 Not yet demonstrated. Future plans included state and federal level reports with stratification by large age groups (≥65 years old, and <65), and perhaps by sub-diagnostic groups (the first strata demonstrated in the application, the latter only suggested).

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

- N/A Measure has not been implemented

Potential harms

- Lack of specificity of the treatment to the disease, false positive events may suggest better quality than is really being delivered. Moreover, only 1 service is required to achieve the measure even as much more service use is likely desireable. Drugs and therapy are also treated equally.

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?
- Could be too general to be actionable or sensitive regarding underlying problems? Perhaps, for example, excessive psychotropic use masks the need for more psychosocial interventions?
    - Developer narrative notes that therapy-to-diagnosis precision may not be evidence-based (e.g., a pharm treatment might be used to treat a condition absent clinical trial support for such an indication)

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

RATIONALE:

- Refinement of measures in sub-components of illness and treatment might be most revealing and useful toward beneficial system change.

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

4a1. Use - Accountability and Transparency
Comments:

** Very concerned about the definition of mental health services as "1 visit etc". For duals with a "need" for non-acute mental health care, 1 visit is hardly a "mental health service." Perhaps there needs to be some clarification of what is meant by "need".

** It is my understanding that CMS and state Medicaid agencies routinely use this type of measure to assess performance of plans. In addition, CMS required measures such as part of the dual eligible Financial Alignment Demonstrations in recent years.

** NA since it's new. The mechanics worked adequately in Washington state.

** Had appropriate feedback.

** New measure.

** hs measure is not being publicly reported. CMS intends to use it as a core measure in its Financial Alignment Initiative. In additin to teh feedback from the Technicla Expert Panel, there was a public comment period of 3 weeks from 2 health plans hosted on CMS' online public comment system. The feedback was considerd but not incorporated.

** No performance results presented: new measure, not implemented.

** Not currently used; TEP supported but suggested dementia might warrant exclusion but the n did not affect the total. It is unclear what the impact of this measure would be (e.g. one follow up in a year seems like a low threshold).

4b1. Usability – Improvement

Comments:

** Benefits given the indication that duals seem not to be receiving non-acute mh services are clear.

** The need for quality improvement in this population is overwhelming -- both with respect to behavioral health and overall health outcomes.

** Woked OK in Washigton.

** I am not sure this is actually useful. Measuring having one visit in a year if the patient had a MH diagnoses would not be useful in my settings.

**Seems too vague/non-specific. i think better measures involve looking at screening and actively referring to appropriate services and providing appropriate treatment.

** This measure only looks at access to care and not quality of care. It will inform health plans about the success or lack of access which could help in eliminating gaps. This measue is not designed to assess the clinical appropriateness of care and there could be unintended consequences of misidentification and missed service needs which could lead to inaccuracies in use.

** Should not be interpreted to indicate quality of care, only any contact with mh services among those broadly defined as having any need which could include a prior psychiatric hospitalization/12 months. It's a bit of a stretch to infer that performance results could be "used to further the goal of high-quality, efficient healthcare.

** Uncertain if doing well on this measure is a measure of good care. States could do well even if the care provided was not in line with current treatment guidelines re: appropriate intensity.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0576 Follow-up after hospitalization in mental illness, this may be relevant as a comparative measure. Though this measure does not directly compete with the current measure, it does share substantial conceptual overlap with the measure numerator and denominator.
- No comparison to previous measures was conducted to assess construct validity.

Harmonization

None presented thus far.

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

5. Related and Competing

Comments

** No concerns.

**A measure developed by Washington State is mentioned in the submission. I would not anticipate any challenges with harmoniztion.

**None.

**No issues.

**No.

**Mental health service penetration rates (WA Dept Social and Health Services, state leg mandated measure). Not harmonized because uses 12 month "look back"after measurement year (instead of 6 month).

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/22/2019

There have been no comments or support/non-support choices as of this date.

Developer Submission

Additional evaluations and submission materials attachments…

Brief Measure Information

NQF #: 3451

Corresponding Measures:

De.2. Measure Title: Non-Acute Mental Health Services Utilization for Dual Eligible Beneficiaries

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

De.3. Brief Description of Measure: The percentage of dual eligible beneficiaries with a mental health service need who received a non-acute mental health service in the measurement year.

1b.1. Developer Rationale: Appropriate access to and use of evidence-based mental health services can reduce the probability that individuals diagnosed with mental health conditions suffer prolonged distress and may help prevent unintended consequences caused by untreated mental illness. Depression, anxiety disorders, schizophrenia/other psychotic disorders, and other bipolar disorders are among the most common behavioral health conditions among dual-eligible beneficiaries (MedPac/MACPAC, 2015).

Ensuring appropriate use of ongoing, non-acute treatment for individuals with mental illnesses could significantly improve population health and quality of life. Measurement of mental health service use for dual eligible beneficiaries with mental health needs provides important information to health plans, consumers and other stakeholders as to how well a system of care helps individuals access the resources necessary to treat their mental illness. The health plan can play a central role in improving access to timely and affordable mental health services through encouraging integration of mental health services into primary care, ensuring an adequate number of mental health professionals in their provider networks, and ensuring accurate information about these professionals is provided to individuals with mental health service needs.

Reference:

MedPAC & MACPAC. (2015). Data Book: Beneficiaries Dually Eligible for Medicare and Medicaid. Retrieved February 10, 2017 from https://www.macpac.gov/wp-content/uploads/2017/01/2015-Dually-Eligible-Beneficiaries-Data-Book.pdf

S.4. Numerator Statement: The number of dual eligible beneficiaries receiving at least one non-acute mental health service in the 12-month measurement year. The following services are included as non-acute mental health services:

- Outpatient service with a mental health provider for a mental health diagnosis
- Mental health outpatient encounter
- Mental health condition management in primary care

S.6. Denominator Statement: The number of dual eligible beneficiaries age 21 and older with a mental health service need in the 18-month identification window (the 12-month measurement year plus six months prior to the measurement year).

S.8. Denominator Exclusions: None

De.1. Measure Type: Process

S.17. Data Source: Claims

S.20. Level of Analysis: Health Plan

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:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? Not applicable.

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

Del18a_Duals_12_NQFEvidence_FINAL_10.26.18.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: Non-Acute Mental Health Services Utilization for Dual Eligible Beneficiaries

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

Date of Submission: 12/14/2018

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 subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- 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*):

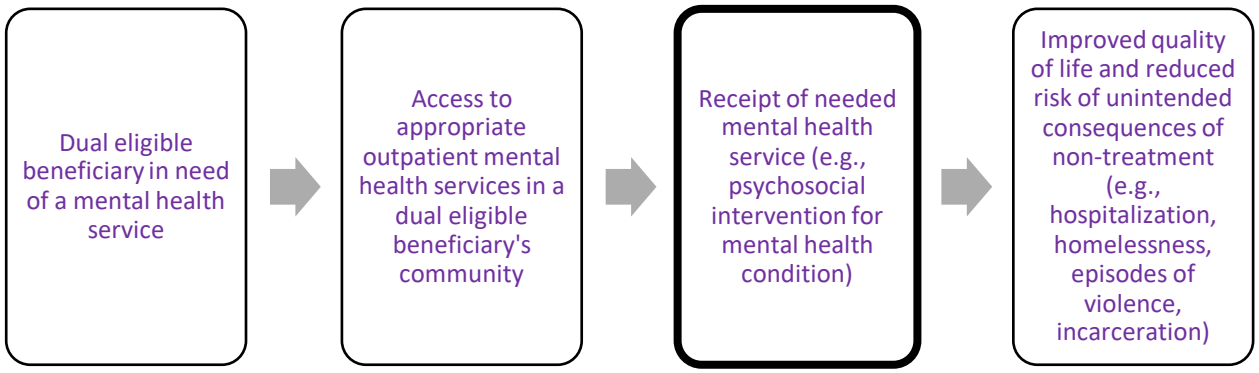
☒ Process: Receipt of appropriate mental health services for dual eligible beneficiaries with mental health needs

☐ 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.



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. This measure is not derived from a patient report.

****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.

Not applicable. This measure is not an outcome measure.

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)

☐ 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 (proceed to section 1a.4)

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.

Depression, anxiety disorders, schizophrenia/other psychotic disorders, and other bipolar disorders are among the most common behavioral health conditions among dual eligible beneficiaries (MedPac/MACPAC, 2015).[1] According to a report from the Congressional Budget Office (CBO), approximately 30% of the 7.1 million full dual eligible beneficiaries in 2009 were diagnosed with a mental illness (CBO, 2013). The proportion of dual

[1] Dual-eligible beneficiaries are people who are enrolled in Medicare and Medicaid at the same time and who are eligible to receive benefits from both programs. "Full duals" qualify for full benefits from both programs "partial duals" qualify for full benefits from Medicare but only partial benefits from Medicaid (meaning that Medicaid pays some of the expenses they incur under Medicare, such as premiums, but does not cover additional health care services, such as long-term services and supports).

eligible beneficiaries with a mental health diagnosis varied by age; approximately 37% of beneficiaries under age 65 were diagnosed with a mental illness, compared to 25% of beneficiaries age 65 or older (CBO, 2013).

