PQM Measure Evaluation Rubric

Note: Rubric items correspond to items in the measure submission form and provide the information needed to evaluate each of the five Rubric domains.

The requirements for initial and maintenance measure endorsement are indicated as, "*[For initial endorsement]*" or "*[For maintenance],*" within each domain of PQM Measure Evaluation Rubric. If neither distinctions are listed for a rubric requirement, then it applies to both initial and maintenance endorsement.

The PQM Measure Evaluation Rubric does not include must-pass criteria, nor algorithms for assigning a rating. Rather, the PQM Measure Evaluation Rubric guides reviewers to a rating of "Met", "Not Met, but Addressable", or "Not Met" based on the criteria listed for each. As part of its continuous quality improvement of the E&M process, Battelle considers whether changes to the domains, criteria, and/or additional guidance, such as an algorithm, are needed.

Importance
Attach a logic model depicting the relationship between structures and processes and the desired outcome.
Summarize evidence of measure importance from the literature linking the structure/process/intermediate outcome to the outcome
[For initial endorsement] If implemented, what is the measure's anticipated impact on important outcomes?
[For maintenance] Provide evidence of performance gap or measurement gap by providing performance scores on the measure as specified (current and over time) at the specified level of analysis
Explain why existing measures/quality improvement programs are insufficient for addressing this health care need?
Provide evidence the target population (e.g., patients) values the measured outcome, process, or structure, and finds it meaningful. Describe how and from whom you obtained input.

Not Met:

- Evidence is about something other than what is measured OR
- Empirical evidence submitted without literature review or grading OR
- Empirical evidence includes only selected studies from the literature review[2] OR

[2] A literature review could include a systematic review, clinical practice guidelines, observational studies, case studies, etc. The purpose of the literature review is to identify relevant studies to support the measure's logic model. Developer/stewards should provide a summary of the evidence for the committee's consideration. An evaluation of the quality of evidence should also be conducted. Often clinical practices guidelines conduct systematic reviews. If a literature review is not possible, a rationale as to why would be considered by the committee.

- Evidence is not graded high quality or strong recommendation OR
- Literature review conclusion is that consistency is low or controversial; moderate/high certainty that the net benefit (i.e., improved outcomes, adverse events and/or costs avoided due to the measure's anticipated impact) is null or small; or grade of weak OR
- There is low confidence/certainty that there is an adequate business case[3] (the anticipated impacts of the measure on patient outcomes and/or costs/resource use justify the measure and its use), where "adequate"=there is a net benefit to measurement OR
- There is low confidence/certainty that there is evidence of a performance gap, as determined by variation in performance or less than optimal performance for the overall target population and/or subpopulations OR
- There is no description of other existing measures or programs or no search conducted to identify other existing measures or programs OR
- Proposed measure has the same measure focus and target population as existing measures and offers no advantage in terms of addressing disparities, feasibility, potential use, or scientific acceptability OR
- Patient input does not support the conclusion that the measured outcome, process, or structure is meaningful or it does so with a low degree of certainty.

Not Met but Addressable:

- Criterion is not met (see above), but the reviewer can identify changes to specifications that may strengthen the measure's importance such that the criterion could be met.

Met:

- Literature review concludes with at least moderate certainty that a net benefit (i.e., improved outcomes, adverse events and/or costs avoided due to the measure's anticipated impact) is at least moderate AND
- There is at least moderate confidence/certainty that there is an adequate business case (i.e., the anticipated impacts of the measure on patient outcomes and/or costs/resource use justify the measure and its use), where "adequate"=there is a net benefit to measurement AND
- There is at least moderate confidence/certainty that there is evidence of a performance gap, as determined by variation in performance or less than optimal performance for the overall target population and/or subpopulations AND

[3] For more information on how to consider the business case for a measure, please refer to the CMS Measure Management System Blueprint

- Description of existing measures or programs justifies the proposed measure's focus among the proposed measure's target population and/or the proposed measure is superior[4] to identified related or competing measures AND
- Description of patient input supports the conclusion that the measured outcome, process, or structure is meaningful with at least moderate certainty.

