

THE NATIONAL QUALITY FORUM

'Hospital Care Outcomes & Efficiency' Technical Advisory Panel Meeting February 4-5, 2009

A meeting of the 'Hospital Care Outcomes & Efficiency' Technical Advisory Panel (TAP) was held on February 4-5, 2009 in Washington, DC.

Technical Advisory Panel members present: Gabriel Escobar, MD (Co-Chair); Pat Stone, PhD, FAAN (Co-Chair); Larry Glance, MD; Lisa Iezzoni, MD, MSc; Nikolas Matthes, MD, MPH, MSc, PhD; Sean O'Brien, PhD; Martha Radford, MD; Denise Remus, PhD, RN; Amy Rosen, PhD

NQF Staff Present: Helen Burstin, Karen Pace, Eric Colchamiro

Measure Stewards Represented In-person or by Phone: CMS, Health Benchmarks, Premier, 3M, Society for Vascular Surgery, The Leapfrog Group

WELCOME, INTRODUCTIONS, AND DISCLOSURE OF INTERESTS

Dr. Escobar and Dr. Stone welcomed the TAP members who then introduced themselves and stated any conflicts of interest¹.

The purpose of the meeting was to:

- describe the hospital outcomes and efficiency project, NQF consensus development process, and the role of the TAP and Steering Committee;
- evaluate the candidate measures using the NQF standard evaluation criteria;
- make recommendations to the Steering Committee on which measures are suitable as voluntary consensus standards (based on the results of the evaluation); and
- begin to identify issues regarding evaluation of outcome measures that need clarification or guidelines.

INTRODUCTION

After the audience and those on the conference call line introduced themselves, NQF staff provided background information.

¹ Sean O'Brien – contractor for Society of Thoracic Surgeons database, but no measures in this project; Martha Radford – ACC/AHA Task Force on Performance Measures; Denise Remus – Previously worked at AHRQ on the quality indicators, previously worked at Premier, but before the collaboration with CareScience; Amy Rosen – Evaluating the AHRQ patient safety indicators, however no financial interest in the measures. During the course of the meeting a concern was raised about potential conflict of interest for two TAP members who were asked to review their prior association with the 3M measures. Dr. Iezzoni was the principal investigator about 9 years ago on a grant funded by AHCPR that developed the complications screening program (CSP), which are incorporated into the 3M PPCs. Later, she conducted a validation study in CA and CT using medical records; however, she has not been involved with CSPs since 1994 and was never involved with the 3M system/measures. Dr. Remus noted she was exposed to the 3M APR-DRGs in her work with the AHRQ quality indicators, but not the PPCs. The state of FL is using the 3M PPRs and she has seen that data. Her health system was a participant in the CMS Premier Hospital Quality Incentive Demonstration, but has not yet seen the measures from the CareScience methodology.

Strategic Issues for NQF

Dr. Burstin spoke of four strategic issues for NQF: driving toward high performance, shifting toward composite measures, moving toward outcomes measurement, and measuring disparities in all we do. The measurement framework currently out for vote, that's been developed over the last couple of years at NQF, has focused on the idea that we need to increasingly get to measures that examine shared accountability rather than trying to isolate single entity accountability. The measurement framework calls for measurement across patient-focused episodes of care including outcome measures, appropriateness measures, and cost/resource use measures coupled with quality measures. Dr. Burstin also noted the work of the National Priorities Partnership has led to the development of six core priority areas which guide NQF's work: patient and family engagement; population health; care coordination; safety; palliative and end-of-life care; and eliminating waste.

Intellectual Property

Dr. Burstin also reviewed the new NQF Intellectual Property policy that was adopted in 2008. Under this policy, proprietary measures may be considered for NQF endorsement. Throughout the process, transparency is of paramount importance. Detailed specifications must be made available to Steering Committees and the TAP during the Consensus Development Process (CDP). While charges are allowed for complex measures, there is still a very clear, strong preference on the part of the Board whenever possible, for measures that are free of charge. Although no measurement is truly free, additional charges for using the measures should be considered as one factor under feasibility. Additionally, if a measure is NQF-endorsed, measure stewards will provide a limited license to view the measure and all of the specifications by anyone who wishes to review the measure before entering into an arrangement to use the measure.

Hospital Outcomes Project and the Role of the TAP

Dr. Pace briefly reviewed the Hospital Outcomes project, timeline, and role of the TAP. The TAP, comprised of technical experts, serves in an advisory capacity to the Steering Committee. The TAP has been convened to primarily evaluate the measures against NQF's criteria for scientific acceptability of measure properties, but also was asked to evaluate the criteria for usability and feasibility. The TAP's evaluations will be provided to the Steering Committee, which is comprised of the full range of stakeholders and is responsible for the final evaluation and recommendations that will be presented for public comment. The timeline for this project requires the Board endorsement decision by the end of July. The TAP will be reconvened later this year to provide assistance to NQF in developing more specific guidance for evaluation of scientific acceptability of outcome measures.