Data from 2008 to 2011 indicates that the average yearly health care expenditures for dual eligible beneficiaries age 18 to 64 who received treatment for their behavioral health conditions were $16,803—twice as high as average health care expenditures among non-dual eligible beneficiaries who received treatment for behavioral health conditions ($7,860) (SAMHSA, 2014). This higher cost of care in the dual eligible population could be reflective of higher use of costly inpatient and emergency department care to treat conditions which should be managed in a community setting.

This measure assesses use of non-acute mental health care services among individuals with a mental health service need. Use of non-acute mental health care services can lead to improved quality of life and reduced risk of unintended consequences of non-treatment (e.g., hospitalization, homelessness, episodes of violence, incarceration). Below we present the evidence for the efficacy of non-acute mental health services included in this measure to treat mental health conditions.

Research on Efficacy of Mental Health Treatments

There is substantial evidence to support the use of non-acute care to treat mental health conditions. We have organized such services into four categories and describe each category in more detail below: psychosocial interventions, pharmacotherapy treatment, electroconvulsive therapy (ECT), and behavioral integration into primary care. These categories of treatment correspond to the non-acute mental health services identified in the measure numerator.

Psychosocial Interventions

Psychosocial interventions have been demonstrated to be effective in treating patients with mental illnesses (England et al., 2015). As summarized in a recent consensus study report by the National Academy of Medicine (formerly the Institute of Medicine), the efficacy of a large range of psychosocial interventions have been established through hundreds of randomized controlled trials, and includes interventions such as psychotherapies (e.g., psychodynamic therapy, cognitive-behavioral therapy (CBT), interpersonal psychotherapy, problem solving therapy), community-based treatment (e.g., assertive community treatment, first episode psychosis interventions), vocational rehabilitation, peer support services, and integrated care interventions (England et al., 2015). These interventions are effective in improving outcomes for vulnerable patients with complex conditions (e.g., improved quality of life and reduced risk of unintended consequences of non-treatment such as hospitalization, homelessness, episodes of violence and incarceration) (England et al., 2015). Furthermore, psychosocial interventions are often preferred over psychotropic treatments by patients when outcomes have similar efficacy (England et al., 2015).

For the treatment of post-traumatic stress disorder (PTSD), a review of 269 meta-analyses found that CBT, exposure therapy, and eye movement desensitization and reprocessing (EMDR) are the most effective treatments for the disorder (with each strategy being relatively comparable in terms of effectiveness and compliance) (Hofman et al., 2008).

For the treatment of bipolar disorder, a systematic review of 16 randomized-controlled trials (RCTs) found that for medicated patients, psychosocial interventions (i.e., psychoeducation, family-focused psychoeducation and CBT) in adjunct to medication seem to be the most efficacious interventions for the prevention of bipolar episode recurrences (Brenner, 2010; Fountoulakis, 2010). Additional effective non-pharmacotherapy strategies for bipolar disorder include CBT (Brenner, 2010), interpersonal and social rhythm therapy (Reinares, 2014, Cosgrove, 2013; Mizushima, 2011; Bottai, 2010).

Evidence also supports the use of psychosocial interventions for the treatment of schizophrenia. Systematic reviews of 18 randomized controlled trials conducted by the Cochrane Schizophrenia Group were consistent in their findings that family interventions and CBT are effective in decreasing relapse and readmission rates (Adams, 2000). A review summary of 41 articles concluded that there is a moderate effect size for the treatment of schizophrenia using CBT (Morrison et al., 2009). In addition to CBT, social skills training, family

psychoeducation, and cognitive rehabilitation have been shown to be effective in reducing symptoms for schizophrenia and are viewed as particularly important for the adjustment of patients when moving from institutional settings to the community (Bellack 2001; Adams 2000; Morrison et al., 2009).

Psychosocial interventions are also indicated for the treatment of anxiety disorders and depression. Using the random effects model, a meta-analysis of 27 studies found that the pooled effect size was 0.73 (95% confidence interval, 0.88–1.65) for those receiving CBT with severe anxiety, and 0.45 (90% confidence interval, 0.25–0.65) for those receiving CBT with severe depressive symptoms. CBT showed the strongest effect sizes for obsessive-compulsive disorder and acute stress disorder (Hofmann et al., 2012).

Pharmacotherapy Interventions

Six systematic reviews and a two-year longitudinal randomized controlled trial study-have shown that antipsychotic medications are efficacious in the acute and maintenance treatment of positive psychotic symptoms of schizophrenia (Davis, Chen, & Glick, 2003; Glick et al., 2011; Leucht et al., 2003; Leucht et al., 2009; Leucht, Pitschel-Walz, Abraham, & Kissling, 1999; Marder et al., 2003; Tuunainen, Wahlbeck, & Gilbody, 2002). Pharmacotherapy interventions are most effective when they are paired with effective medication management in the outpatient setting to monitor patient adherence and side effects. Continuation of medication therapy with support of medication management is associated with improved outcomes for depression (Glue et al., 2010), bipolar disorder (Sylvia et al., 2014) and schizophrenia (Jaeger et al., 2012).

Electroconvulsive Therapy

ECT has been established as a safe and effective non-pharmacological treatment for patients with mood disorders and neuropsychiatric disorders including catatonia, bipolar mania, schizophrenia and dementia with behavioral disturbance in four systematic reviews and two comprehensive literature reviews. (Dierckx et al., 2012; Loo et al., 2011; McGirr et al., 2015; Pompili et al., 2013; Wilkins et al., 2008). A systematic review consisting of 6 studies assessed the efficacy of ECT and found that the overall remission rate for unipolar depression was 50.9 percent among patients with unipolar depression, and 53.2 percent among patients with bipolar depression (Dierckx et al., 2012). Additionally, a systematic review consisting of 11 studies supported the safety of ECT for patients with unipolar, bipolar, or psychotic depression (Brunoni et al., 2014).

ECT remains effective regardless of increasing age, and data indicate that the use of ECT in the treatment of psychiatric disorders in the elderly population has increased in patients who are refractory to pharmacologic management or suffer from adverse events from medications (Wilkins et al., 2008), which may be particularly important for a dual eligible population who are on average older than non-dual Medicaid beneficiaries.

Behavioral Health Integration into Primary Care

Behavioral health integration into the primary care setting is widely considered an effective strategy for improving outcomes for individuals with mental or behavioral health conditions.. An RCT consisting of 1,801 patients suggested that integrated approaches are more comprehensive than typical primary care due to the addition of care management support for patients receiving behavioral health treatment and regular psychiatric inter-specialty consultation (Press et al., 2016; Unützer et al., 2002). Furthermore, integrated mental health care in primary care makes it easier for individuals to access mental health care and can reduce stigmas associated with seeking mental health care exclusively (Hardy et al., 2015).

Role of the Health Plan in Increasing Access to Mental Health Services

The health plan can play a central role in improving access to timely and affordable mental health services. Health plans can ensure that their network includes an adequate number of mental health professionals, and that accurate information on these professionals is provided to individuals with mental health service needs. Encouraging the integration of mental health services into primary care is another approach health plans can take to improve access to mental health services. These models include case management by a nurse or social worker with training in behavioral health. Regular psychiatric consultation allows for more regular access to mental health services without relying on exclusive treatment by psychiatrists (Goodrich, 2014). Patients with

mental health service needs may be more likely to seek care for mental health conditions in the primary care setting (Bartels, 2004).

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

Evidence was identified through a strategic review of literature available from PubMed and Google and the consultation of key literature cited. Literature relevant to the measure and related mental health treatments was incorporated and synthesized.

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

Adams, C., Wilson, P., & Bagnall, A. (2000). Psychosocial interventions for schizophrenia. *Quality in Health Care: QHC*, 9(4), 251-256.

Bartels, S., Coakley, E., Zubritsky, C., et al. (2004). Improving access to geriatric mental health services: a randomized trial comparing treatment engagement with integrated versus enhanced referral care for depression, anxiety, and at-risk alcohol use. *American Journal of Psychiatry*, 161(8), 1455-1462.

Bellack, A.S. (2001). Psychosocial treatment in schizophrenia. *Dialogues in Clinical Neuroscience*, 3(2), 136-137.

Bottai, T., Biloa-Tang, M., Chistrophe, S., Dupuy, C., Jacquesy, L., Kochman, F., Meynard, J.A., Papeta, D., Rahioui, H., Adida, M., Fakra, E., Kaladjian, A., Pringuey, D., & Azorin, J.M. (2010). Interpersonal and social rhythm therapy (IPSRT). *Encephale*, 6, S206-S217.

Brenner, R., Madhusoodanan, S., Puttichanda, S., & Chandra, P. (2010). Primary prevention in psychiatry – adult populations. *Annals of Clinical Psychiatry*, 22(4), 239-248.

Brunoni, A.R., Baeken, C., Machado-Vierira, R., Gattaz, W.F., & Vanderhasselt, M.A. (2014). BDNF blood levels after electroconvulsive therapy in patients with mood disorders: a systematic review and meta-analysis. *World Journal of Biological Psychiatry*, 15(5), 411-418.

CBO. (2013). Dual-Eligible Beneficiaries of Medicare and Medicaid: Characteristics, Health Care Spending, and Evolving Policies. Retrieved February 6, 2017 from: https://www.cbo.gov/sites/default/files/113th-congress-2013-2014/reports/44308_DualEligibles2.pdf.

