NATIONAL QUALITY FORUM

RESOURCE USE STEERING COMMITTEE CONFERENCE CALL

December 5, 2011 (12-2pm ET)

Committee Members Participating: Bruce Steinwald, MBA (Co-Chair), Independent Consultant; Tom Rosenthal, MD, UCLA School of Medicine (Co-Chair); William Golden, MD, Arkansas Medicaid; Ann Hendrich, RN, MSN, Ascension Health; Thomas Lee, MD, Partners Healthcare System, Inc.; Jack Needleman, PhD, FAAN, University of California, Los Angeles School of Public Health; Mary Kay O'Neill, MD, MBA, CIGNA HealthCare; David Redfearn, PhD, WellPoint; Jeffrey Rich, MD, Mid-Atlantic Cardiothoracic Surgeons Ltd.; William Rich, MD, Northern Virginia Ophthalmology Associates; Barbara Rudolph, PhD, MSSW, The Leapfrog Group; Joseph Stephansky, PhD, Michigan Health and Hospital Association; Dolores Yanagihara, MPH, Integrated Healthcare Association

NQF Staff Participating: Helen Burstin, MD, MPH, Senior Vice President, Senior Director; Taroon Amin, MPH, MA, Senior Director; Ashlie Wilbon, RN, MPH, Senior Project Manager; Sarah Fanta, Project Manager; Lauralei Dorian, Project Manager; Evan M. Williamson, MPH, MS, Project Analyst.

Others present via phone: Gail Amundson, Caterpillar & Quality Quest; Cheryl Demars, The Alliance; Ben Hamlin, NCQA; Theresa Helle, The Boeing Company; Tom Lynn Ingenix; Dena Mendehlsohn, Pacific Business Group on Health; Amy Moyer, The Alliance; Kim Ritten, Health Partners; Marsha Smith, CMS; Cheri Zelinski, Ingenix; Tom Zumtobel, Culinary Health Fund.

MEETING PROCESS

The purpose of this conference call is for the Committee to:

1. Discuss comments received during the public and member comment period.
2. Provide input on responses to comments.
3. Provide recommendations on changes to the Cycle 2 draft report prior to the member voting period.
4. Determine whether reconsideration of any measures is warranted.

PROJECT UPDATES

Ms. Wilbon welcomed the Committee to the call and provided an overview of the call and a summary of the activities to date.

The CSAC's final discussion of Cycle 1 measures will occur during their upcoming call on December 12, 2011.

The Cycle 2 draft report closed for public and member comment on November 21, 2011; four measures were recommended for endorsement in this report:

- (1609) ETG-based hip/knee replacement cost-of-care measure (Ingenix)
- (1611) ETG-based pneumonia cost-of-care measure (Ingenix)

- (1560) Relative Resource Use for People with Asthma (NCQA)
- (1561) Relative Resource Use for People with COPD (NCQA)

Additionally, one measure had no consensus:

- (1595) ETG-based diabetes cost-of-care measure (Ingenix)

Comments from David Hopkins, PhD, Pacific Business Group on Health (PBGH)

Dr. Hopkins addressed the Committee on several concerns on behalf of PBGH and the Consumer/Purchaser Disclosure Project (CPDP). He commended the Committee on its work and efforts on this very important, yet complex, topic and acknowledged that NQF is working quickly to meet the need for cost measures but feels that industry is far ahead in the use of these measures. In order to not impede the progress of these measures in practice, Dr. Hopkins feels that recommendations in the report need to be clarified in some areas, many of which are reflected in the submitted comments and themes identified in the memo:

- Remove language indicating a preference for standardized pricing over actual pricing;
- Inclusion of certain statistical information is too restrictive, this should be left up to the user of the measure
- Measurement is inherently complex – understandability should not be used as a criteria for evaluation

Discussion of Comment Themes

NQF received 87 comments on the draft report from public and NQF members. All comments were subject to discussion; however, due to the volume of comments, staff aggregated comments by theme to facilitate the Committee's discussion. The nine major themes of the comments are listed below. In response to these themes, responses were drafted for the Committee to discuss.

Theme 1- Application of Costing Approaches

Description. Comments submitted expressed strong views on the usefulness of cost measures of actual prices paid for comparison of prices in markets nationally. While standardized pricing allows for comparison of resource use holding costs constant, pairing these measurement approaches to understand costs and resource use provides valuable information on the margin between prices paid and resource use.