Dr. Pace reviewed the agenda and work plan for the meeting. The TAP will first have a discussion regarding methodological issues that need to be considered when evaluating outcome measures, then move into a discussion of the candidate measures. Dr. Pace noted that TAP members have been assigned as primary and secondary reviewers for each measure and have completed an initial evaluation of the measures prior to the meeting. During this meeting, the assigned reviewers will present their evaluations and the full TAP will discuss each measure and vote on the evaluation ratings. The measure stewards/developers were invited to attend and provided time to make brief introductory remarks to the TAP, as well as respond to questions that arise during the TAP discussion. Prior to the formal review and evaluation of each individual measure, the TAP will meet in two groups – one for complications measures and one for the mortality and readmission measures. The task of the workgroups is to identify and discuss harmonization, advantages and

disadvantages of competing, measures, and cross-cutting evaluation issues. The audience is free to observe the work group discussions.

Measure Evaluation and Methodological Issues for Outcome Measures

Dr. Pace reviewed the NQF measure evaluation criteria and introduced three methodological issues (risk adjustment, rare occurrence, and time windows) that were included in the briefing memo.

Risk adjustment is intended to account for differences in intrinsic health risks that patients have at the start of their healthcare and aims to “level the playing field” for valid comparisons. The update to NQF’s measure evaluation criteria provides some guidance (e.g., factors should be evidence-based, present prior to the start of care, clinical factors rather than those associated with inequalities in care such as race). Dr. Pace noted that although the Steering Committee agreed with the goals of hierarchical modeling (e.g., address small sample size and clustering of patients within hospitals) it did not think it appropriate to limit consideration to only outcome measures that employed hierarchical models.

Another methodological issue with some outcome measures is the rare occurrence of numerator events (e.g., death, harm resulting from care). Although the denominator population size may be substantial, the low occurrence of the numerator event makes risk adjustment and determining statistically significant differences difficult. The Steering Committee for this project agreed that a rate for a low incidence event was not very useful. It also identified another method that should be considered for rare occurrences. Rather than reporting numerator occurrences and denominator patients, the measure could represent some element of time elapsed since last occurrence or events (e.g., number of falls per 1000 patient days, number of medication errors per 1000 doses, etc.).

Outcomes and risks need to be framed in the context of a specific time window. Time windows may be specified in terms of fixed periods (e.g. one year) or events (hospital stay). Fixed time periods reduce the confounding effect of different practice patterns (e.g., in-hospital mortality may be higher for hospitals with longer average hospital stays, but 30-day mortality may be the same). Another consideration with time windows is that the more time that elapses from the care provided, the more likely that other factors besides the specific processes of care will influence outcomes.

CROSS-CUTTING OUTCOME MEASURE EVALUATION ISSUES

The TAP discussed outcome measure evaluation issues and during the course of the work group discussions and individual measure evaluations some additional issues were identified. The TAP will be able to address these more fully when it is reconvened to assist with developing more specific guidance on evaluating outcome measures.

In response to question to explain what is meant by harmonization, Dr. Pace noted the definition in the evaluation criteria:

Measure harmonization refers to the standardization of specifications for similar measures on the same topic (e.g., *influenza immunization* of patients in hospitals or nursing homes), or related measures for the same target population (e.g., eye exam and HbA1c for *patients with diabetes*), or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are dictated by the evidence. The dimensions of harmonization can include numerator, denominator, exclusions, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.

A question was raised as to whether harmonization of two measures is necessary if one measure is not publicly reported.

The TAP discussed hierarchical modeling. Some TAP members identified that we need to find ways to explain hierarchical modeling to clinicians and consumers. The TAP agreed that hierarchical models are appropriate, but are not the only way to address clustered data and small volumes. Restricted cohort selection where hospitals are compared only for the same type of patients may be an alternative for some measures. The issue of clustering also can be addressed in regular regression with robust variance estimators. One TAP member noted that hierarchical modeling can be used with other types of measures besides outcomes, so need to discuss separately from risk adjustment. He suggested that hierarchical modeling is sometimes over-represented as the answer to the issue of small volume; however, it may give a false sense of making a precise estimate for small volume providers. Such an estimate is only close to true on average because it is based on so much information from the pool of providers. The rankings resulting from those estimates can be extremely unstable and do not necessarily result in better conclusions about whether a provider is better or worse than its peers. It may be better to present the information (risk-adjusted rate and sample size) and make it clear about the uncertainty about a provider's performance.

NQF's evaluation criteria suggest that race and socioeconomic status be used for stratification rather than risk adjustment, a TAP member questioned whether not adjusting for SES status in the hospital readmission measures may put hospitals that serve an impoverished population at a disadvantage. This issue was not discussed further but perhaps highlights the need for some analysis to justify the inclusion of risk factors.

Reliability (precision) of estimates is affected by sample size and has implications for the level of aggregation (e.g., clinician level), small providers, and the typical number of outcome events. If the outcome event is relatively rare and the sample volume is low, there might not be enough information to discriminate between high and low quality providers.

There was substantial agreement that it's important to provide the confidence intervals and sample size information for the scores on outcome measures, especially in the context of public reporting. TAP members noted some good examples – CMS' latest display on Hospital Compare and the VT/Oxford Neonatal Network. The TAP discussed that it is difficult to separate how measure results are reported, particularly the amount of uncertainty in the score, from the endorsement of the measure. Dr. Burstin noted that typically NQF endorses measures and does not have control over how implemented, however the TAP should provide any guidance on reporting that it thinks is important.

