Thank you for the opportunity to review and comment upon the Partnership for Quality Measurement (PQM) first Endorsement and Maintenance (E&M) Guidebook. I submit these comments as an individual member of PQM with extensive consensus-based experience, as follows:

- I currently serve on the Scientific Methods panel (since 2019)
- Member, National Quality Forum (NQF) Variation in Measure Specifications Advisory Group (2016-2017)
- Co-Chair, Technical Expert Panel on Composite Performance Measure Evaluation, National Quality Forum (2012-2013)
- Member, National Quality Forum Task Force on Measure Testing (2010-2011)
- Member, National Quality Forum Task Force on Usability (2011-2012)
- Chair, Expert Advisory Panel, National Quality Forum Measure Specification Coding Maintenance Project (2009-2010)
- Member, Surgery and Anesthesia Technical Advisory Panel, National Voluntary Consensus Standards for Hospital Care: Additional Priorities, 2007, National Quality Forum (2007-2008)
- Member, Safe Practices Maintenance Committee, National Quality Forum (2008-2014, intermittent)
- Member, National Quality Forum Ad Hoc Advisory Committee on Evidence and Performance Measure Grading (2005-2006)
- Member, National Quality Forum Workshop on Child Healthcare Quality Measurement and Reporting (2004)
- Developer or co-developer of many consensus-based entity (CBE) endorsed measures of hospital quality and patient safety, including several electronic clinical quality measures (eCQMs) and claims-based measures, such as the Patient Safety Indicators (PSIs) originally developed by AHRQ and now stewarded by CMS in the form of PSI 90, the Patient Safety and Adverse Events Composite

Most importantly, I appreciate Battelle's efforts to streamline the E&M process so the decision-making process can be expedited while maintaining the transparency and multi-stakeholder participation that characterize the current process. The most notable and advantageous enhancements include:

- Retirement of the Consensus Standards Approval Committee, which currently adds little value to the E&M process;
- Tightening the calendar between the Intent to Submit and the full submission, limiting the amount of information required with the Intent to Submit to those elements necessary for proper assignment of the measure to a project, and recruitment of needed reviewers; and
- Reducing the number of E&M committees to ensure more equitable distribution of effort and to increase the number and diversity of voters on each committee (while increasing Battelle's efficiency in managing the process).

However, I do have several concerns regarding the proposed process, and whether it will actually achieve the goals of CBE review.

First, the Scientific Methods Panel (SMP) would be limited to enhancing "all measures by focusing on novel and the most difficult methodological challenges faced by measure developers." The Guidebook is otherwise silent on the composition, activities, and procedures of the SMP, demonstrating how it would

be marginalized under the new Guidebook. Specifically, the SMP's role would be entirely advisory, and would have no direct role in the E&M process. Although there is no need for the SMP to review every submitted measure, or even every "complex" measure, it should be engaged to address particularly important or novel methodologic questions on individual measures, or to help resolve questions regarding consistent treatment of similarly situated measures. For example, the project E&M committee chairs, in consultation with CBE staff, could refer measures to the SMP when additional methodologic input is needed for their scientific acceptability review. Alternatively, project committees could have an option of referring a measure to the SMP before making a final decision. Although such processes could potentially pull a measure off the 6-month endorsement track, deferring a final decision to the next cycle, this process would apply to a small minority of submitted measures, and the total duration of the process would not exceed the duration of the current process (for ALL measures). The SMP's involvement in individual measure review would be expected to decrease over time, as the measure evaluation criteria become more precise and better understood, and as project committees develop greater methodologic expertise, but eliminating any option for involvement in individual measure review seems imprudent at this time of transition and uncertainty.