> *Proposed Committee Response*: Standardized pricing allows users to compare the use and intensity of health services while holding actual paid amounts constant. Resource use measures that apply standardized prices allow for comparison of resource use units across regions and markets, while actual prices allow for comparison of prices paid within regions and markets. The Committee agreed that both approaches could be appropriate for different applications. However, the Committee's decision to recommend (or not recommend) individual measures should not be interpreted as driven by simply the measure's costing approach. A measure-by-measure decision was made on the appropriateness of the costing approach given other measure characteristics. Reliability

and validity was examined through the interaction of the measure's specified level of measurement, risk adjustment model, and other measure characteristics. There was agreement that actual prices paid by health plans to individual clinicians is important to measure and report; for example, regional comparisons at the individual clinician level where environmental factors may not be as prominent, or nationally at higher levels of measurement (i.e. health plan level). The Committee did, however, express concern over applying an actual price approach for national comparisons at an individual clinician level. Specifically, the Committee noted the potential for misinterpreting clinician resource use in national reporting. This pricing approach includes environmental factors (i.e., local facility and wage index) that may be outside of an individual clinician's control. The Committee agreed that when actual prices paid are reported, utilization counts should be reported as well.

Committee Discussion: The Committee agreed that both costing approaches were appropriate in different applications and that a measure's costing approach was not a disqualifying factor for endorsement. Furthermore, the Committee referenced the fact that measures with actual costing had been recommended.

Theme 2- Splitting costing approaches into separate measures

Description. Comments submitted questioned the need to separate costing approaches into separate measures, arguing the need for both approaches to be included in one measure.

Proposed Committee Response: The Committee agreed early in the evaluation process that a single measure should allow for only one costing approach (actual prices paid or standardized pricing) to ensure consistent and accurate comparisons of measure results. For use as a national consensus standard, measure results should unambiguously reflect differences in performance for an accountable entity, not differences in the type of data an entity choses to submit (actual prices or standardized prices). As such, developers that allowed for user flexibility in the costing approach were asked to split their measures into two separate measures where only one approach is specified in a single measure. Endorsing measures with a single costing approach, does not preclude the use of both measures as a pair. Developers also had the option to select a single costing approach to be applied to the measure. Health Partners elected to split their measure, while Ingenix selected actual price paid.

Committee Discussion: The Committee reiterated the general consensus of the group, with small exception, that individual measures should only specify one costing approach, noting that while measures are used in the commercial sector with user flexibility in the costing approaches, as a national consensus standard, combining approaches in to one measure run the risk of ambiguity when reporting and comparing results. A single costing approach per measure ensures valid comparisons in that measure results using standardized prices are not compared to measure results using actual prices and the meaning of the results are not misinterpreted.

Theme 3- Higher bar for resource use measure evaluation

Description: Commenters expressed concern that the report appeared to describe an evaluation standard that was higher than that used for quality measures, arguing that the evaluation of resource use measures should be held to the same standard as quality measures.

Proposed Committee Response: The resource use measure evaluation criteria are the same criteria used for quality measures; specifically, importance to measure and report, scientific acceptability of the measure properties, usability and feasibility. In order to customize the evaluation to specific components in resource use measures, the Steering Committee, in its first phase of work, sought to identify how resource use measure should be specified, and how to evaluate reliability and validity in these types of measures. The result of this effort is the NQF resource use measure evaluation criteria and the resource use specification modules.

The Committee identified five "modules" to describe the way resource use measures should be specified including data protocol, clinical logic, construction logic, adjustments for comparability, and reporting. The modules sought to provide developers with a familiar framework in which resource use measures are often constructed. The submission process was mirrored after the modules and vetted by most developers who submitted measures to the project (including Ingenix and NCQA).

While the measure evaluation sub-criteria were adapted for resource use, including importance and usability subcriteria, the remaining criteria remained unchanged from the criteria that are applied to quality measures. When evaluating the measures, the Committee applied the same criteria to all submitted measures in the same manner while taking into consideration some of the unique constructs of resource use measures and the nature of the interactive components of the specifications.

Both quality and resource use measures must demonstrate adequate reliability and validity testing at the lowest specified level of analysis. The Committee's determination of adequate testing and results relied on expert judgment of the Technical Advisory Panels and members of the Steering Committee to consider: (1) if the developers test was appropriate for the specified measure; (2) if the scope of testing including the representiveness and sample size was adequate for the specified level of analysis; and (3) if the results indicate an acceptable level of reliability (and validity). This standard is consistent across both types of measures.

Committee Discussion: The committee made the point of distinguishing between quality measures and resource use measures and the fact that the same evaluation criteria for quality measures were adapted for use with resource use measures.

Theme 4- Measures in use should be endorsed

Description: Commenters argue that measures that are already widely in use should meet the field testing requirements and this should be taken into consideration when making

recommendations for endorsement. Because the measure is in use it is inherently usable and feasible.