Some TAP members discussed the actionability of outcome measures for quality improvement, questioning how much providers control outcomes and noting that outcome results do not pinpoint what should be done to improve. Actionability by providers may be a desirable measure attribute, but the increasing focus on measures where accountability is shared needs to be considered. It was noted that the purpose of risk adjustment is to control for variability in patient factors that influence the outcome. Variability in risk-adjusted outcomes where some providers are performing better than others should indicate that lower performers can improve. Outcome measures indicate areas where performance could improve (at least up to peers), but providers will need to investigate what needs to be considered for improvement and that can vary from

hospital to hospital. Dr. Burstin discussed that outcome measures such as 30-day readmission rates can drive toward shared accountability resulting in much bigger improvement over the silo approach to measurement and accountability.

A related concern was the usability of outcome measures with composite endpoints such as the global complications and readmission measures under consideration. The overall rates subsume many different types of complications or readmissions, which could vary greatly from one hospital to another. Hospitals would need to have a way to discover its specific problems to initiate quality improvement activities and consumers may find specific information more useful for making decisions.

The TAP discussed the limitations of using administrative data. Reliance on administrative data for the outcome element (e.g., complications) presents more concerns than using those data for risk factors. You can forgive some inaccuracy in risk variables, but it's more problematic when there are inaccuracies in the outcome itself. The validity of the score results must be assessed – that variability in scores represents variability in the outcome vs. variability in coding practices. It was suggested that false negative rates from not coding a particular complication also needs to be assessed. The TAP discussed that variability in data sets in terms of the number of secondary diagnoses included also affects comparability because it is likely that more complications will be identified when more diagnosis codes are analyzed. The TAP discussed that the present-on-admission (POA) code will help to sort out complications from comorbidities, however it is new in many states and reliability is still questionable. For future measure maintenance, complications measures should be updated and analyzed using the POA indicator. It is especially important during measure maintenance to evaluate the reliability of outcome variables and important risk adjustment characteristics.

Some noted that measures of complications previously were considered screens or indicators rather than confirmed performance because of the limitations of using codes from claims data. Dr. Burstin stated that measures endorsed by NQF are considered suitable for public reporting, not just quality improvement. A TAP member commented that endorsed measures may quickly be adopted for pay-for-performance initiatives. Dr. Burstin noted that continuing to rely on intense medical record review for quality measures is untenable. We are in a transition period before clinical data are available in and can be extracted from electronic medical records for quality measurement. NQF currently has projects on clinically enriched claims data and identifying a quality data set for electronic medical records. A TAP member commented that clinical registry data also is transitional because it is not currently extracted from electronic medical records.

The TAP made the following suggestions to consider for its future task on assisting NQF develop guidance for outcome measure evaluation.

- Set expectations for submitters by providing more specific guidance and examples.
- Provide a checklist of what risk model performance metrics should be reported. There was discussion, but not complete agreement, around minimum submission requirements including standard statistical measurements (e.g., R-squared, C statistic/ROC curve, calibration, overfitting, sensitivity to missing data, development and validation sample). The TAP discussed that the model performance metrics need to be understood in the context of the data set used. If the data sets are not comparable, it's problematic to compare these metrics. It was suggested and agreed that NQF should try to provide a reference database that could be used to run the proposed measure and risk models for claims-based measures. If that's not possible,

the submitter should put the performance in perspective with comparison to other models for the outcomes cited in the literature.

- Ask submitters to provide a context or framework for the submitted measure, the purpose, and how the measure was developed.
- Provide justification for the risk variables included/excluded from the model – conceptually as well as data on variability of risk factors across providers and relative contribution of the factors. Sensitivity analyses for bad/missing model data should be considered.
- Clarify the different types of reliability: reliability of the data elements used in constructing the measure and reliability (precision) of the score produced by the measure, which pertains to true vs. random variation. Be more specific of expectations for reliability and validity testing.
- Ask for information on the outcomes – the rates expected and whether they are large enough to be able to detect significant differences.
- NQF should consider pre-screening measure submission by statisticians.
- Measure and model testing should use the most current data available.
- Ask for information on improvement in a proposed outcome.

COMMENTS

NQF members and public audience members were given the opportunity to make comments. The following points were addressed.

- Variability in the number of diagnosis codes and positions and presence of POA in a data set affects the results of a measure, so the measure specifications need to address that.
- Some ICD-9 codes are more meaningful than others and that also needs to be evaluated. For example, only about a tenth of the actual occurrence of DVT/PE is reported in the administrative data.
- Comments on the PCI 30-day mortality measures HOE-009/HOE-010 included a concern regarding the probabilistic matching, that non-procedure related mortality is included, and that all hospitals performing PCI should be included when the measure is implemented. A letter from the Society for Cardiovascular Angiography and Interventions was distributed to the TAP.
- There is difference in the global complications measures discussed and the AHRQ indicators that focus on specific complications with an assessment of the codes that are used, so it's possible to use ICD-9 codes for complications measures if adequately tested.
- Related to global readmissions measures, there are not a lot of evidence-based practices to reduce readmission beyond the CHF patient population.