Feasibility
[For Initial Endorsement] Describe the feasibility assessment showing you considered the people, tools, tasks, and technologies necessary to implement this measure. If an eCQM, please attach your completed eCQM Feasibility Scorecard.
Describe how the feasibility assessment informed the final measure, indicating any decisions made to adjust the measure in response to data availability.
Describe any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, algorithm.

Not Met:

- Feasibility assessment not systematically conducted or described OR
- Long-term or no path is specified to support routine and electronic data capture with an implementable data collection strategy.

Not Met but Addressable:

- Criterion is not met (see above), but the reviewer can identify changes to specifications that may improve feasibility such that the criterion could be met.

Met:

- Near-term paths are specified to support routine and electronic data capture with an implementable data collection strategy OR
- Required data are routinely generated and used during care, required data are available in EHRs or other electronic sources, and the data collection strategy can be implemented.

[4] Measure developers/stewards must document why the proposed measure is superior to any identified and/or competing measures and should include any literature used to support this position. For instance, clinical practice guidelines supporting the proposed measure do not support any existing measures identified; or the proposed measure's intentions vary across programs/payors, which requires the measure to be distinct from other existing measures; or the proposed measure captures a target population at higher risk such at the use of the proposed measure may close care gaps for a higher-risk population.

Scientific Acceptability
Describe the data or sample used for testing (include dates, source). If you used multiple data sources for different aspects of testing (e.g., reliability, validity, risk adjustment), identify how the data or sample are different for each aspect of testing.
Provide descriptive characteristics of measured entities included in the analysis (e.g., size, location, type). If you used a sample, describe how you selected entities for inclusion in the sample.
Identify the number and descriptive characteristics (e.g., age, sex, race, diagnosis), of the unit of analysis, for example, patient, encounter or episode, separated by level of analysis and data source. If you used a sample, describe how you selected the patients for inclusion in the sample. If there is a minimum case count used for testing, you must reflect that minimum in the specifications.
If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), please identify how the data or sample are different for each aspect of testing.
Select the level of reliability testing conducted. ☐ Patient or Encounter-Level (e.g., inter-abstractor reliability) ☐ Accountable Entity Level (e.g., signal-to-noise analysis)
For each level of reliability testing conducted, describe the method of reliability testing and what it tests.
Provide the statistical results from each level of reliability testing conducted and at the measure's level of analysis (e.g., clinician, health plan, facility).
Provide your interpretation of the results in terms of demonstrating reliability (i.e., How do the results support an inference of reliability for the measure?)
Select the level of validity testing conducted. ☐ Patient or Encounter-Level (e.g., sensitivity and specificity) ☐ Accountable Entity Level (e.g., criterion validity)
Select the type of validity testing conducted. ☐ Empirical validity testing (e.g., data element testing, empirical testing of measure score) ☐ Systematic assessment of face validity of performance measure score as an indicator of quality or resource use (i.e., the score is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance).
For each level of testing conducted, describe the method of validity testing and what it tests.
Provide your interpretation of the results in terms of demonstrating validity (i.e., How do the results support an inference of validity for the measure?)
Check all methods used to address risk factors. ☐ Statistical risk model with risk factors (___Specify number of risk factors) ☐ Stratification by risk category (___Specify number of categories) ☐ Other (___Specify) ☐ No risk adjustment or stratification

Scientific Acceptability
Attach a conceptual model illustrating the pathway between patient risk factors (social, functional status-related, and clinical factors), quality of care, and the measured outcome. Explain the rationale for the model.
Provide descriptive statistics on the distribution across the measured entities of the risk variables identified in the conceptual model.
If using statistical risk models, provide detailed risk model specifications (query or algorithm), including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.
Detail the statistical results of the analysis used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.
Provide the approach and results of calibration and discrimination testing. Describe any over- or under-prediction of the model for important subgroups.
If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate there is no need to control for differences in patient characteristics (i.e., case mix) to achieve fair comparisons across measured entities.