Cosgrove, V.E., Roybal, D., & Chang, K.D. (2013). Bipolar depression in pediatric populations: epidemiology and management. *Paediatric Drugs*, 15(2), 83-91.

Davis, J., Chen, N., & Glick, I. (2003). A meta-analysis of the efficacy of second-generation antipsychotics. *Archives of General Psychiatry*, 60, 553-564.

Dierckx, B., Heijnen, W.T., van den Broek, W.W., & Birkenhager, T.K. (2012). Efficacy of electroconvulsive therapy in bipolar versus unipolar major depression: a meta-analysis. *Bipolar Disorders*, 14(2), 146-150.

England, M.J., Butler, A.S., & Gonzalez, M.L. (2015). Psychosocial Interventions for Mental and Substance Use Disorders: A Framework for Establishing Evidence-Based Standards. Washington, DC: The National Academies Press. Retrieved February 6, 2017 from: http://www.nap.edu/catalog/19013/psychosocial-interventions-for-mental-and-substance-use-disorders-a-framework.

Fountoulakis, K.N. (2010). An update of evidence-based treatment of bipolar depression: where do we stand? *Current Opinion in Psychiatry*, 23(1), 19-24.

Glick, I.D., Correll, C.U., Altamura, A.C., Marder, S.R., Csernansky, J.G., Weiden, P.J., et al. (2011). Mid-term and long-term efficacy and effectiveness of antipsychotic medications for schizophrenia: a data-driven, personalized clinical approach. *Journal of Clinical Psychiatry*, 72(12), 1616-1627.

Glue, P., Donovan, M.R., Kolluri, S., & Emir, B. (2010). Meta-analysis of relapse prevention antidepressant trials in depressive disorders. *Australian and New Zealand Journal of Psychiatry*, 44(8), 697-705.

Goodrich, D., Kilbourne, A., Nord, K., & Bauer, M. (2014). Mental health collaborative care and its role in primary care settings. *Current Psychiatry Reports*, 8(5), 383-388.

Hardy, L., Rosenblatt, A., Holdren, J., & Boiling, P. (2015). Integrating mental health care in a medical home for dual eligibles. *The American Journal of Geriatric Psychiatry*, 23(3), S172.

Hofmann, S.G., Asnaani, A., Vonk, I.J.J., Sawyer, A.T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cognitive Therapy and Research*, 36(5), 427-440.

Hofmann, S.G., & Smits, J.A. (2008). Cognitive-behavioral therapy for adult anxiety disorders: a meta-analysis of randomized placebo-controlled trials. Journal of Clinical Psychiatry, 69(4), 621-632.

Jaeger, S., Pfiffner, C., Weiser, P., et al. (2012). Adherence styles of schizophrenia patients identified by a latent class analysis of the medication adherence rating scale (mars): a six-month follow-up study. *Psychiatry Research*, 200(2-3), 83-88.

Leucht, S., Barnes, T., Kissling, W., Engel, R., Correll, C., & Kane, J. (2003). Relapse prevention in schizophrenia with new-generation antipsychotics. A systematic review and exploratory meta-analysis of randomized, controlled trials. *American Journal of Psychiatry*, 160, 1209-1222.

Leucht, S., Corves, C., Arbter, D., Engel, R.R., Li, C., & Davis, J.M. (2009). Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis. *Lancet*, 373(9657), 31-41.

Leucht, S., Pitschel-Walz, G., Abraham, D., & Kissling, W. (1999). Efficacy and extrapyramidal side-effects of the new antipsychotics olanzapine, quetiapine, risperidone, and sertindole compared to conventional antipsychotics and placebo. A meta-analysis of randomized controlled trials. *Schizophrenia Research*, 35(1), 51-68.

Loo, C., Katalinic, N., Mitchell, P.B., & Greenberg, B. (2011). Physical treatments for bipolar disorder: a review of electroconvulsive therapy, stereotactic surgery and other brain stimulation techniques. *Journal of Affective Disorders*, 132(1-2), 1-13.

Marder, S.R., Glynn, S.M., Wirshing, W.C., Wirshing, D.A., Ross, D., Widmark, C., et al. (2003). Maintenance treatment of schizophrenia with risperidone or haloperidol: 2-year outcomes. *American Journal of Psychiatry*, 160(8), 1405-1412.

McGirr, A., Berlim, M.T., Bond, D.J., Neufeld, N.H., Chan, P.Y., Yatham, L.N., & Lam, R.W. (2015). A systematic review and meta-analysis of randomized controlled trials of adjunctive ketamine in electroconvulsive therapy: efficacy and tolerability. *Journal of Psychiatric Research*, 62, 23-30.

MedPAC & MACPAC. (2015). Data Book: Beneficiaries Dually Eligible for Medicare and Medicaid. Retrieved February 10, 2017 from: https://www.macpac.gov/wp-content/uploads/2017/01/2015-Dually-Eligible-Beneficiaries-Data-Book.pdf

Mizushima, H. (2011). Psychoeducation and interpersonal and social rhythm therapy for bipolar disorder. *Seishin Shinkeigaku Zasshi*, 113(9), 880-885.

Morrison, A.K. (2009). Cognitive behavior therapy for people with schizophrenia. *Psychiatry (Edgmont)*, 6(12), 32-39.

Pompili, M., Lester, D., Dominici, G., Longo, L., Marconi, G., Forte, A., Serafini, G., Amore, M., & Girardi, P. (2013). Indications for electroconvulsive treatment in schizophrenia: a systematic review. *Schizophrenia Research*, 146(1-3), 1-9.

Press, M., Howe, R., Schoenbaum, M., Cavanaugh, S., Marshall, A., Baldwin, L., & Conway, P. (2016). Medicare payment for behavioral health integration. *New England Journal of Medicine*, 376, 405-407.

Reinares, M., Sanchez-Moreno, J., & Fountoulakis, K.N. (2014). Psychosocial interventions in bipolar disorder: what, for whom, and when. *Journal of Affective Disorders*, 156, 46-55.

Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality. (2014). The CBHSQ Report: Behavioral Health Conditions and Health Care Expenditures of Adults Aged 18 to 64 Dually Eligible for Medicaid and Medicare. Rockville, MD: SAMHSA. Retrieved August 17, 2018 from: http://www.samhsa.gov/data/sites/default/files/SR180/SR180.html.

Sylvia, L.G., Reilly-Harrington, N.A., Leon, A.C., et al. (2014). Medication adherence in a comparative effectiveness trial for bipolar disorder. *Acta Psychiatrica Scandinavica*, 129(5), 359-365.

Tuunainen, A., Wahlbeck, K., & Gilbody, S. (2002). Newer atypical antipsychotic medication in comparison to clozapine: a systematic review of randomized trials. *Schizophrenia Research*, 56(1-2), 1-10.

Unützer, J., Katon, W., Callahan M., et al. (2002). Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *Journal of the American Medical Association*, 288(22), 2836-2845.

Wilkins, K., Ostroff, R., & Rajesh, T. (2008). Efficacy of electroconvulsive therapy in the treatment of nondepressed psychiatric illness in elderly patients: a review of the literature. *Journal of Geriatric Psychiatry and Neurology*, 21(1), 3-11.

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)*

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

Appropriate access to and use of evidence-based mental health services can reduce the probability that individuals diagnosed with mental health conditions suffer prolonged distress and may help prevent unintended consequences caused by untreated mental illness. Depression, anxiety disorders, schizophrenia/other psychotic disorders, and other bipolar disorders are among the most common behavioral health conditions among dual-eligible beneficiaries (MedPac/MACPAC, 2015).

Ensuring appropriate use of ongoing, non-acute treatment for individuals with mental illnesses could significantly improve population health and quality of life. Measurement of mental health service use for dual eligible beneficiaries with mental health needs provides important information to health plans, consumers and other stakeholders as to how well a system of care helps individuals access the resources necessary to treat their mental illness. The health plan can play a central role in improving access to timely and affordable mental health services through encouraging integration of mental health services into primary care, ensuring an adequate number of mental health professionals in their provider networks, and ensuring accurate information about these professionals is provided to individuals with mental health service needs.

Reference:

MedPAC & MACPAC. (2015). Data Book: Beneficiaries Dually Eligible for Medicare and Medicaid. Retrieved February 10, 2017 from https://www.macpac.gov/wp-content/uploads/2017/01/2015-Dually-Eligible-Beneficiaries-Data-Book.pdf

1b.2. Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis*. (<u>This is required for maintenance of endorsement</u>. 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.*

These measures were tested in a sample of 40 Medicare-Medicaid Plans (MMPs) including 77,497 dual eligible beneficiaries continuously enrolled in fully integrated MMPs from October 1, 2015 to September 30, 2016, and for five months between April 1, 2015 to September 30, 2015. This sample included 28 MMPs with 42,397

beneficiaries between ages 21 and 64, and 36 MMPs with 34,935 beneficiaries ages 65 and older. We removed MMPs in some states that were not suitable for testing this measure given low enrollment during the time periods for which we tested data or because data was incomplete. The health plans included provide an integrated Medicare and Medicaid benefit for dual eligible beneficiaries in eight states under the Financial Alignment Initiative (FAI) demonstration.