The measure stewards/developers were invited to present introductory remarks on their measures to the TAP. The following representatives did so, and also were available during the TAP's discussion of the individual measures to respond to questions that arose.

- Dr. Bruce Hall, from the Washington University and Barnes Hospital in St. Louis; HOE-015, the LEB bypass mortality/complications measure developed in conjunction with CMS
- Dr. Jephtha Curtis, from the Yale-New Haven Health System; HOE-009 and HOE-010 PCI 30-day mortality measures developed in conjunction with CMS
- Judy Chen, from Health Benchmarks; HOE-004, the Risk Adjusted 30-day Readmission Rate for Heart Failure
- Barbara Rudolph, Director of Leaps and Measures for the Leapfrog Group; HOE-013, Survival Predictor

- Gene Kroch, Chief Scientist with Premier, spoke on measures HOE-006 and HOE-018, complications and morbidity indexes
- Dr. Anton Sidawy, President-elect of the Society for Vascular Surgery (SVS); HOE-017, Postoperative Death or Stroke in Patients Undergoing Carotid Endarterectomy

EVALUATION OF INDIVIDUAL MEASURES

The evaluation of each measure began with an assessment by the primary and secondary reviewers, followed by discussion and voting by the entire TAP. Questions that arose were referred to the measure stewards/developers. The following table provides the TAP ratings and recommendations and a summary of the major points related to its evaluation. Four measures were recommended by the TAP to advance in the consensus process; however, all measures will be evaluated by the Steering Committee.

THE NATIONAL QUALITY FORUM

Hospital Outcomes & Efficiency Technical Advisory Panel – February 4-5, 2009 Summary of Review of Measures

NQF Evaluation Criteria: I=Importance to measure and report; S=Scientific acceptability of measure properties; U=Usability; F=Feasibility
Importance to measure and report: this is a threshold criterion and the Committee votes: Y=yes, N=no, or A=abstain. Measures that do not pass the importance criterion are not further evaluated and not recommended for consensus standards.

Remaining Criteria: Extent to which the NQF evaluation criteria are met: H=high; M=moderate; L=low. The Committee votes or reaches consensus on ratings.

Recommendation: The Committee/TAP votes on the overall recommendation for endorsement: Y=yes, N=no, or A=abstain.

Cross-Cutting Issues

The committee identified several cross-cutting issues that were considered: accurately identifying complications from ICD-9 codes in administrative data and testing to identify that scores reflect variation in complications vs. variation in coding, evaluation of variability in false negative rates across hospitals, the effect of present-on-admission (POA) coding, and the effect of the number of diagnosis codes that are included in a dataset (e.g., MEDPAR includes 10 dx codes, CA includes 25 dx codes).

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
HOE-011-08 Measure of the Occurrence of deep-vein thrombosis/pulmonary embolism (DVT/PE) Following Hip or Knee Replacement Surgery (Johnson & Johnson Health Care Systems, Inc.)	<p>Measure Evaluation criteria: I: Yes (SC) S: H-0;M-0;L-9;-A- U: cannot determine-9 F: cannot determine-9</p> <p>Recommend for Time-Limited Endorsement: Y-0;N-9;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: I: DVT is an important topic of measurement and relates to NPP goal. Information was provided on impact, but not variability in performance.</p> <p>S: Measure is untested. The measure is intended to identify treatment for DVT/PE 30 days after discharge; however the specifications do not provide any detail for ambulatory coding and linking index hospitalization to post-discharge hospital and ambulatory claims. No risk adjustment strategy is planned because the steward states that DVT/PE is considered "preventable for all patient risk profiles"; however, in item #19 of the submission form, the steward identified factors associated with disparate outcomes including cancer, obesity, age, previous VTE, oral contraception, which indicates the need for some method of risk adjustment or risk stratification.</p> <p>F: Only feasible if able to link claims across time and settings.</p>
HOE-015-08 Postoperative Respiratory Failure (PSI #11) (Agency for Healthcare Research and	<p>Measure Evaluation criteria: I: Yes (SC) S: H-3;M-6;L-;-A- U: H-9;M-;-L-;-A- F: H-9;M-;-L-;-A-</p> <p>Recommend for Endorsement w/Condition: Y-7;N-2;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: I: Although there was some question of the source of estimates for variability 2.3-29.2% and whether wide confidence intervals would negate much variability, because this measure is being used, the committee thought it warranted further evaluation.</p>

