Comments on the Endorsement and Maintenance (E&M) Guidebook

The Outpatient Measures team at Acumen, LLC (hereafter referred to as Acumen) thanks the CBE entity for the opportunity to provide feedback on the Endorsement and Maintenance (E&M) Guidebook Version 3.0, published by the Partnership for Quality Measures (PQM) in June 2025. Our comment is structured into two sections: 1) a discussion of the improvements made to the E&M process covered in the latest Guidebook, and 2) a discussion of several remaining areas for further improvements we believe the CBE entity should consider addressing prior to finalization.

Comments

1. Improvements

Acumen appreciates several updates made to the Guidebook. Specifically, Version 3.0 of the Guidebook:

- Adds some specificity with respect to reliability rubrics and ways for developers to address validity concerns;
- Notes that space will be provided for developers to describe accountability program context under the Use and Usability evaluation rubric.

2. Areas for Further Improvement or Clarification

Despite the improvements noted above, Acumen finds that in other instances there are issues that it believes the CBE entity can provide further clarity, or consider further changes. Our comments encompass new Guidebook text, as well as text that has not been updated in this version.

2.1. Inadequately defined and justified reliability standards

We commend the Guidebook's effort to provide standardized benchmarks and methodologies for determining measure reliability. However, the current guidance does not adequately delineate recommended methodologies for calculating reliability or give sufficient justification for the methods it does propose. Several instances of such guidance include:

- The basis for having "not met" scientific acceptability standards is listed as having reliability scores "<0.6 for 70% or more of the accountable entities" (p. 49), but the converse criteria for having "met" reliability standards reads "[reliability] ≥0.6 for 70% or more of the accountable entities" (p. 51). We believe the intention in the "not met" case was to convey having reliability scores <0.6 for *30%* or more of accountable entities.
- While the PQM Measure Evaluation Rubric (Appendix D) lists failing reliability thresholds as grounds for having "not met" scientific acceptability standards, Table 4 suggests that a measure that failed the reliability thresholds could be endorsed with proposed conditions on a 3-year maintenance cycle. We would suggest more clarity on how the reliability standards would be applied.
- The Guidebook consistently lists standards for split-half reliability based on a certain percentage of accountable entities meeting the 0.6 threshold. However, split-half reliability is not usually calculated individually for each provider given that it estimates the theoretical correlation between multiple observations of the entire set of accountable entity scores. Thus, this criterion is incoherent and should be revised to give a threshold for a single ICC

value or to recommend a different reliability method (e.g., variance ratios). In the alternative, guidance should be provided as to which method(s) of calculating split-half reliability at the individual entity level are recommended.

Beyond these inconsistencies, we are concerned that there is insufficient justification given for the Guidebook's reliability standards. The updated guidance does not discuss any rationale for the specific thresholds of reliability listed (e.g., 0.6 for 70% of accountable entities), nor does it cite an external source of guidance or consensus for these exact thresholds. In addition, the blanket recommendation for non-binomial measures to be assessed using split-half reliability over other methods such as signal-to-noise ratios is not given an explicit rationale. For binomial measures, a forthcoming paper which does not seem to be publicly available (Aume et al, 2025) is cited with no indication of how it is being referenced. The Guidebook also continues to cite the National Quality Forum's (NQF) Scientific Methods Panel (SMP) June 2022 meeting as the source for establishing reliability thresholds (p. 48). However, this meeting did not set reliability thresholds and explicitly noted that the next steps would be to hold a public comment period, which ultimately did not occur.[1]

Finally, reliability standards must recognize a tradeoff with validity standards. It is straightforward to construct highly reliable measures that are clearly invalid (e.g., a measure that assigns every provider's episodes a numerical score based on the first letter of the provider's name and then takes the average across episodes). More realistically, certain statistical methods may be used to enhance reliability metrics by reducing validity (e.g., shrinkage, outlier removal, increasing sample sizes by expanding inclusion criteria). It is also straightforward to construct highly valid measures that have low reliability (e.g., defining a narrow clinical cohort for a measure denominator increases the validity of most measures, while generally reducing the reliability – this effect is familiar to statisticians, as it is akin to an estimator that is unbiased [i.e., valid] but imprecise [i.e., less reliable], sometimes described as bias-variance trade-off). The Guidebook should address how Committee members should evaluate such tradeoffs.

2.2. Lack of objective standards for validity

The E&M process lacks objective validity standards. The Guidebook does not define explicitly define validity, instead leaving its evaluation entirely up to the reviewers' discretion: the validity criterion is not met if the "[r]eviewer determines the methodology to assess validity is inadequate/inappropriate; OR the analytic approach is inadequate/inappropriate" (p. 46).

We recommend that the CBE entity develop objective validity standards that can be applied by reviewers. We propose that the definition of validity should be the same as the Measures Management System definition for validity: "In measure development, the term 'validity' has a specific application known as test validity, which refers to the degree to which evidence, clinical judgment, and theory support interpretations of a measure score. Stated more simply, test

[1] The reliability standards have evolved over time. There remain many unresolved aspects of SMP's years of considering reliability standards (e.g., how firmly to apply this line in recognition that any number is arbitrary, and whether additional validity testing would offset lower reliability results).

validity is an empirical demonstration of the ability of a measure to record or quantify what it purports to measure."