Not Met:

Sampling

- Sampling is used and sampling strategy is not determined by the measure's analytic unit OR sample does not represent variety of entities whose performance will be measured OR sample does not include adequate numbers of units of measurement for the selected statistical method OR

For Patient or Encounter Level Reliability[5]

- Internal consistency < 0.7 OR
- Inter-rater agreement < 0.4 OR
- Test-retest reliability (Intraclass correlation or Pearson correlation) < 0.5 OR
- Linear relationship < 0.6 OR

For Accountable Entity Level Reliability[5,6]

- Signal to noise/Inter-unit Reliability < 0.6 OR

[5] Reliability thresholds were established by the Scientific Methods Panel and confirmed at the June 14, 2022 advisory meeting.

[6] For accountable entity level reliability testing, the associated thresholds apply to the accountable entity (e.g., facility, clinician, health plan), not the mean or median across all entities.

- Split-half reliability (ICC) < 0.6 OR

Validity

- Face validity is inadequate[7] OR is the only type of validity discussed and the measure is undergoing maintenance review OR
- Reviewer determines the methodology to assess validity is inadequate/inappropriate[8] OR the analytic approach is inadequate/inappropriate OR
- Reviewer disagrees with the assertion that the measure can distinguish quality with limited or no threats to validity present OR

Risk Adjustment

- Factors in the risk model do not influence the measured outcome OR are not present at the start of care OR the risk model includes factors that are associated with differences or inequities in care without sufficient rationale based on the conceptual model OR
- Analysis does not demonstrate:
    - Variation in prevalence of risk factors across measure entities AND
    - Contribution to unique variation in the outcome AND
    - Impact of risk adjustment for providers at high or low extremes of risk OR
    - Results do not demonstrate acceptable model performance.

Not Met but Addressable:

- Criterion is not met but the reviewer can identify:
    - Improvements to the sampling methodology OR
    - Changes to the methodology/analytic approach that could improve assessment of reliability OR
    - Changes to the methodology/analytic approach that could improve assessment of validity OR
    - Changes to the specifications that could improve validity and/or address threats to validity OR

[7] Face validity is accomplished through a systematic and transparent process, in which developers/stewards disclose identified relevant experts (e.g., clinicians, accountable entity representatives, those [patient, caregivers] with lived experience) 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.

[8] As part of the validity testing methodology, developers/stewards should empirically assess, as appropriate, the impact of missing data and/or measure exclusions.

- Changes to the risk model that could improve model appropriateness or performance.

Met:

Sampling

- If a sample is used, the sampling strategy is determined by the measure's analytic unit AND sample represents the variety of entities whose performance will be measured AND sample includes adequate numbers of units of measurement for the selected statistical method AND

For Patient or Encounter Level Reliability[5]

- Internal consistency $\geq$ 0.7 OR
- Inter-rater agreement $\geq$ 0.4 OR
- Test-retest reliability (ICC or Pearson correlation) $\geq$ 0.5 OR
- Linear relationship $\geq$ 0.6 AND

For Accountable Entity Level Reliability[5,6]

- Signal to noise/Inter-unit Reliability $\geq$0.6 OR
- Split-half reliability (ICC) $\geq$ 0.6 AND

Validity

- Face validity is adequate[7] and the measure is undergoing initial review OR
- Reviewer determines methodology employed[8] is adequate and the analytic approach presented is appropriate and thorough AND
- Reviewer determines results of empirical testing adequately demonstrate that the measure is valid AND
- Reviewer determines the interpretation of the empirical results supports an inference of validity AND

Risk Adjustment

- Factors in the risk model influence the measured outcome AND are present at the start of care AND the risk model does not include factors that are associated with differences or inequities in care unless justification provided based on the conceptual model AND
- Analysis demonstrates:
    - Variation in prevalence of risk factors across measured entities AND
    - Contribution to unique variation in the outcome

- Impact of risk adjustment for providers at high or low extremes of risk AND
- Results demonstrate acceptable model performance.