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
Quality)	<p>S: Criterion validity is high, suggesting that the measure is identifying a high number of true positives or true events. The risk-adjustment methodology appears sound, has been used in numerous indicators and settings, and takes into account clustering within hospitals. The indicator is used specifically to examine the quality of care within a specific hospitalization, so that measurement is relatively precise. The measure has been used in several settings with comparable results and high positive predictive validity. Someone questioned whether the false negative rates had been evaluated; however others pointed out that has not been a requirement for testing and this measure has had other appropriate reliability and validity testing.</p> <p>F: Use of administrative makes feasible.</p> <p>The TAP recommended this measure on the condition that the results of current validation testing are reported as soon as possible. A suggestion also was made that at the time of maintenance review, an assessment of the use the POA indicator be included.</p>
HOE-017-08 Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Endarterectomy (Society for Vascular Surgery)	<p>Measure Evaluation criteria: I: Y-9;N-0;A- S: H-;M-5;L-4;A- U: H-0;M-2;L-7;A- F: H-0;M-0;L-9;A-</p> <p>Recommend for Time-Limited Endorsement: Y-0;N-9;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: I: This measure submission had been inadvertently missed. The TAP agreed it met the importance criterion. These are important outcomes and measure also would encourage selection of appropriate patients for the procedure.</p> <p>S: The measure is untested. The measure would require physician claims be linked to hospital claims in order to have the information in the G-code that indicates the patient was asymptomatic for a year prior to the procedure. Although the measure would not need risk adjustment if restricted to the asymptomatic patients, testing of the reliability and validity of the G-code, especially for under-reporting is necessary. The TAP also did not think that a cumulative lifetime rate for individual physicians was a sound approach for performance measurement and that other approaches to deal with small volume should be explores (e.g., rolling time periods).</p> <p>F: G-code not yet established and G-codes not used in hospital claims.</p> <p>These issues do not warrant granting time-limited endorsement – the measure should be tested and then brought back to NQF.</p>
HOE-018-08 Inpatient Co-morbidity Adjusted Complication Index (Premier, Inc) HOE-006-08 Inpatient Co-morbidity Adjusted Morbidity Index (Premier, Inc)	<p>HOE-018-08</p> <p>Measure Evaluation criteria: I: Y-8;N-0;A-1 S: H-0;M-8;L-1;A- U: H-0;M-0;L-9;A- F: H-1;M-5;L-3;A-</p> <p>Recommend for Endorsement: Y-0;N-9;A-</p> <p>HOE-006</p> <p>Measure Evaluation criteria: I: Yes (SC) S: H-0;M-3;L-6;A- U: H-0;M-0;L-9;A- F: H-1;M-5;L-3;A-</p> <p>Recommend for Endorsement: Y-0;N-9;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: Measures HOE-018 and HOE-006 are both measures of complications using the same methodology. HOE-018 includes all complications; HOE-006 includes severe complications. Please note that the measure steward was notified that for this project, specific measures need to be proposed and evaluated. Although the Premier classification system facilitates drilling down into the data for various levels of analyses, the measures being evaluated for potential endorsement are for total complications or total severe complications across all</p>

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
	<p>hospitalized patients.</p> <p>I: Measure HOE-018 submission had been inadvertently missed and not previously reviewed by the Steering Committee. The TAP agreed it met the importance criterion.</p> <p>HOE-006 was reviewed previously by the Steering Committee. Relates to NPP goals and is a relevant outcome for patients with variability in performance.</p> <p>Issues for further evaluation include: usability due to broad focus; adequacy of risk adjustment; ability to identify co-morbid conditions; lack of specificity for target population – can be applied to any population, but NQF only endorses discrete, specific measures.</p> <p>S: A primary concern was the replicability of the classification system. The definition of what constitutes a complication is dependent upon an evaluation of principal-secondary diagnosis pairs selected by volume and reviewed by physician panels using modified Delphi consensus techniques to determine the probability that the secondary dx is a complication rather than a comorbidity. Complications also are classified by severity on a 5-point Likert scale (A-E) by internal panels of clinicians and those rated D&E are used to denote the severe complications for HOE-006. The risk adjustment model includes race and income variables, which the NQF evaluation criteria suggest should not be used in risk adjustment. The developer stated these are considered proxies for access to care. The risk model also includes valid procedure (not sure what that includes) and discharge status (which occurs after care is provided). Risk model performance metrics for a development and validation sample were not provided. The number of diagnoses included in a data source (e.g., CA-25, MEDPAR-9) can affect rates of complications unless hospitals only compared within the same data source. Only face validity is addressed, and there was no testing to determine if variability in scores reflects variability in complication rates or in coding practices. The TAP discussed that use of administrative ICD-9 codes to identify an outcome such as all complications vs. variables used in risk models, necessitates an understanding of the reliability of those data.</p> <p>U: The TAP discussed whether a global complications measure based on diagnosis codes can be used for public reporting because of the issues identified above, although using such methods as screening tools for quality improvement activities might be helpful. In addition, the overall scores subsume many different complications so that the type of complications could differ greatly from one hospital to another. There also was some discussion of whether a global complications measure can be used for quality improvement without data on the specific complications. However, risk-adjusted complication rates from coded data could identify situations that require further investigation. The classification system used to compute the measures also can be used in the QI investigation to identify patients or various groups of patients (e.g., by diagnosis) with the complications. The developer stated that hospitals would need to do that analysis on their own and it could be done in a simple spreadsheet.</p> <p>F: The measure is based entirely on administrative data. The measure steward plans to make the measure freely available. The TAP agreed that such a measure is useful for screening and the system a useful tool for QI investigations, but it is not ready for publicly benchmarking performance.</p>
<p>HOE-007-08 3M™ Potentially Preventable</p>	<p>Measure Evaluation criteria: I: Y-9;N-0 S: H-0;M-4;L-5;A- U: H-0;M-6;L-3;A- F: H-0;M-2;L-7;A-</p> <p>Recommend for Endorsement: Y-0;N-9;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: Please note that the measure steward was notified that for this project,</p>