Testing of the measure shows low performance and statistically significant variation across MMPs, suggesting that there is a gap in access to non-acute mental health services for dual eligible beneficiaries with a mental health service need. Fewer than 43% of the overall number of dual eligible beneficiaries with a mental health need accessed non-acute mental health services. Overall, the average performance of MMPs was 32.4%, with an average MMP performance of 49.6% for dual eligible beneficiaries age 21 to 64 and an average MMP performance of 21.9% for dual eligible beneficiaries ages 65 and older. The following data are restricted to MMPs with at least 30 beneficiaries in the denominator, as a minimum of 30 obsevations is required to created stable estimates for a plan's performance (12 MMPs excluded due to this requirement) . The mean performance rate presented is unweighted. See section 2b4 of the testing attachment for additional data on statistical significance testing of MMP performance rates.

MMP Performance overall:

Mean	StD	Min	Max	IQ Range	10%	20%	30%	40%	50%	60%	70%	80%	90%
32.4	15.6	5.3	72.0	20.5-41.1	13.5	17.8	23.7	27.2	31.7	37.1	38.7	43.0	52.2

MMP Performance for dual eligible beneficiaries ages 21-64:

Mean	StD	Min	Max	IQ Range	10%	20%	30%	40%	50%	60%	70%	80%	90%
49.6	11.3	19.8	72.2	44.2-55.5	36.9	40.0	45.0	48.4	49.6	51.5	53.9	57.6	63.5

MMP Performance for dual eligible beneficiaries ages 65+:

Mean	StD	Min	Max	IQ Range	10%	20%	30%	40%	50%	60%	70%	80%	90%
21.9	10.8	5.3	57.3	13.2-28.3	10.1	11.8	15.3	18.1	20.9	22.6	26.3	28.7	34.4

StD – standard deviation

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.

A similar measure, Mental Health Service Penetration developed by Washington State's Department of Social and Health Services, has reported performance for Medicaid enrollees in the state. Publicly reported data indicates less than half of Medicaid managed care beneficiaries with a mental health service need received mental health care in 2017 (44.2% of 356,222 enrollees) (Washington State Department of Social and Health Services, 2018).

Studies demonstrate that a significant number of dual eligible beneficiaries with mental illness or behavioral conditions are receiving care in inpatient institutional settings instead of the outpatient setting. One study found that in any given year, 25% of dual eligible beneficiaries with a behavioral health condition were hospitalized and approximately 12% were hospitalized two or more times – higher than the hospitalization rate in Medicare-only beneficiaries with similar conditions (Frank et al., 2014; The SCAN Foundation, 2013).

High use of inpatient care could be related to problems accessing timely and affordable outpatient care for mental health conditions.

References:

Frank, R.G., & Epstein, A.M. (2014). Factors associated with high levels of spending for younger dually eligible beneficiaries with mental disorders. Health Affairs, 33(6), 1006-1013.

The Scan Foundation. (2013). Data brief: Medicare beneficiaries with severe mental illness and hospitalization rates. Available at: http://www.thescanfoundation.org/sites/default/files/1pgdatabrief_no36_medicare_beneficiaries_with_severe_mental_illness_and_hospitalization_rates.pdf.

Washington State Department of Social and Health Services. (2018). Cross-system outcome measures for adults enrolled in Medicaid. Available at: https://www.dshs.wa.gov/sesa/research-and-data-analysis/cross-system-outcome-measures-adults-enrolled-medicaid-0.

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.*

Beneficiary age is a common risk factor for patient outcomes that providers are unable to affect. Therefore, we tested the beneficiary-level performance on the measure. Our sample consisted of dual eligible beneficiaries continuously enrolled in fully integrated MMPs from October 1, 2015 to September 30, 2016 and for 5 months between April 1, 2015 to September 30, 2015. We found dramatic and statistically significant differences in beneficiary-level performance by age. Dual eligible beneficiaries age 65 and older with a mental health service need were less than half as likely (26.8%) as dual eligible beneficiaries age 21 to 64 (55.1%) to have had a non-acute mental health service during the measurement period (note: these average beneficiary rates reflect the average score for all beneficiaries included in each category and is not limited to MMPs with a minimum of 30 beneficiaries in the measure denominator). We also tested MMP-level performance stratified by age and found that the beneficiary-level differences in performance by age were consistently reflected in the MMP-level performances (see 1b.2).

We also explored whether there was statistically significant variation in beneficiary-level performance by sex and race/ethnicity. We found that women had lower performance on the measure (38.8%) compared to men (48.4%). In addition, we found disparities by race/ethnicity. Overall, White dual eligible beneficiaries consistently had higher performance on the measure (44.1%) compared to Asian (23.5%), Black (41.1%), and Hispanic (37.2%) dual eligible beneficiaries. This observation held true for dual eligible beneficiaries age 65 and older, with White dual eligible beneficiaries having a higher rate (29.7%) than Asian (12.5%), Hispanic (13.5%), and Black (25.6%) dual eligible beneficiaries. Within the population of dual eligible beneficiaries age 21 to 64, White (56.7%), Hispanic (56.7%), and Asian (56.5%) appeared to have similar performance rates, while Black dual eligible beneficiaries experienced a lower rate (51.7%). These data suggest that there are disparities in performance for sub-groups of the measure population.

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

There are racial disparities in access to and utilization of mental health services. In the U.S. Surgeon General's first report and supplement on mental health, it was reported that compared to White individuals, individuals of racial and ethnic minorities were less likely to receive needed care, more likely to receive poor-quality care, and overall had less access to mental health services (U.S. Department of Health and Human Services, 2001). A

2001 study, which looked at survey data to assess quality of care for alcoholism, drug abuse, and mental health conditions found that African Americans (25%) and Hispanics (22.4%) were less likely to be receiving active treatment compared to non-Hispanic whites (37.6%) (Wells et al., 2001). Further, they reported that among individuals with a perceived need for mental health services, compared to Whites, African Americans were more likely to have no access to care, and Hispanics were more likely to have delayed or inadequate care. Another study looking into factors associated with detection of mental health problems found that physicians were less likely to detect mental health problems in African Americans compared with Whites (Borowsky et al., 2000).

References:

Borowsky, S.J., Rubenstein, L.V., Meredith, L.S., Camp, P., Jackson-Triche, M., & Wells, K.B. (2000). Who is at risk of nondetection of mental health problems in primary care? Journal of General Internal Medicine, 15(6), 381-388.

U.S. Department of Health and Human Services, Office of the Surgeon General. (2001). Mental health: culture, race, and ethnicity. A supplement to mental health: a report of the Surgeon General. Available at: http://www.ncbi.nlm.nih.gov/books/NBK44243/pdf/Bookshelf_NBK44243.pdf.

Wells, K., Klap, R., Koike, A., & Sherbourne, C. (2001). Ethnic disparities in unmet need for alcoholism, drug abuse, and mental health care. American Journal of Psychiatry, 158(12), 2027-2032.

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.)*

Currently not available.

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

This is not an eMeasure **Attachment:**

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

Attachment: Attachment FINAL_-_7.18.18_-_Duals12_ValueSets.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.

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.

Not applicable.

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) 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 dual eligible beneficiaries receiving at least one non-acute mental health service in the 12-month measurement year. The following services are included as non-acute mental health services:

- Mental health outpatient encounter/services
- Mental health condition management in primary care

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).

Include in the numerator all dual eligible beneficiaries receiving at least one non-acute mental health service (defined below) in the 12-month measurement year:

Non-Acute Mental Health Service Definition

A non-acute mental health service use is identified by the occurrence of any of the following three criteria:

1. Any claim with from a mental health provider with a primary diagnosis code for mental health diagnosis (Mental Health Diagnosis Value Set) AND servicing provider taxonomy code is in the set: 101Y00000X, 101YM0800X, 101YP2500X, 103G00000X, 103T00000X, 103TB0200X, 103TC0700X, 103TC1900X, 103TC2200X, 103TF0000X, 103TH0100X, 103TP0016X, 103TP0814X, 103TP2700X, 103TP2701X, 103TR0400X, 104100000X, 1041C0700X, 106H00000X, 163WP0809X, 2080P0006X, 2084A0401X, 2084F0202X, 2084N0400X, 2084N0402X, 2084N0600X, 2084P0015X, 2084P0800X, 2084P0802X, 2084P0804X, 2084P0805X, 2084S0012X, 2084V0102X, 251S00000X, 261QM0801X, 273R00000X, 283Q00000X, 323P00000X, 363LP0808X, 364SP0808X

2. Any claim with a mental health service procedure code in the following value sets (MPT IOP/PH Group 1, MPT Stand Alone Outpatient Group 1, Electroconvulsive Therapy, Transcranial Magnetic Stimulation) OR any procedure code in the following set: 90791, 90792, 90801, 90802, 90804, 90805, 90806, 90807, 90808, 90809, 90810, 90811, 90812, 90813, 90814, 90815, 90816, 90817, 90818, 90819, 90821, 90822, 90823, 90824, 90825, 90826, 90827, 90828, 90829, 90832, 90833, 90834, 90836, 90837, 90838, 90839, 90840, 90845, 90846, 90847, 90849, 90853, 90857, 90862, 90889, 96101, 96102, 96103, 96110, 96111, 96116, 96118, 96119, 96120, 90867, 90868, 90869, 90870, 90875, 90876, 96127, G0155, G0176, G0177, H0004, H0023, H0025, H0027, H0030, H0031, H0032, H0035, H0036, H0037, H0038, H0039, H0040, H0046, H1011, H2011, H2012, H2013, H2014, H2015, H2016, H2017, H2018, H2019, H2020, H2021, H2022, H2023, H2035, H2027, H2030, H2031, H2033, M0064, Q5008, S9480, S9482, S9484, S9485, T1025, T1026, T2038, T2048

3. Any claim from a primary care provider with a primary diagnosis code for mental health diagnosis (Mental Health Diagnosis Value Set) AND procedure code is in the set: 99201-99215 (Office), 99241-99255 (Consultation), or?99441-99444 (telephonic or online)

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

The number of dual eligible beneficiaries age 21 and older with a mental health service need in the 18-month identification window (the 12-month measurement year plus six months prior to the measurement year).