Second, the proposed project structure would benefit from some clarification of the areas covered to improve balance across committees and to ensure that each committee is competent to review the measures assigned to it. To elucidate this problem, it would be helpful to enumerate the currently endorsed measures that would fall within the domain of each project. For example:

- Primary prevention, which is typically defined as efforts to prevent the development of disease, intervening before health effects occur, is primarily within the domain of public health and therefore motivates relatively few measures requiring or referred for CBE review. Relatively few CBE-endorsed measures focus on primary prevention, and one of the examples provided (i.e., cervical cancer screening) is clearly NOT primary prevention. I suggest broadening the concept to include elements of secondary prevention, such as screening for diseases (e.g., cervical, breast, and colorectal cancer; alcoholism as in CBE#2152).
- Initial recognition and management should cover more than signs and symptoms; it should cover the entire diagnostic process including diagnostic safety and diagnostic error, which have recently been recognized as critically important gaps in the current quality measurement enterprise. For example, laboratory testing and imaging are critical components of the diagnostic process; a committee in this domain should have strong representation from disciplines such as radiology, pathology, and laboratory medicine.
- The 3rd and 4th projects are not clearly delineated and would require hugely divergent expertise; for example, consumer assessment of hospice care is extremely different from management of pediatric hemodialysis. I suggest instead dividing projects according to the objective of the measure, which may also align with the well-accepted Institute of Medicine/National Academy of Medicine domains for high-performing care: for example, "patient or caregiver experience" versus "timely and effective care for acute and chronic conditions" versus "patient safety." My suggested approach would ensure that the first committee has experts in survey research and patient experience, while the second has experts in chronic disease management and process measurement, and the third has experts in patient safety.

Third, measures would undergo maintenance reviews every 3 years, although developers/stewards may request extension for up to 1 year, based on unclear criteria. I suggest instead adopting a consistent 5-

year timeline for maintenance review (subject to the provision for emergency/off-cycle review if needed), which would significantly reduce the burden on measure developers and the entire PQM. There is no benefit to triennial review, and the proposed approach seems problematic in the absence of clear criteria for proposing or accepting an extension.

Fourth, the removal of endorsement would require "75% or greater agreement for endorsement removal by the E&M committee," if the steward resubmits the measure with evidence of a meaningful gap. This standard is very problematic because it dramatically lowers the bar from the initial endorsement decision. In other words, a measure would require 75% support for endorsement, but only 25% (+1) support for maintenance. In other words, is the "default assumption" or "base case" that measures should be sunset after 5 years, and continued if they demonstrably continue to meet E&M criteria, or is it that measures should be continued and used forever? I would argue that the standard for maintenance should be the same as the standard for endorsement, and that raising the bar so dramatically will make it very difficult to sunset measures that the vast majority (up to 75%) of experts no longer support.

Fifth, it is important to allow flexibility for criteria such as "no significant change in measure results for accountable entities over time." Just because performance on a measure has not improved (yet) does not mean that it cannot improve, or that it will never improve. In some cases, there may be obstacles to improving performance that are difficult to overcome with currently available human, organizational, and financial resources, but continued attention to the measure will help to address these obstacles. Some problems in health care require continued attention and focus, even if progress has been difficult.

Sixth, the proposed "75% or greater agreement for endorsement" threshold appears to be a significant change from the current 60% threshold, but it is hard to evaluate how it will actually work in practice, based on Appendix F. More transparency and clarity regarding this proposed change is necessary. For example, the column headings in Appendix F are undefined and uninterpretable. The footnote to the table ("threshold for consensus is 0.95") appears to contradict the 75% threshold described elsewhere and is unrealistically high. It is not clear what the proportions within the table cells represent. The concept of "total available range of variance" is not defined or illustrated by example. It is unclear what estimator of variance will be used based on the Measure Evaluation Rubric in Appendix D; for example, will committee members be asked to rate measures on the 1-9 ordinal scale used in Davies et al. (2011), or will they simply be asked to classify measures as "not met," "not met but addressable," or "met"? The latter classification is categorical, not ordinal, because addressability is a complex judgment that is conditioned on available time and resources (and requires developer input).