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
<p>Complications (PPCs) (3M Health Information Systems)</p>	<p>specific measures need to be proposed and evaluated. Although the 3M classification system facilitates drilling down into the data for various levels of analyses, the measure being evaluated for potential endorsement is for total complications across all hospitalized patients.</p> <p>I: This measure was not previously reviewed by the SC. The TAP agreed it met the importance criterion.</p> <p>S: This measure builds on the AHRQ Patient Safety Indicators (PSI) and the Complications Screening Program (CSP). The measure will be sensitive to present-on-admission (POA) coding practices, and the developers point out that hospitals have 2 incentives to increase POAs: 1) to decrease complication rate and 2) increase severity of illness. It was developed using CA data where POA has been implemented. The number of diagnoses included in a data source (e.g., CA-25, MEDPAR-9) can affect rates of complications unless hospitals are only compared within the same data source. A TAP member noted that CA data tends to be quite different and that validation with other data sets from other states or a national set would be desirable. Only face validity is addressed, and there was no testing to determine if variability in scores reflects variability in complication rates or in coding practices. Risk adjustment is accomplished by indirect standardization using APR DRGs further subdivided by 4 severity of illness subclasses and 4 risk of mortality subclasses (developed by and iterative process of formulating clinical hypotheses and then testing the hypotheses with historical data). The TAP discussed that use of administrative ICD-9 codes to identify an outcome such as all complications vs. variables used in risk models, necessitates an understanding of the reliability of those data. It agreed that POA coding will assist with distinguishing complications from co-morbidity; however POA coding is relatively recent in many states and measure scores are subject to variability in coding practices.</p> <p>U: The TAP discussed whether a global complications measure based on diagnosis codes can be used for public reporting because of the issues identified above, although using such methods as screening tools for quality improvement activities might be helpful. In addition, the overall scores subsume many different complications so that the type of complications could differ greatly from one hospital to another. There also was some discussion of whether a global complications measure can be used for quality improvement without data on the specific complications. However, risk-adjusted complication rates from coded data could identify situations that require further investigation. The classification system used to compute the measures also can be used in the QI investigation to identify patients or various groups of patients (e.g., by diagnosis) with the complications.</p> <p>F: The measure is based entirely on administrative data. The measure steward intends to charge for use of the measure, which would require both the PPC system and APR DRGs (stated PPC is roughly half the cost of APR DRG). The TAP agreed that such a measure is useful for screening and the system useful tool for QI investigations, but it is not ready for publicly benchmarking performance.</p>
<p>HOE-008-08 Hospital specific risk-adjusted measure of mortality or one or more major</p>	<p>Measure Evaluation criteria: I: Yes (SC) S: H-4;M-5;L-;A- U: H-0;M-6;L-3;A- F: H-0;M-4;L-5;A-</p> <p>Recommend for Time-Limited Endorsement: Y-8;N-1;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: I: All sub-criteria were met. One committee member questioned whether it was high enough volume to be considered high impact or best for internal QI only. Another committee member thought it is also an indicator of appropriate pre-operative patient selection. In response to a question regarding variability Bruce Hall, NSQIP stated that 17% experience an event and variability of risk-adjusted predicted:expected ratio</p>

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
<p>complications within 30 days of a lower extremity bypass (LEB). (Centers for Medicare and Medicaid Services)</p>	<p>is 0.75 to 1.25.</p> <p>S: Submission indicates not fully developed and tested and will be completed within 24 months, however development testing reported is quite extensive. Data fields are well defined, but the developer indicated reliability testing would be completed prior to implementation. The TAP questioned whether reliability would hold up when implemented outside of NSQIP's training and auditing and also noted that the risk model would need to be recalibrated. The measure steward noted that NSQIP currently captures about 90% of cases, so would expect relatively few changes. It was clarified that although the measure was developed using NSQIP database, participation in NSQIP is not a requirement for implementation. The measure has a multiple endpoints because of the low occurrence of each event individually, but is not submitted as a composite measure. The reliability of the functional status risk variable was questioned, as well as the validity of RVU as a risk factor. Others commented that accuracy of risk factors overall are less a problem than accuracy of the outcome data. Creatinine>1.2 is a risk factor - should also consider code for dialysis. Developed using 3 years of data, but anticipate computing yearly rates when implemented. The presentation of interval estimates is a strength of the proposed methodology.</p> <p>U: In response to a question whether rates could be improved, the developer stated that they have seen improvement in NSQIP.</p> <p>F: Uses clinical data that until electronic records are available must be collected and reported (now to NSQIP registry, possibly some other mechanism). Participation in NSQIP is not a requirement. Feasibility cannot be entirely evaluated because a national data collection strategy has not yet been proposed.</p> <p>The TAP recommended time-limited endorsement due to development using registry data vs. implementation intended nationally, no report about reliability testing, and need to recalibrate the risk model when implemented nationally.</p> <p>Developer Response regarding RVU: Years ago in the NSQIP the program attempted to control for "procedural complexity" by creating an in-house scale of complexity developed by a panel of experts, but it became apparent that this same information was largely captured already by the CMS designation of work RVUs, with the added advantage that this was an independent body doing the assessments, and that the assessments were updated periodically. It was demonstrated within NSQIP then that the correlation between work RVUs and the in-house "complexity score" was high, and so the complexity score was dropped and the work RVUs were adopted. Again, the aim was to provide some control for procedural complexity, within or across procedure types. Thus, there is now many years of experience using work RVU as a risk adjuster within the NSQIP, and that experience was carried forward into this project. In this vascular project, wRVU continued to demonstrate explanatory value as a risk adjuster (as reflected in the submitted materials). Keep in mind, however, that this LEB measure only deals with a well-defined subset of vascular procedures; controlling for complexity of procedures is always less important within a small procedure subset than it would be for comparisons across large sets of disparate procedures. Nonetheless, the inclusion of wRVU in this vascular measure did contribute to explanatory power.</p>
<p>HOE-009-08 30-day all-cause risk-standardized</p>	<p>The ratings, recommendations, and rationale apply to both measures.</p> <p>Measure Evaluation criteria: I: Yes (SC) S: H-9;M-;L-;A- U: H-6;M-3;L-;A- F: H-3;M-6;L-;A-</p> <p>Recommend for Time-Limited Endorsement: Y-9;N-0;A-</p>