Proposed Committee Response: The Committee acknowledged that resource use measures have been in use in the commercial/private sector for many years, but have not been subject to the peer review and scrutiny that most quality measures have. In addition to the various complex methods and approaches for measuring the same types of costs/resources, there is limited published peer reviewed literature about the reliability and validity of these measures. This effort marks the first time that many of these measures have been subject to a systematic review of the methodology and scientific acceptability. As such, the wide use of these measurement approaches does not inherently imply the measures are acceptable for endorsement. The Committee also acknowledges the sensitive nature of some of the measures used in markets where financial investments have been made on behalf of purchasers and other users to integrate the measures into their systems for reporting and understanding costs/resource use. The context and process by which measures become endorsed as NQF standards requires that the measures meet each of the four criteria and qualify for use for public accountability and performance improvement purposes. While the current use of the measures is taken into consideration (within the usability criteria) by the Committee during evaluation, it does not imply the measure meets the criteria for endorsement.

Committee Discussion: Current use of the measure was considered in the usability criterion during the evaluation process. While some measurement approaches were developed using internal peer review and are in use by various organizations, this project was the first time that many of these measures have be subject to a public multistakeholder peer review. During this review, it was determined that statistical and clinical logic validity had not been adequately demonstrated. Even though several of these measures are currently in use, it does not imply that they used the most valid or appropriate approach. The Committee reiterated the general consensus, with small exception, that this is a new area of measurement and that a measure "in-use" is not necessarily "mature" and ready for NQF endorsement.

Theme 5- Complexity of the Resource Use measures from an episode grouper

Description: Commenters expressed concern that measures submitted by Ingenix were not endorsed due to their complexity. They argue that resource use and cost measures that use an episode grouper are inherently complex. Alternatively, Commenters also feel that due to the complexity of these measures they should be examined before the typical three year review cycle. This shorter cycle for updating these measures will help to solicit feedback from the field on the implementation process of these measures.

Proposed Committee Response: The Committee recognizes that resource use measures, including those derived from episode groupers are inherently complex. This complexity should not, however, hinder the transparency, clarity, and ability to deconstruct the measure for understanding. Further, the Committee chose to recommend measures based on individual measure characteristics, rather than disregarding any measure due to its

inherent complexity. The Committee agreed that resource use measures should be held to the same standard as quality measures, and evaluated against the same criteria; specifically, importance to measure and report, scientific acceptability of the measure properties, usability and feasibility. NQF will strongly consider a shorter cycle for updating these measures considering the concerns raised.

Committee Discussion: The Committee noted that complexity was not a disqualification for measure endorsement and that the ETG risk adjuster is very complex and still passed endorsement in several measures.

Theme 6- Cost of the measures submitted by Ingenix

Description: One commenter felt very strongly that the Committee should acknowledge the widespread use of Ingenix measures even in light of their costs. While another commenter expressed concern over the cost of the Ingenix measures, include cost of ETGs, ERGs, PEGs and the cost of implementation.

Proposed Committee Response: The Committee considered the cost of the Ingenix product (ETGs, ERGs, PEGs) in the feasibility criterion of the measure evaluation. While some users may find the cost of the episode grouper reasonable, the use of these measurements does not inherently imply the measures are acceptable for endorsement. The issue of the cost of the measures submitted by Ingenix was weighted differently for various stakeholders represented in the Steering Committee. The Committee also weighed the potential burden these costs may carry if these measures were adopted for regional or national reporting programs requiring that organizations take on these costs to participate.

Committee Discussion: The cost of measure use was considered under the feasibility criterion as indicated by the policy on endorsement of proprietary performance measures. This policy is not unique to resource use measures and is applied in the evaluation of proprietary quality measures with fees as well. The Committee agreed that while the issue of cost was taken into consideration, it was not a deciding factor in the recommendations for any of the measures.

Theme 7-Risk adjustment model

Description: Commenters disagreed that factors in the risk adjustment model and severity model should be confirmed to be a contributor to the outcome of the measure. One commenter was very concerned that the Committee was too focused on the scientific validity and the variables used in risk adjustment methods were actually correlated with outcomes (as well as clinically significant).

Proposed Committee Response: The Committee looked to Guidance provided by measure evaluation criteria and the NQF Measure Testing and Evaluation Scientific Acceptability of Measure Properties. For resource use measures and quality measures, an evidence-based risk adjustment strategy (e.g. risk models, risk stratifications) should be based on

patient clinical factors that influence the measured outcome (page 24). When evaluating the validity testing of the measure, the Committee sought to ensure that the data and sample used for development and validation are reflective of its intended measured population. The Committee agreed that measure developers have a responsibility to demonstrate quantitatively, the relative contribution of risk factors, risk model performance metrics and the an assessment of adequacy in the context of norms for risk models. The Committee argued that these testing requests are similar and aligned with quality measures.