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).

Include in the denominator all dual eligible beneficiaries age 21 and older continuously enrolled in the 12-month measurement year and at least 5 months of the 6 months prior to the measurement year with a mental health service need (defined below) in the 18-month identification window.

Mental Health Service Need Definition

Mental health service need is identified by the occurrence of any of the following conditions:

1. Receipt of any mental health service meeting the numerator service criteria in the 18-month identification window
2. Any diagnosis of mental illness (not restricted to primary) in the 18-month identification window. These include diagnoses from the following value sets:
    a) Psychotic Diagnosis Value Set 101
    b) Mania/Bipolar Diagnosis Value Set 102
    c) Depression Diagnosis Value Set 103
    d) Anxiety Diagnosis Value Set 104
    e) ADHD Diagnosis Value Set 105
    f) Disruptive/Impulse/Conduct Diagnosis Value Set 106
    g) Adjustment Diagnosis Value Set 107
    h) Other Mental Health Diagnosis Value Set
3. Receipt of any psychotropic medication listed in the Rx Table (see attached excel spreadsheet) in the 18-month identification window. These medications comprise the following drug therapy classes:
    a) Antianxiety Rx
    b) Antidepressants Rx
    c) Antimania Rx
    d) Antipsychotic Rx
    e) ADHD Rx
4. Any claim with a mental health service procedure code in the following set: 90791, 90792, 90801, 90802, 90804, 90805, 90806, 90807, 90808, 90809, 90810, 90811, 90812, 90813, 90814, 90815, 90816, 90817, 90818, 90819, 90821, 90822, 90823, 90824, 90825, 90826, 90827, 90828, 90829, 90832, 90833, 90834, 90836, 90837, 90838, 90839, 90840, 90845, 90846, 90847, 90849, 90853, 90857, 90862, 90889, 96101, 96102, 96103, 96110, 96111, 96116, 96118, 96119, 96120, 90867, 90868, 90869, 90870, 90875, 90876, 96127, G0155, G0176, G0177, H0004,

H0023, H0025, H0027, H0030, H0031, H0032, H0035, H0036, H0037, H0038, H0039, H0040, H0046, H1011, H2011, H2012, H2013, H2014, H2015, H2016, H2017, H2018, H2019, H2020, H2021, H2022, H2023, H2035, H2027, H2030, H2031, H2033, M0064, Q5008, S9480, S9482, S9484, S9485, T1025, T1026, T2038, T2048

5. Any psychiatric inpatient stay in the following facility types: Community Psychiatric Hospital, Evaluation & Treatment Center

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

None

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.)*

None

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.)*

Measure is stratified by patient age as of the last day of the measurement period:

1. Age 21 to 64
2. Age 65 and older

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Stratification by risk category/subgroup

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 = Higher 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.)*

1. Identify the denominator – individuals with a mental health service need in the measurement year or 6 months prior to the measurement year (see S.7).
2. Stratify individuals in the denominator into age groups (i.e., 18-64, 65+) based on age on the last day of the measurement period (see S.10).
3. Among the remainder denominator population, identify the numerator – individuals who received a mental health service in the measurement year (S.5).
4. For each age group, divide the numerator population (step 3) by the denominator (step 2).

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.

Not applicable.

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.

Not applicable.

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.

Both the numerator and denominator for this measure are based on administrative claims data.

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)*

Health Plan

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

Outpatient Services, Post-Acute Care

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.)*

Not applicable.

2. Validity – See attached Measure Testing Submission Form

Duals12_NQF_Testing_Attachment.docx,Duals12_NQF_Appendix_7.18.18.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*)**:**

Measure Title: Non-Acute Mental Health Service Utilization for Dual Eligible Beneficiaries

Date of Submission: 12/14/2018

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 specifications (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 (*including 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 decision-making) 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 encounter and enrollment data from Medicare-Medicaid Plans (MMPs) in the Integrated Data Repository (IDR) to identify all MMP enrollees eligible for the measure with an indication of a mental health service need (denominator) and utilization of non-acute mental health services (numerator). The IDR is a data warehouse integrating Medicare Parts A, B, C, and D, and durable medical equipment claims; beneficiary and provider data sources; and ancillary data such as contract information and hierarchical condition category risk scores. IDR enrollment files are updated monthly, and encounter records are loaded into the IDR at scheduled times each week. We used data elements related to beneficiary enrollment status; region; institutionalization status; claims or encounters for inpatient stays, emergency department visits, observation stays, prescription drugs claims, outpatient or professional mental health service visits; and use of home and community-based services.

1.3. What are the dates of the data used in testing? April 1, 2015 – September 30, 2016

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:	☐ other:

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

The intended use of this measure is to allow CMS to evaluate the quality of care for dual eligible beneficiaries in MMPs; therefore, the data source was appropriate for the intended level of accountability.

We removed MMPs in two states which were not suitable for testing this measure, given low enrollment during the time periods for which we tested data or due to poor data quality. Specifically, we excluded plans from Rhode Island due to low enrollment in MMPs during the measurement and lookback periods. We also excluded plans from California as a result of guidance from CMS that encounter data from MMPs in this state was incomplete. We also excluded MMPs with a denominator size of fewer than 30 dual eligible beneficiaries with a mental health service need (12 MMPs), as a minimum of 30 observations is required to create stable estimates for a plan's performance. In the display of results stratified by age group, we further excluded MMPs with a denominator size of fewer than 30 dual eligible beneficiaries in each stratum. Our final analytic file included results from 40 plans in 8 states (Illinois, Massachusetts, Michigan, New York, Ohio, South Carolina, Texas, and Virginia). Table 1 describes the number of dual eligible beneficiaries included in this analysis by state. Table 2 describes the MMPs that were included in this analysis.

Table 1. Number of dual eligible beneficiaries meeting denominator criteria, by state

	Total number of unique dual eligible beneficiaries meeting denominator criteria	Total number of MMPs	Proportion of total beneficiaries
Total	77,497	40	100.0%
Illinois	16,410	8	21.2%
Massachusetts	8,538	2	11.0%
Michigan	3,244	5	4.2%
New York	1182	9	1.5%
Ohio	32,579	5	42.0%
South Carolina	352	3	0.5%
Texas	4,887	5	6.3%
Virginia	10,305	3	13.3%

MMP = Medicare-Medicaid plan.

Source: Mathematica analysis of dual eligible beneficiaries continuously enrolled in fully integrated Medicare-Medicaid plans (MMPs) from October 1, 2015 – September 30, 2016 and for at 5 months between April 1, 2015 – September 30, 2015 with a mental health need.

Note: Data included in this table are limited to MMPs with a minimum of 30 beneficiaries in the measure denominator.

Across the eight states, all 40 MMPs in operation during the analytic period were included in the testing sample (note that fewer MMPs were included in the stratified rates due to MMPs not having a minimum sample size of 30 dual eligible beneficiaries in the denominator for each age strata). Table 2 describes the number of dual eligible beneficiaries meeting the denominator criteria in MMPs by each proposed age strata.

Table 2. Descriptive characteristics of MMPs included in testing

	Number of MMPs	Total Denominator Size Across MMPs	Average Denominator Size Within MMPs	Range in Size of MMPs
Overall	40	77,497	1,983	30-8,215
21-64 Strata	28	42,397	1,514	30-7,467
65+ Strata	36	34,935	970	30-4,294

MMP = Medicare-Medicaid plan.

Source: Mathematica analysis of dual eligible beneficiaries continuously enrolled in fully integrated Medicare-Medicaid plans (MMPs) from October 1, 2015 – September 30, 2016 and for at 5 months between April 1, 2015 – September 30, 2015 with a mental health need.

Note: Data included in this table are limited to MMPs with a minimum of 30 beneficiaries in the measure denominator for each beneficiary age category.

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 MMP sample included 77,497 dual eligible beneficiaries from eight states across both age strata. The descriptive characteristics (i.e., age, sex, and race/ethnicity) of the beneficiaries meeting the denominator criteria included in the analysis are listed in Table 3 below.

Table 3. Descriptive characteristics of dual eligible beneficiaries included in testing

	21-64 Strata	65+ Strata
	Proportion of Total Strata Denominator	Proportion of Total Strata Denominator
Female	56.3%	73.5%
Male	43.7%	26.5%
White	55.9%	59.5%
Black	34.6%	28.5%
Hispanic	6.3%	6.3%
Asian	1.0%	3.5%
Other	1.1%	1.4%
Unknown	0.9%	0.6%
North American Native	0.2%	0.1%

Source: Mathematica analysis of dual eligible beneficiaries continuously enrolled in fully integrated Medicare-Medicaid plans (MMPs) from October 1, 2015 – September 30, 2016 and for at 5 months between April 1, 2015 – September 30, 2015 with a mental health need.