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
<p>percutaneous coronary intervention (PCI) mortality rate for patients without ST segment elevation myocardial infarction (STEMI) and without cardiogenic shock (Centers for Medicare and Medicaid Services)</p> <p>HOE-010-08 30-day all-cause risk-standardized Percutaneous Coronary Intervention (PCI) mortality rate for patients with ST segment elevation myocardial infarction (STEMI) or cardiogenic shock (Centers for Medicare and Medicaid Services)</p>	<p>Rationale for ratings (I, S, U, F)/recommendation: Measures HOE-009 and HOE-010 are basically the same except for the denominator populations (with or without STEMI/cardiogenic shock), which are clearly distinct, both from a clinical standpoint as well as from a data collection standpoint. The measures were discussed and voted on together.</p> <p>I: All sub-criteria met.</p> <p>S: Submission indicates not fully developed and tested and will be completed within 24 months, however development testing was reported. Data fields are well defined, but the developer indicated reliability testing would be completed prior to implementation. It was clarified that the measure was developed using NCDR CathPCI registry database, but participation in the registry is not a requirement for implementation. The measure submitted requires matching registry data to Medicare claims and enrollment data. The developer indicated that availability of patient identifier would improve the measure through ability to link with actual outcome (rather than probabilistic matching to outcomes). The TAP agreed that probabilistic matching to endpoint would not be acceptable for a publicly reported measure. The TAP also agreed that 30-day mortality was preferable to in-hospital mortality and that the use of clinical data as in the registry is preferred to administrative data alone. The definition of cardiogenic shock needs reliability testing and may need refinement.</p> <p>U: The endpoint for this measure is easily comprehensible to both the general public as well as to clinicians; it is useful for both consumers and providers.</p> <p>F: Measure is based on an existing registry in which the majority of hospitals that perform PCI already participate and for whom feasibility is high. Those not already participating will need to allocate staff for data collection. Although all data elements are in the electronic registry, they are not currently extracted from an electronic medical record. Feasibility cannot be entirely evaluated because a national data collection strategy has not yet been proposed. The TAP recommended time-limited endorsement due to development using registry data vs. implementation intended nationally, probabilistic matching for testing vs. unique identifiers for implementation, need to recalibrate the risk model when implemented nationally, and no report about reliability testing for key cohort identification and risk adjustment variables.</p> <p>Developer Response regarding additional testing: In our NQF applications for the PCI mortality models, we indicated we would conduct additional testing within 24 months (Page 1, Question D) because CMS plans to refit the models once a national dataset with patient identifiers is assembled for public reporting (this was probably a conservative interpretation of the question). As you know, because we were not able to use direct patient identifiers during the process of measure development, we used a probabilistic match to merge CathPCI registry data with administrative data available on the subset of Medicare patients. The characteristics of patients who matched are virtually identical to those of patients excluded from measure development because they were not matched (Table 5 of the technical report). Accordingly, we are confident that the patients in our analysis are representative of the larger cohort of Medicare patients. In the course of measure implementation, we will have to refit the models using direct identifiers in all PCI patients. However, we would consider these steps part of measure maintenance as opposed to measure development. We have reviewed the criteria for “adequate field testing” set forth in NQF’s guidance on time-limited endorsement and believe that the PCI measures meet these criteria. The measures were developed in a large, representative cohort of PCI patients. Specifically, we analyzed data from more than 125,000 patients undergoing PCI at more than 600 hospitals that submit data to the American College</p>