Committee Discussion: The Committee recalled that the risk adjustment approaches from each developer was evaluated independently for each measure using the same criteria within scientific acceptability to determine its reliability and validity.

Theme 8-Preference for specification compared to guidelines

Description: Commenters felt that the Steering Committee favored specifications over guidelines. The concerns specifically referenced Emerging Principle 1 favoring specifications for the resource use measure construct.

Proposed Committee Response: The Committee did not express preference for specifications or guidelines. The submission process required that the measure clinical logic, construction logic, and adjustments for comparability details be submitted as specifications, however, all submission items within the data protocol and reporting modules allowed for flexibility. The Committee intentionally designed the measure submission with this flexibility in these modules of the measure.

Committee Discussion: The Committee reiterated their position on this topic. The reporting module allowed for flexibility and the Committee expressed no preference for specifications or guidelines.

Theme 9- Burden of validity testing

Description: Commenters expressed concern that the validity testing requirements are overly prescriptive and should not require a chart review as a necessary validity check. Chart reviews are expensive and are also susceptible to deficiencies that limit the accuracy of data extraction.

Proposed Committee Response: The Committee agreed that adequate validity testing is required for resource use measures in addition to quality measures, relying on guidance from the NQF Measure Testing and Evaluation Scientific Acceptability of Measure Properties. Validity testing can be done at the data element or the measure score level. If the developers choose to demonstrate data element validity, patient-level information on individual patients (e.g., count of medication provided) should demonstrate that the data elements are correct and the correctly identify differences in resource use (page 14; page 31). However, data element validity does not need to be conducted for every single data element. Testing can include only those critical data elements. Developers also have the option of measure score validity testing where developers can demonstrate

correlation of measure score results with another valid indicator of resource use. Developers have the responsibility to demonstrate the data elements and/or measure score are reliable and valid in their testing. Emerging principle 7 should not be interpreted as chart reviews are a necessary validity check, but rather, when demonstrating validity data elements they should be evaluated against an authoritative source (e.g. a similar measure that has been validated, or a validated tool).

Committee Discussion: The Committee confirmed that validity testing can be completed at the data element level or the measure score level. They further stated that during the measure evaluation, distinguishing between the two testing approaches (score or data element level) was not a major discussion for any of the measures.

Measure Specific Comments on Recommended Measures

(1560) Relative Resource Use for People with Asthma (NCQA)
(1561) Relative Resource Use for People with COPD (NCQA)
Description: Comments received for the two NCQA measures were similar. Commenters disagreed with the Committee's request for sample size requirements of 400 for NCQA measures. They argue that sample size requirements are overly restrictive and measure developers should have enough sample size to demonstrate reliability of 0.7. Moreover, commenters were concerned about this measure's use of administrative data as they are notoriously inaccurate, implementation of the measure may be overly burdensome, and problems with the use of diagnostic codes to distinguish between asthma and COPD in older persons. Commenters encourage the developers to use historical data to confirm and distinguish between COPD and asthma.

Proposed Committee Response: The Committee evaluated these measures based on a minimum sample size submitted as guidelines by the developer; it was not required. Specifically, the developer noted that measure testing demonstrated reliability with a minimum sample size of 400. The Committee, nor NQF, requires a minimum sample size for resource use measure endorsement; the submission process allows developers to submit this information as specifications, guidelines or not at all. The Committee agreed that measure developers need to demonstrate adequate testing and results and considered: (1) if the developers test was appropriate for the specified measure; (2) if the scope of testing including the representiveness and sample size was adequate; and (3) if the results indicate an acceptable level of reliability (and validity). The Committee, nor NQF, is prescriptive of the type of testing approach or any cut-off for reliability testing scores.

Further, the Committee recognizes that the use of administrative claims data presents certain limitations for measuring resource use performance; these limitations are present in quality performance measurement as well. While administrative data are the primary data source used for measuring resources at this time, the Committee encourages developers to integrate the data gathered through EHRs and other clinical data to measure resource use.

Committee Discussion: The identification of this sample size was submitted as guidelines and as part of the testing analysis by the developer and was not required by NQF or the Steering Committee.

(1609) ETG based hip/knee replacement cost of care measure (Ingenix)

Description: Some commenters expressed support of this measure, noting the measure's ability to capture actual costs at the individual clinician level. Another commenter questioned the measure's clinical logic since this hip fracture measure is based on a non-representative population and the developer submission lacks information on why low-cost outliers are excluded, but high cost outliers were windsorized. Further, the measure fails to capture important and costly complications of comorbidity such as post-op delirium, pulmonary embolus or dementia.