Note: Data included in this table are NOT limited to MMPs with a minimum of 30 beneficiaries in the measure denominator.

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. There were no differences in the data or sample used.

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.

We did not analyze social risk factors for two reasons: (1) this measure focuses exclusively on a population with social risk (i.e., dual eligible beneficiaries) and (2) patient-reported data and patient community characteristics were not available in the testing data source of administrative claims. Analysis of area-level SES-indicators (i.e., zip code) was not within the scope of measure testing. We note findings from a recent two-year National Quality Forum (NQF) effort 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 (NQF, 2017). Future measure testing may wish to examine further whether area-level SES-indicators have an impact on performance in the dual eligible population specifically.

Reference:

National Quality Forum. 2017. All-Cause Admissions and Readmissions 2015–2017. Technical Report. Available at http://www.qualityforum.org/Publications/2017/04/All-Cause_Admissions_and_Readmissions_2015-2017_Technical_Report.aspx.

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*)

This measure is a process measure of non-acute mental health service utilization among individuals with a mental health need. The intent of the measure is to provide a plan-level metric of access to non-acute mental health services for dual eligible beneficiaries with a mental health need, as an indicator of high quality plan-level care management. The measure is stratified by age group (i.e., 21 to 64 years and 65 years and older) to account for the differences in the populations enrolled in MMPs across states (i.e., some states enroll only younger dual eligible beneficiaries whereas others enroll only the older population). Testing demonstrated significant disparities in performance by age, therefore risk-stratification by age group was supported by the Technical Expert Panel (TEP) that advised on this measure. The intended use of this measure is to allow CMS to evaluate the quality of care for dual eligible beneficiaries in MMPs.

We tested reliability of the performance measure score using a signal-to-noise (SNR) analysis of the performance measure score, which quantifies the degree to which variation results from performance versus case mix and random error. In signal-to-noise reliability analyses, we calculate the ratio of signal to noise, which is the ratio of the variation in MMP-level performance rates to the total variation of the measure (which includes random fluctuation). This type of assessment addresses whether differences in measure results between MMPs are due to differences in their underlying performance or due to chance or other sources of variation. The signal variance characterizes the magnitude of differences in underlying performance between MMPs. The total variation is calculated by summing the signal variance and other random variation (noise variance)—for example, due to sampling.

We estimated SNR reliability for the measure using a beta-binomial model, which is suitable for binary measures (Adams, 2009). The measure is binary because beneficiaries receive the binary designation of either receiving or not receiving non-acute mental health services meeting the numerator definition. The beta-binomial model assumes the numerator size follows a binomial distribution conditioning on the entity's true

value, which comes from the beta distribution (ranging from 0 t o1). We calculated SNR reliability in three steps (Adams, 2009; Adams, 2014; NQF, 2011; NQF, 2016):

First, we calculated plan-specific measure variance ("noise") as a function of the measure "passing rate" at the plan level, $\hat{p}$ (passed/eligible) and the sample size, n:

$$\sigma^2_{within} = \frac{\hat{p}(1-\hat{p})}{n}; \text{ (1)}$$

Second, we used version 2.2 of the BETABIN SAS macro to fit the beta-binomial model to the measure dataset (Wakeling, n/d). The macro produced the estimated average pass rate across all plans, as well as the Alpha (α) and Beta (β) parameters that describe the shape of the fitted beta-binomial distribution. We calculated the "signal" (between-plan variation of the measure) using these parameters, as follows:

$$\sigma^2_{between} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}; \text{ (2)}$$

Third, we calculated the SNR reliability as the ratio of the between-plan variance and the total variance (i.e., the sum of the between-plan and within-plan variances) of the measure rate:

$$SNR = \frac{\sigma^2_{between}}{\sigma^2_{between}+\sigma^2_{within}}. \text{ (3)}$$

References:

Adams, J.L. "The Reliability of Provider Profiling: A Tutorial." TR-653-NCQA. Santa Monica, CA: RAND Corporation, 2009.

Adams, J.L. 2014. Reliability-Testing Concepts. National Quality Forum presentation. www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=74717 (accessed February 23, 2017).

National Quality Forum. Measure Testing Task Force Report (January 2011). www.qualityforum.org/Publications/2011/01/Measure_Testing_Task_Force.aspx (accessed February 23, 2017).

National Quality Forum. NQF-Endorsed Measures for Cardiovascular Conditions, 2015-2016. Final Report. http://www.qualityforum.org/Publications/2016/05/Cardiovascular_Conditions,_2015-2016_-_Final_Report.aspx (accessed February 23, 2017).

Wakeling, Ian (n/d). SAS Macro for fitting Beta-Binomial models, written by Ian Wakeling. http://www.qistats.co.uk/BetaBinomial.html (accessed February 23, 2017).

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

Table 4. Distribution of MMP reliability scores, by beneficiary age stratification

	Mean	Median	Min	25th Pctile	75th Pctile	Max
Beneficiaries age 21-64	0.94	0.98	0.61	0.97	0.99	1.00
Beneficiaries age 65+	0.92	0.97	0.62	0.87	0.99	0.99

Source: Mathematica analysis of dual eligible beneficiaries continuously enrolled in fully integrated Medicare-Medicaid plans (MMPs) from October 1, 2015 – September 30, 2016 and for at 5 months between April 1, 2015 – September 30, 2015.

Note: The total measure score reflects the unweighted average of all included MMP scores. Data included in this table are limited to MMPs with a minimum of 30 beneficiaries in the measure denominator for each beneficiary age category with a mental health need.

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 SNR statistic, R (ranging from 0 to 1), summarizes the proportion of the variation between entity scores that is due to true differences in quality of care as opposed to chance or other source of variation (for example, due to measurement or sampling error). If R=0, there is no variation on the underlying performance across the measure reporting 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.

Table 4 displays the distribution of MMP-level SNRs for each age cohort and potential numerator definition. Although mean SNRs are slightly lower for the over 65 population (0.92-0.95) compared to the 21 to 64 population (0.93-0.95), the mean SNR was at least 0.92 for all numerator definitions. The 25th percentile of MMP-level reliability scores was 0.97 for all numerator definitions for the 21 to 64 population, and at least 0.87 for the over 65 population. The generally-accepted threshold for being able to reliably distinguish between group-level performance is 0.7 (Adams, 2009; Adams, 2014). These data indicate that that this measure can reliably discern performance between plans for this population.

References:

Adams, J.L. "The Reliability of Provider Profiling: A Tutorial." TR-653-NCQA. Santa Monica, CA: RAND Corporation, 2009.

Adams, J.L. 2014. Reliability-Testing Concepts. National Quality Forum presentation. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=74717 (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.

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)*

This measure is a process measure to track non-acute mental health service utilization. The intent of the measure is to provide a plan-level metric of access to non-acute mental health services for dual eligible beneficiaries with a mental health need, as an indicator of high quality plan-level care management. The measure is stratified by age group (i.e., 21 to 64 years and 65 years and older) to account for the differences in the populations enrolled in MMPs across states (i.e., some state enroll only younger dual eligible beneficiaries whereas others enroll only the older population).

We solicited input from the TEP on the face validity of this measure (see Appendix Table A.1 for the 2017 TEP member list). Face validity is a subjective assessment by experts about whether the measure reflects what it is intended to assess. It addresses whether performance scores resulting from the measure can be used to distinguish good from poor quality. Following a discussion among TEP members about the measure's face validity, we asked them to individually complete a formal survey after the call to quantify the group's overall assessment of the following questions:

- Is the denominator for this measure appropriate, given the intent?
- Is the numerator for this measure appropriate, given the intent?
- Are the exclusions appropriate, given the intent of this measure?

- Would high performance on this measure indicate that a health plan is providing higher quality care management?
- Do you think that performance scores on this measure will distinguish between good and poor performance?

The assessment was conducted through an online review process using a web-based questionnaire (developed using SurveyMonkey®). Face validity of the measure score as an indicator of quality was systematically assessed as follows: After the measure was fully specified and tested, the expert panel members were asked to rate, based on a 4-point Likert-type scale, their level of agreement with the following statement: "The measure appears to measure what is intended." The 4-point Likert-type scale was defined as follows: 1=Strongly Disagree; 2=Disagree; 3=Agree; 4=Strongly Agree

We also received feedback on the measure's validity from a workgroup which advised on the technical details of the measure (i.e., numerator and denominator specification – see Appendix Table A.2). Feedback on the measure's validity was also received through a three-week public comment period hosted on CMS's online public comment system. The public comment period was open and broadcast to all interested parties. Overall, commenters offered general support for the measure.

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

The results of the Technical Expert Panel rating of face validity are listed below by the statements posed.

"This denominator is appropriate given the intent of the measure," on a scale of 1-4.

N=6 panel members, Mean Rating=4

Response	% of TEP	Number of TEP Members
Strongly Agree	100%	6
Agree	0%	0
Disagree	0%	0
Strongly Disagree	0%	0

1. "The numerator is appropriate given the intent of the measure," on a scale of 1-4.
   N=5 panel members, Mean Rating=3.2

Response	% of TEP	Number of TEP Members
Strongly Agree	20%	1
Agree	80%	4
Disagree	0%	0
Strongly Disagree	0%	0

2. Quality of care. "Would high performance on this measure indicate that a Medicare-Medicaid plan is providing higher quality care management than a Medicare-Medicaid plan with low performance on this measure?" on a scale of 1-4.