Furthermore, the examples provided (p. 66) with respect to correlation/causation are problematic. First, a specific correlation coefficient (0.5) is described as 'lower than expected' without any guidance explaining how such specific expectations should be established. It is not clear whether these expectations are determined based on established benchmarks/thresholds from the CBE entity, or if the researcher should pre-specify a rough expected correlation (i.e., high, medium, low, positive, negative) based on existing data, literature, or a conceptual model. Second, the suggested follow-up analysis to explore causal explanation for the correlation is more suited to researchers and/or practitioners who are able to exploit additional primary data. Measure developers that typically conduct these sort of correlation studies using generally available measures/data may face challenges beyond noting clear and well-known policy factors that could be responsible. The CBE entity should provide a wider range of options/examples for addressing such validity concerns.

2.3. Inconsistent Analysis requirements for Risk Adjustment

The Guidebook includes updates to risk adjustment expectations for new versus maintenance measures. Specifically, formal analyses for risk adjustment are only required during maintenance (p. 52). However, we believe this should be more clearly stated in the description of Scientific Acceptability related to risk adjustment on page 48. PQM should also provide a rationale for why risk adjustment analyses are not required during initial endorsement. Additionally, to ensure that measures are fairly evaluated, PQM should establish that measure submissions that choose to omit optional analyses are not negatively evaluated by committee members.

2.4. Inappropriately uses "acceptable model performance" as a Scientific Acceptability criterion

As currently written, this requirement creates the possibility for the reviewers to be subjective and inconsistent across measures. We strongly recommend removing requirements for the risk adjustment model to "demonstrate acceptable model performance" (p. 52), unless a clear definition and accompanying criteria for determining this standard can be established. Should such a definition be established, we further recommend an additional public comment period so measure developers/stewards can provide input on the definition.

Currently, PQM panelists often interpret "acceptable model performance" based on explained variance metrics alone (e.g., R-squared). Such metrics are not appropriate to differentiate a good measure from a bad measure, as they will improve when additional covariates are included in a risk adjustment model regardless of their conceptual validity. In our view, metrics such as predictive ratios (the ratio of observed to expected outcomes), calculated for various subsets of the measured populations (e.g., risk deciles, patients or providers with different characteristics, etc.) are much more informative about model performance and the scientific acceptability of a measure.

2.5. Unclear Criteria for "Closing Care Gaps"

We commend the CBE entity for its attention to how quality measures can play an important role in addressing efforts to reduce gaps in care across particular sets of identifiable patient

populations. As part of this effort, the criteria for maintenance states one of the requirements is the developer "*describe or provide* evidence indicating how accountable entities can utilize these results to close gaps in health care delivery and outcomes for the identified groups" (p. 43, emphasis added). However, the criteria for 'Met' only notes that there is evidence provided, rather than the more inclusive option to instead describe how providers could close care gaps using the results of the measure. We believe maintaining the option to either describe or provide evidence is an important consideration as this domain becomes required in future cycles, in part because certain processes or outcomes may be difficult to directly measure through evidenced impact, such as in instances where these sorts of characteristics are not available within the data sources used to construct the measure.

Furthermore, we would encourage the CBE entity to adopt language that broadens the set of characteristics that could be selected from for testing gaps in care, so that there are clear and feasible standards for this domain.

2.6. Ambiguous Default Endorsement Status for Measures Newly Converted to eCQMs

Please clarify if a new eCQM version of an existing endorsed [non-eCQM] measure is automatically considered to be endorsed. The opening sentence of the General Requirements states that it is, but the following text suggests it is not automatically considered to be endorsed. Our suggestion would be that a measure would keep its endorsement status in situations where the data elements are the same before and after the transition to an eCQM.

2.7. Feasibility and Burden Concerns Related to eCQM Testing

We recommend that the CBE entity take advantage of all available opportunities to reduce burden and cost for eCQM developers. Otherwise, as the guidebooks continue to become longer and more complex, the cost of developing eCQMs may quickly spiral and become out-of-reach for developers, thereby running counter to their original intent of reducing cost and burden within quality measurement.

The updated Guidebook states on pg. 63 that feasibility and person-or encounter-level testing must occur at a certain number of sites (three sites within at least one EHR vendor for initial endorsement and five sites within at least two EHR vendors for maintenance). Our team is concerned about the feasibility of this blanket requirement. Our prior experience indicates measures developed for use in more specialized and/or resource-burdened settings, such as Ambulatory Surgical Centers and Rural Emergency Hospitals, may be especially challenging to recruit testing partners given it can be both timely and costly for them to participate. To alleviate this concern, we would recommend the CBE entity develop additional guidance to permit cross-setting testing where test partner recruitment becomes impractical.

Furthermore, we would recommend that the requirement for multiple EHR vendors during maintenance should be eliminated. Many eCQMs also have a voluntary reporting period and a slow build-up of reporting over several years, meaning this requirement could disproportionally contributes to high development cost. Furthermore, any benefits in having multiple EHR vendors is likely decline over time as Fast Healthcare Interoperability Resources (FIHR) continues to emerge as a common interoperability standard.

Lastly, data elements testing should be limited to those that have not been used by another endorsed measure before to increase efficiency by reducing duplicative testing. Re-establishing the feasibility and availability of commonly used elements such as ED encounters brings marginal value but can incur significant costs to developers.