Equity*
Describe how this measure contributes to efforts to address inequities in health care. Provide a description of your methodology and approach to empirical testing of differences in performance scores across multiple sociocontextual variables (e.g., race, ethnicity, urbanicity/rurality, SES, gender, gender identity, sexual orientation, age). Provide an interpretation of the results, including interpretation of any identified differences and consideration of negative impact or unintended consequences on subgroups.

**The Equity domain is optional, as Battelle recognizes some measures are not designed to advance health equity. Battelle continues to explore this, but to align with national priorities, Battelle encourages developers and stewards to address this domain, if and when possible.*

Not Met:

- Reviewer determines equity is not sufficiently assessed OR the measure does not contribute to efforts to address inequities in health care.

Not Met but Addressable:

- Criterion is not met but reviewer can identify changes to the assessment of equity OR changes to the measure specifications that would address inequities in health care.

Met:

- Reviewer determines sufficient assessment of equity was conducted (i.e., methodology provided, differences in scores tested across multiple categories, and interpretation of results) AND the measure contributes to efforts to address inequities in health care.

Use and Usability
[For initial endorsement] Check all planned uses and provide the name of the program and sponsor, URL, purpose, geographic area and percentage of accountable entities and patients included, and level of analysis and care setting. ☐ Social Security Act modifications under the Patient Protection and Affordable Care Act and related accountability applications ☐ Quality Payment Program (QPP) Merit-based Incentive Payment System (MIPS) and Qualified Clinical Data Registries (QCDRs)

Use and Usability
☐ Specialty society clinical data registrations ☐ Certification programs ☐ Employer insurance plans ☐ Medicaid ☐ Other use:
[For maintenance review] Check all current uses. ☐ Social Security Act modifications under the Patient Protection and Affordable Care Act and related accountability applications ☐ QPP MIPS and QCDRs ☐ Specialty society clinical data registrations ☐ Certification programs ☐ Employer insurance plans ☐ Medicaid ☐ Other (specify):
What are the actions measured entities can take to improve performance on this measure? How difficult are those actions to achieve?
[For maintenance only] Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how you obtained feedback.
[For maintenance only] Describe how you considered the feedback when developing or revising the measure specifications or implementation, including whether you modified the measure and why or why not.
[For maintenance only] Discuss any progress on improvement (trends in performance results, including performance among sub-populations, if available, number and percentage of people receiving high-quality health care, geographic area, number and percentage of accountable entities and patients included). If use of the measure demonstrated no improvement, provide an explanation.

Not Met:

For initial endorsement

- There is no plan for use in at least one accountability application after initial endorsement but before the measure's first maintenance review OR
- Performance scores do not yield actionable information that can be used to improve performance among measured entities.

For maintenance

- The measure is not currently in use in at least one accountability application OR
- Performance scores do not yield actionable information that can be used to improve performance among measured entities OR
- Reviewer determines, based on the information provided regarding feedback on measure performance, the measure is not usable.

Not Met but Addressable:

For initial endorsement and maintenance

- Criterion is not met (see above), but the reviewer can identify changes to specifications that may strengthen the measure's ability to yield actionable information or usability.

Met:

For initial endorsement

- There is a plan for use in at least one accountability application after initial endorsement but before the measure's first maintenance review AND
- Performance scores yield actionable information that can be used to improve performance among measured entities.

For maintenance

- The measure is currently in use in at least one accountability application AND
- Performance scores yield actionable information that can be used to improve performance among measured entities.
- Reviewer determines, based on the information provided regarding feedback on measure performance, the measure is usable.