   N=6 panel members, Mean Rating=3.33

Response	% of TEP	Number of TEP Members
Strongly Agree	33%	2
Agree	67%	4
Disagree	0%	0
Strongly Disagree	0%	0

3. Distinguishing performance. "Do you think that scores on this measure will distinguish between good and poor Medicare-Medicaid plan performance?" on a scale of 1-4.

N=6 panel members, Mean Rating=3.33

Response	% of TEP	Number of TEP Members
Strongly Agree	33%	2
Agree	67%	4
Disagree	0%	0
Strongly Disagree	0%	0

In addition to the TEP feedback, we received comments on this measure during public comment. The majority of commenters supported the *Non-Acute Mental Health Service Utilization for Dual Eligible Beneficiaries* measure and proposed specific recommendations. One commenter did not support the measure, citing concerns about measure intent and use, and provided feedback that the measure definition is too broad to provide information useful to determine quality or meaningful outcomes. However, another commenter noted that the measure attempts to address a "unique and currently unmet niche"—mental health needs—and that most other measures have a more restricted focus on follow-up care, medication management, and integration of behavioral health and physical health. We also received several comments about the measure specification, including potential populations for exclusion or stratification, and the time period during which mental health needs are identified for the measure. We incorporated this feedback into our testing plan.

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?*)

In summary, 100% of the TEP members responded "agree" or "strongly agree" with the statements that the measure has face validity. In public comment, 4 out of 5 (or 80%) of commenters supported the measure. The results indicate strong support of the face validity of the measure.

_____

2b2. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

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*)

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*)

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*)

_____

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 2b5.

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

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with 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.

Not applicable. No statistical risk model was used.

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?

Not applicable.

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?
Not applicable.

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.**

Not applicable.

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*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not applicable.

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable.

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable.

2b3.9. Results of Risk Stratification Analysis:

Not applicable.

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*)

Not applicable.

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*)

_____

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)*

This measure is stratified by age for two reasons: 1) to allow for more appropriate comparison of MMP performance by accounting for differences in the population eligible to enroll in MMPs by state[2]; and, 2) to assist MMPs in targeting improvements in access to mental health services toward the elderly population, where the performance gap is greatest.

We evaluated variation in MMP-level performance for the 21 to 64 age cohort and the 65 and older age cohort (Table 6) by calculating the 95 percent confidence interval of the measure score for each MMP and compared the confidence interval range to the weighted average of all MMPs. To identify statistically significant differences in an MMP's performance compared to the mean, we chose a global mean (weighted average) to reflect the experience of the entire population. The weighted average of MMP-level performance reflects the average score for all beneficiaries included in each MMP's denominator for MMPs with at least 30 beneficiaries in the measure denominator.

If the entire range of the confidence interval for an MMP score is lower than the weighted average for all MMPs included in the analysis, then the MMP score is statistically worse than average. Similarly, if the entire range of the confidence interval for an MMP score is higher than the weighted average for all MMPs included in the analysis, then the MMP score is statistically better than average. Finally, if the confidence interval for an MMP score includes the weighted average for all MMPs included in the analysis, then the MMP score is statistically no different than average.

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*)

Table 5. MMP measure performance, by beneficiary age

	# Beneficiaries in the denominator	# MMPs	Average MMP rate	MMP interquartile range
Total	77,332	39	32.4%	20.5-41.2%
Age: 21-64	42,397	28	49.6%	44.2-55.5%
Age: 65+	34,935	36	21.9%	13.2-28.3%

Source: Mathematica analysis of dual eligible beneficiaries continuously enrolled in fully integrated Medicare-Medicaid plans (MMPs) from October 1, 2015 – September 30, 2016 and for at 5 months between April 1, 2015 – September 30, 2015.

Note: The average MMP rate reflects the unweighted average of all included MMP scores. Data included in this table are limited to MMPs with a minimum of 30 beneficiaries in the measure denominator for each beneficiary age category with a mental health need.

[2] In South Carolina, the under-65 population is excluded from MMP enrollment, and in Massachusetts the over-65 population is excluded from MMP enrollment.

Table 6. Summary of MMP measure performance

	Age 21-64	Age 65+
Number of MMPs	28	36
MMP denominator interquartile range	549.5 - 2,177.75	134.75 - 1,326
Weighted average MMP performance rate	55.1%	26.8%
Unweighted average MMP performance rate	49.6%	21.9%
MMP performance rate interquartile range	44.2-55.5%	13.3-28.1%
…		
MMPs performing significantly better than weighted MMP average	7	7
MMPs performing no different from weighted MMP average	5	10
MMPs performing significantly worse than weighted MMP average	16	19

CI = confidence interval

MMP = Medicare-Medicaid plan.

Source: Mathematica analysis of dual eligible beneficiaries continuously enrolled in fully integrated Medicare-Medicaid plans (MMPs) from October 1, 2015 – September 30, 2016 and for at 5 months between April 1, 2015 – September 30, 2015.

Note: Data included in this table are limited to MMPs with a minimum of 30 beneficiaries in the measure denominator for each beneficiary age category with a mental health need.

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?*)

We observed dramatic differences in dual eligible beneficiary-level performance by age (Table 5). Dual eligible beneficiaries age 65 and older with a mental health service need were less than half as likely as younger dual eligible beneficiaries to have had a non-acute mental health service during the measurement period (average performance 26.8% in 65 and older population compared to 55.1% in the 21 to 64 population).

For the 21 to 64 age cohort, at least three-fourths of the MMPs had scores that were statistically significantly different than the weighted average for all numerator definitions. For the 65 and older age cohort, at least two-thirds of the MMPs had scores that were statistically significantly different than the weighted average for all numerator definitions.

Although statistically significant variation in MMP-level performance is greater among the 21 to 64 age cohort compared to the 65 and older age cohort, the results indicate that there is substantial variation in measure performance for both cohorts at the MMP level. (Detailed MMP-level results are provided in Appendix Tables B.1 and B.2.)

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. Only one set of specifications provided.

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. Only one set of specifications provided.

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. Only one set of specifications provided.

_____

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 evaluated the proportion of encounter and enrollment records where key data needed to calculate the measure were missing to determine whether the measure is feasible for MMPs to implement. We found that none of the key data elements were missing from the encounter data in any of the time periods studied. Claims data are large and real-time collected repositories of data that have been used for many years to determine which services were provided, and we are using them as such without reference to any other standard.

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*)

Not applicable. No key data elements were missing.

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*)

Not applicable. No key data elements were missing.

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

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.

None of the key data elements were missing from the encounter data in any of the time periods studied, indicating that the measure is feasible for MMPs to implement.

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)*.

Value sets included in the attached Value Set Directory (VSD) are developed by and are owned by the National Committee for Quality Assurance (NCQA). NCQA holds a copyright in the value sets and may rescind or alter the value sets at any time. Users shall not have the right to alter, enhance or otherwise modify the value sets, and shall not disassemble, recompile or reverse engineer the value sets. Anyone desiring to use or reproduce the value sets without modification for a non-commercial purpose may do so without obtaining any approval from NCQA. All commercial uses or requests for alteration must be approved by NCQA and are subject to a license at the discretion of NCQA. The value sets are provided "as is" without warranty of any kind.

Proprietary coding is contained in the attached list of codes. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets.

Current Procedural Terminology (CPT) codes copyright 2018 American Medical Association (AMA). All rights reserved. CPT is a trademark of the AMA. 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.

The American Hospital Association (AHA) holds a copyright to the Uniform Bill Codes (UB) contained in the measure specifications. The UB Codes in the HEDIS specifications are included with the permission of the AHA. The UB Codes contained in the HEDIS specifications may be used by health plans and other health care

delivery organizations for the purpose of calculating and reporting HEDIS measure results or using HEDIS measure results for their internal quality improvement purposes. All other uses of the UB Codes require a license from the AHA. Anyone desiring to use the UB Codes in a commercial Product(s) to generate HEDIS 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.

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).

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)
Public Reporting Quality Improvement (external benchmarking to organizations) 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?*)

This is a new measure which has not been implemented yet.

Use :

The measure under evaluation is a newly developed measure. The measure is intended for use in Medicare-Medicaid Plans (MMPs) and was tested with the full available sample of MMPs. Testing used all available MMP data that met certain quality standards and included eight states. Only MMP data from California and Rhode Island were excluded; California for data quality and Rhode Island for sample size. Results from testing suggest that the measure is feasible, usable, and effective.

Safety (Unintended Consequences):

As with all measures, there may be unintended consequences. For this measure, unintended consequences could include: 1) the potential for misidentification and missed service needs, and 2) harms of overuse for therapies that have a narrow therapeutic interval, such as pharmacotherapy.

If a member does not have the criteria that this measure identifies as a mental health service need, he/she may not be given the attention intended. In order to minimize the possibility of misidentification, the measure includes various options for the member to meet denominator criteria.

In contrast, the misidentification of individuals into the denominator may present harms of overuse. Various treatment options (i.e. pharmacotherapy and ECT), have a narrow therapeutic interval and should be prescribed and/or performed judiciously. The quality measure is not designed to assess the clinical appropriateness of treatment use, and should not supersede shared decision making with patients about the risks and benefits of proposed treatments. However, we note that several mental health treatment options that satisfy the measure's numerator criteria are without known harms.