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
<p>HOE-013-08 Survival Predictor (6 individual mortality measures - CABG, AVR, PCI, AAA, Esophagectomy, Pancreatectomy) (Leapfrog Group)</p>	<p>of Cardiology's CathPCI Registry.</p> <p>Measure Evaluation criteria: I: Yes (SC) S: H-0;M-3;L-6;A- U: H-1;M-5;L-3;A- F: H-0;M-0;L-8;CA-1;A-</p> <p>Recommend for Endorsement: Y-0;N-8;A-1</p> <p>Rationale for ratings (I, S, U, F)/recommendation: Although only one measure submission form was submitted that referred to a survival predictor and listed 12 component measures, the measure steward clarified that there are 6 separate mortality measures (CABG, AVR, PCI, AAA, Esophagectomy, Pancreatectomy).</p> <p>I: All components are NQF-endorsed so already determined to be important.</p> <p>S: Although the TAP agreed that the Bayesian methodology and modeling is elegant and cutting edge, it agreed it was not ready for endorsement. The proposed composite measures are a weighted average of a facility's mortality and the expected mortality given its volume ($E[m]=a+b*\log[\text{volume}]$). Facilities with a small number of cases then get weighted more heavily towards the expected mortality given their volume. This expected mortality is likely to be higher than the mean across all facilities since lower volumes are generally associated with higher mortality. The TAP noted the controversy surrounding the volume-outcome relationship and questioned the premise of using a volume-predicted mortality rate as a component of the composite. Although the methodology employed by this measure was recently published in a prestigious journal (Medical Care), the panel noted that publication of a single article often marks the beginning, not the end, of a discussion of a controversial subject. The panel expects that this paper will trigger much discussion as well as the publication of counter-examples and critiques, and that this process will take some time before consensus is reached on the volume-outcome relationship. Another issue identified was the lack of standardization regarding risk adjustment - the specifications allow for either risk-adjusted or raw mortality rates. The developer stated risk adjustment makes no difference in predicting future risk-adjusted rate. The competing NQF-endorsed mortality measures are all risk-adjusted.</p> <p>U: Because there are already NQF-endorsed mortality measures for the six procedures, the question is whether these represent additive value or superior methodology. The measure steward noted that the current NQF-endorsed cardiovascular measures from STS and ACC/AHA are not currently publicly reported.</p> <p>The TAP did not think these measures were ready to replace the existing endorsed measures.</p>
<p>HOE-004-08 Risk-Adjusted 30-Day Readmission Rate For Heart Failure (Health Benchmarks, Inc)</p>	<p>Measure Evaluation criteria: I: Yes (SC) S: H-0;M-8;L1-;A- U: H-0;M-7;L-2;A- F: H-0;M-8;L-1;A-</p> <p>Recommend for Endorsement: Y-0;N-9;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: I: SC already agreed on importance of readmission in Phase I.</p> <p>S: A number of issues were identified. It appears the risk models are fit to each plan rather than one that applies to all plans. Some exclusions (hospice) and risk factors (discharge to nursing home) occur after discharge and may inappropriately exclude or adjust for outcomes that are the result of care. There is conflicting information on how age is used (dichotomous, categorical). The comorbidity index includes some of the other individual risk factors (e.g., COPD, renal failure).</p> <p>U: Although this measure would apply to potentially all patients vs. the competing endorsed measure that applies only to Medicare patients, it would be limited to health plans because of the need to link claims over time.</p> <p>The TAP agreed this measure was not strong scientifically.</p>

Meas# / Title/ (Owner)	Steering Committee Discussion/Evaluation
<p>HOE-012-08 3M™ Potentially Preventable Readmissions (PPRs) (3M Health Information Systems)</p>	<p>Measure Evaluation criteria: I: Yes (SC) S: H-0;M-2;L-7;A- U: H-0;M-0;L-9;A- F: H-0;M-1;L-8;A-</p> <p>Recommend for Time-Limited Endorsement: Y-0;N-9;A-</p> <p>Rationale for ratings (I, S, U, F)/recommendation: Please note that the measure steward was notified that for this project, specific measures need to be proposed and evaluated. Although the 3M classification system facilitates drilling down into the data for various levels of analyses, the measure being evaluated is for total preventable readmissions across all hospitalized patients.</p> <p>I: SC already agreed on importance of readmission in Phase I.</p> <p>S: Although there is some appeal to isolating preventable readmissions, the TAP questioned the reproducibility and validity of the designation of preventable readmissions by clinical panels (the developer indicated “each of the 98,596 cells contain a specification of whether the combination of the base APR DRG for the Initial Admission and for the readmission were clinically-related and therefore potentially preventable”). A question also was raised about the stability of the empiric estimates for the PPRs based on one state (FL). The submission form indicates any risk adjustment method could be used but APR DRGs is recommended; however, one method would need to be specified to result in a standard measure. A limitation of the risk adjustment method is reliance on ICD-9 codes without POA indicator and whether can adequately distinguish what was present at the start of care from conditions that developed during care.</p> <p>U: There is an NQF-endorsed risk-adjusted measure for all readmissions. Comparison of results and rankings from this candidate measure of preventable readmissions with the endorsed risk-adjusted all readmission measure is needed to justify the complexity of this measure.</p> <p>F: The measure is based entirely on administrative data. The measure steward intends to charge for use of the measure, which would require both the PPR system and APR DRGs (stated PPR is roughly half the cost of APR DRG).</p>