Proposed Committee Response: Concerns related to the clinical logic related to this measure were considered in TAP and Steering Committee discussions; however the Committee will reconsider these concerns on its upcoming December 5, 2011 conference call.

Committee Discussion: The Committee acknowledged the commenters' concerns, but determined that the recommendation for this measure should remain.

(1611) ETG based pneumonia cost of care (Ingenix)

Description: Commenters expressed concern over the validity of the clinical logic, specifically identifying the measure population using administrative claims data with limited ability to distinguish between different types of pneumonia. The inability to distinguish between community-acquired and healthcare-acquired pneumonia will result in the inclusion of costs for episodes of very distinct types of pneumonia into this measure. Further, Commenters also believed that there was insufficient information provided to the TAP to determine scientific acceptability. Other commenters disagreed that inclusion of costs six months prior to the pneumonia episode is an inappropriate approach to assigning costs.

Proposed Committee Response: The Committee considered the TAP discussion and concern of the inability to distinguish between different types of pneumonia. However, ultimately they agreed that this measure should be recommended noting the current limitations of administrative data, limitations that would apply to quality measures as well. The Committee will consider concerns on inclusion of six months of costs prior to the pneumonia episodes in its upcoming December 5, 2011 conference call.

Committee Discussion: The Committee acknowledged the commenters' concerns, but determined that the recommendation for this measure should remain.

Measure Specific Comments on the Split Vote Measure

(1595) ETG based diabetes cost of care measure (Ingenix)

Description: Commenters were generally supportive of this measure. One commenter encouraged the Committee and developers to further understand and describe the risk adjustment/stratification approach to ensure that comparisons are reasonable and accurate.

> *Proposed Committee Response*: Because the Committee's initial vote on this measure resulted in a split vote, the Committee will reconsider these concerns its upcoming December 5, 2011 conference call and determine the final disposition of this measure.
>
> *Committee Discussion:* It was agreed that re-voting or reconsidering the measure would likely not result in a substantial difference in Committee stance on the measure. As such, the Committee determined that the split vote should remain and be forwarded to CSAC as is.

Comments on Measures Not Recommended

(1591) ETG-based congestive heart failure (CHF) cost of care measure (Ingenix)
(1594) ETG-based coronary artery disease (CAD) cost of care measure (Ingenix)
(1599) ETG-based non-condition specific cost of care measure (Ingenix)
(1603) ETG-based hip fracture cost of care measure (Ingenix)
(1605) ETG-based asthma cost of care measure (Ingenix)
(1608) ETG-based chronic obstructive pulmonary disease cost of care measure (Ingenix)

Description: Commenters expressed concern over the Committee's decision not to recommend these measures. Commenters believe that all of these measures meet the NQF criteria and should be recommended for endorsement. They also suggest the Committee's rationale for not recommending endorsement for these measures was insufficient.

Proposed Committee Response: The Committee considered each measure submitted to this project individually. The Committee encourages identifying specific supportive or clarifying information related to the clinical logic and construction logic concerns raised. All measures recommended for use as a national consensus standard must meet the same four criteria as quality measures; specifically, importance to measure and report, scientific acceptability of the measure properties, usability and feasibility. Further, the Committee agreed that all measures must meet current standards for reliability and validity testing outlined by the NQF Measure Testing and Evaluation Scientific Acceptability of Measure Properties report.

> *Committee Discussion*: The Committee acknowledged the commenters' concerns, but determined that the recommendations for these measures should remain.

FURTHER COMMITTEE DISCUSSION

Following discussion of the themes and individual measures, several general comments were made by members of the committee:

- The Committee supports any efforts to shorten the review cycle for these measures due to the rapid advancement in the field, and suggested language to this affect be included in the report. NQF agreed that resource use measures will be evaluated again sooner than the 3-year maintenance cycle.
- The concern over Committee members' voting rationale was raised; specifically the concern there was a lack of detail in the report around why certain measures failed upon re-voting after initially passing. The committee agreed that having this information for developers would be desirable, but did not think it would be feasible to have each Committee member provide specific justification for each vote. This is not currently the process for Steering Committees.

PUBLIC COMMENT

There were no public comments.

NEXT STEPS

Staff will update the Committee responses to the comments and the draft report based on feedback from today's call and distribute for Committee review. Other important dates coming up for the project include:

- December 12, 3-5 pm ET: CSAC discussion of cycle 1 recommended measures