Seventh, the role of measure developers and stewards in the E&M process must be clarified and strengthened. The current Guidebook appears to remove developers and stewards from the process, except insofar as they may be asked to address questions from E&M staff, as described on page 13. It is important that measure developers and/or stewards continue to be available to address questions raised by committee co-chairs and members (not just E&M staff), to respond to public comments, and to address criteria flagged as "not met but addressable" by committee members.

Eighth, the roster targets in Table 3 appear to grossly underestimate the importance of individuals with clinical expertise in the E&M process, especially given how this roster category includes all types of licensed health professionals. For example, review of process measures requires deep knowledge of how specific diseases should be treated, based on current professional guidelines and published evidence.

For example, cancer-related measures require input from medical oncologists, radiation oncologists, surgical oncologists, primary care providers involved in cancer care, oncology nurses, therapists involved in cancer treatment and recovery, radiologists or pathologists involved in diagnosis and follow-up surveillance, etc. Although strong representation of other stakeholders, as described, is essential, they cannot provide the critical review and interpretation of clinical evidence that is required for process-of-care measures, in particular.

Ninth, the Guidebook proposes a new criterion for appeal based on "evidence that the appellant's interests are directly and materially affected by the measure, and that the CBE's endorsement of the measure has had, or will have, an adverse effect on those interests." This criterion appears to be borrowed from the judicial sphere, is conceptually problematic, and will prove to be impossible to implement fairly. Specifically, this criterion will disqualify any appellant that is not a provider, purchaser, or payer of healthcare. The CBE is not an appellate court. Any stakeholder, including patients, patient/caregiver advocacy organizations, and researchers, should be able to appeal a CBE decision based on the existence of a procedural error, overlooked evidence, misapplication of the measure evaluation criteria, or other failure of the review process. The proposed criteria would preclude any researcher or former patient (for example) from appealing a decision. In addition, I would strongly recommend that the Appeals Committee include at least some independent voices and recuse the involved committee co-chairs from decision where they have a conflict of interest. Fair consideration of appeals requires limiting the participation of those who have a vested interest in rejecting the appeal, and including impartial reviewers.

Finally, some of the specific Measure Evaluation Criteria proposed in Appendix D do not appear to be well justified based on either measurement theory or published literature. For example, it is unclear why the proposed threshold is so low for inter-rater agreement (0.4) and somewhat higher for test-retest reliability (0.5). I am not aware of evidence that test-retest reliability is more important, or easier to achieve, than inter-rater agreement. Specifically, test-retest reliability estimates are very sensitive to the interval between test and retest, as the underlying phenomenon may change between test and retest. On the other hand, inter-rater agreement is usually tested at a single point in time, or based on the same source material (e.g., images, text, video of a patient encounter), so it isolates the impact of the assessor. The threshold for inter-rater agreement should be equal to or higher than the threshold for test-retest agreement, consistent with SMP discussion in 2022. It should also be clarified that the threshold for accountable entity-level reliability (0.6) refers to a measure of central tendency (median or mean), as these reliability estimates vary widely according to the volume/size of the entity. With respect to risk-adjustment, it is sometimes appropriate for risk-adjustment models to include features that do not significantly influence the measured outcome, if they are reasonably EXPECTED to influence that outcome, based on the conceptual framework and published literature. Similarly, it is often appropriate to include features that do not vary significantly in prevalence across measured entities, because they COULD vary in prevalence, given a larger and more diverse testing sample, and because they are clearly associated with the outcome of interest. In other words, the definition of confounding provides a reasonable foundation for identifying potential risk-adjustment features, but it is often appropriate to include features that do not meet the strict definition of a confounder.

Thank you for this opportunity to comment on the E&M Guidebook, and I look forward to following this process to its conclusion.

Patrick S. Romano, MD MPH

Professor of Internal Medicine and Pediatrics

University of California, Davis

Member, Scientific Methods Panel