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.*)

This measure is planned for potential implementation in CMS Financial Alignment Initiative (FAI) core measure set for Medicare-Medicaid Plans (MMPs). This set of measures is used to monitor and evaluate the quality of care provided in MMPs participating in the FAI. These measures will be publicly reported and used for quality improvement. At a future point, this measure could also be used for payment incentives as part of a quality withhold arrangement and for states participating in the Managed Fee For Service component of the FAI demonstration.

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.

Measure specification and performance results from testing have been presented to a Technical Expert Panel (TEP) and expert workgroup. These groups include representatives from national associations of state health officers, two individuals representing two states out of the eight in the MMP demonstration, and two representatives from health plans, selected based on convenience and prior experience as participating key experts.

The measure specifications also received feedback from two health plans through a three-week public comment period hosted on CMS's online public comment system.

Measure performance results specific to each Medicare-Medicaid Plan (MMP) were not provided back to the MMPs. However, representatives from MMPs participated in the TEP and expert workgroup described above and provided feedback on the measure importance and construction.

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.

During measure development, the expert workgroup was convened twice to provide input on the measure specification and testing results, and the TEP was convened once to provide input on the measure specification following testing. Members were presented with the measure description, intent, detailed specifications, and findings (from testing). Materials posted for public comment included the measure specification and justification.

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.

Feedback from the expert workgroup and TEP was obtained with open discussion following presentation of the measure specification, testing results. For the TEP, we also distributed a survey to the members following the presentation to ask about potential alternative methods of constructing the measure.

Feedback on the measure was also received through a three-week public comment period hosted on CMS's online public comment system. The public comment period was open and broadcast to all interested parties.

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

The expert workgroup recommended additional conditions and value sets to include in both the measure denominator and numerator. They also supported a reduction in the lookback period from 12 months to 6 months, making the denominator window 18 months total. Both the TEP and expert workgroup agreed with the recommendations by the measure development team to not add exclusions to the measure specifications, and to stratify by age. All (100%) f the TEP members responded "agree" or "strongly agree" with the statements that the measure has face validity.

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

From public comment, the majority of respondents supported the measure while proposing specific recommendations. One commenter did not support the measure, citing concerns about measure intent and use, and provided feedback that the measure definition is too broad to provide information useful to determine quality or meaningful outcomes. However, another commenter noted that the measure attempts to address a "unique and currently unmet niche"—mental health needs—and that most other measures have a more restricted focus on follow-up care, medication management, and integration of behavioral health and physical health. We also received several comments about the measure specification, including potential populations for exclusion (dual eligible beneficiaries with dementia or delirium) or potentioal stratifications (dual eligible beneficiaries using LTSS), and the denominator time period during which mental health needs are identified for the measure.

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.

Feedback received from the TEP, expert workgroup and public comment were incorporated into the testing plan and final measure specifications. Additional value sets, conditions, and health services identified by the expert workgroup were added to the measure specifications. We explored a denominator time period of 18 months, as supported by the workgroup and public comment, and revised the measure specification to use a six-month look back period prior to the measurement year instead of 12 months.

Other suggestions from public comment (i.e., the measure should exclude dual eligible beneficiaries with dementia or delirium) were included in testing and discussed with the measure TEP. After testing, we recommended against incorporating these exclusions in the measure specification because we believe these beneficiaries can appropriately be included in the measure. We found that excluding dual eligible beneficiaries with dementia or developmental disorder diagnosis from the measure had a negligible impact (less than 0.1 percentage points) on average MMP performance. There were very few beneficiaries with a developmental disorder in the MMP population, and differences in MMP performance for the population with dementia reflected differences in performance by age rather than dementia status. The TEP supported our recommendation to not exclude these beneficiaries, and also agreed with not excluding beneficiaries who were institutionalized or used home and community-based services from the measure. The final measure specification aligns with these recommendations.

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 not currently implemented, so longitudinal data is not available. Measurement of mental health service use for dual eligible beneficiaries with mental health needs will provide important information to health plans, consumers and other stakeholders as to how well a system of care helps individuals access the resources necessary to treat mental illness. Performance results can be used to help health plans identify and target interventions to dual eligible beneficiaries with mental health needs, encouraging health plans to improve access to high quality outpatient mental health services. The health plan can play a central role in improving access through encouraging integration of mental health services into primary care, ensuring an adequate number of mental health professionals, and ensuring accurate information on these professionals is provided to individuals with mental health service needs. Mental health services use among dual eligible beneficiaries with mental health needs will improve health and quality of life, and reduce the risk of unintended consequences caused by non-treatment.

This measure is specified with the intention to encompass all mental health service needs and utilization, and is intended for reporting at the state and national level. Public readability may be compromised if additional subcomponents, beyond the age strata are added. However, we believe that stratification by condition could be helpful for internal quality improvement programs. In contrast to Medicare Advantage (MA) plans, MMPs are given the freedom to identify their own quality improvement programs to address issues specific to their population. Therefore, adding additional levels of granularity would be most useful on a program-by-program basis and can be explored post-measure implementation.

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.

Not applicable. This measure is not yet implemented.

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

Not applicable. This measure is not yet implemented.

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)

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

Mental Health Service Penetration (Washington State Department of Social and Health Services).

This measure was based on a non-NQF endorsed measure used by the Washington State Department of Social and Health Services. The Mental Health Service Penetration measure evaluates access to behavioral health care among Medicaid enrollees with identified behavioral health service needs and was developed as directed by Engrossed House Bill 1519 (Chapter 320, Laws of 2013) and Second Substitute Senate Bill 5732 (Chapter 338, Laws of 2013) (Mancuso, 2016).

Based on feedback from our TEP, expert workgroup and public comment, the proposed measure varies slightly from the measure developed by Washington State. The proposed measure uses an 18-month denominator period (one measurement year with a six-month look back period), whereas the Washington State measure uses a two-year denominator period (one measurement year and a 12-month look back period). This difference was recommended to expand the eligible population enrolled in MMPs, as dual eligible beneficiaries often experience changes in MMP enrollment. Additionally, the list of conditions for the denominator and list of mental health services for numerator are more inclusive in this measure compared to the Washington state measure. This change was recommended to improve the measure's relevance to the dual eligible population, although it is acknowledged that the measure is limited in that it assesses service utilization but not necessarily appropriateness of service. A representative from the Washington State developer team clarified that the mental health services included in the Washington State measure reflect services covered by Washington's Medicaid program.

It is not anticipated that these differences will increase or at all impact data collection burden.

Reference:

Mancuso, D. 2016. Behavioral health access to care metrics: Illustration of the impact of case-mix adjustment. Washington State Department of Health and Human Services. Available at: https://www.dshs.wa.gov/sites/default/files/SESA/rda/documents/research-3.43_0.pdf.

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?

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

Not applicable. There are no related NQF-endorsed measures.

5b. Competing Measures

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

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Not applicable. This measure does not address both the same measure focus and same target population as another NQF-endorsed measure.

Appendix

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

Attachment Attachment: Del18a_Duals12_NQFEvidence_FINAL_10.22.18.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid 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: Henry, Ireys, HIreys@mathmatica-mpr.com, 202-554-7536-

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.

Innovation Accelerator Program Technical Expert Panel (TEP) – advised on the measure specification following testing:

1. Carol McDaid, Capitol Decisions, Inc.
2. Janice Tufte, Patient-Centered Outcomes Research Institute ambassador
3. Kayte Thomas, Patient-Centered Outcomes Research Institute ambassador
4. Joe Parks, Missouri HealthNet Division (Medicaid)
5. David Mancuso, Washington State Department of Social and Health Services
6. Roxanne Kennedy, New Jersey Division of Mental Health and Addiction Services
7. Alonzo White, Aetna Medicaid
8. Deb Kilstein, Association for Community Affiliated Plans
9. Jim Thatcher, Massachusetts Behavioral Health Partnership, Beacon Health Options
10. Daniel Bruns, Health Psychology Associates
11. Aaron Garman, Coal Country (ND) Community Health Center (and American Academy of Family Practice Comm. on Quality & Practice)
12. Annette DuBard, Aledade
13. Andrew Bindman, University of California San Francisco (incoming director of the Agency for Healthcare Research and Quality)
14. Mady Chalk, Treatment Research Institute
15. Kimberly Hepner, RAND Corporation
16. Benjamin Miller, University of Colorado School of Public Health

17. Alex Sox-Harris, Department of Veterans Affairs
18. Deb Potter, Office of the Assistant Secretary for Planning and Evaluation
19. Laura Jacobus-Kantor, Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality

Non-Acute Mental Health Service Utilization Expert Workgroup – advised on the measure specification and testing:

1. David Mancuso, Washington State Department of Social and Health Services
2. Raina Josberger, New York State Department of Health
3. Deb Kilstein, Association of Community Affiliated Plans
4. Janice Tufte, Patient-Centered Outcomes Research Institute ambassador
5. Robert Roca, Sheppard Pratt Heath System

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? Not applicable. This is a new measure.

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement: Not applicable. This measure is in the public domain.

Ad.7 Disclaimers: None

Ad.8 Additional Information/Comments: None