

From: Garcia, Edward (CMS/OCSQ) [mailto:Edward.Garcia@CMS.hhs.gov]

Sent: Tuesday, July 13, 2010 3:21 PM

To: Performance Measures

Cc: Pratt, Mary J. (CMS/OCSQ); Rapp, Michael T. (CMS/OCSQ); Frederick, Pamela R. (CMS/OCSQ); Halim, Shaheen (CMS/OCSQ); Millar, Fatima S. (CMS/OCSQ)

Subject: NQF Measure Testing Report - CMS COMMENTS

To Whom It May Concern:

Please accept the following comments from the Centers for Medicare and Medicaid Services (CMS), Office of Clinical Standards and Quality (OCSQ), Quality Measurement and Health Assessment Group (QMHAG) regarding the NQF draft document, "Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties."

General Comments:

- We suggest that the intent of the measure and the interest of the "greater good" should reign supreme, and that the proposed criteria should be applied in the context of this.
- The intended use of this document should be clarified – is this guidance to help educate steering committee members only or will it be used strictly to determine endorsement status of measures being evaluated? The latter has tremendous implications for measure developers.
 - The stated intent of the proposed guidelines is to provide backbone for the systematic and standardized evaluation of candidate quality measures as well as guidance for the development of candidate measures. Although we agree that scientific rigor provides the foundation for all measure development and implementation, the guidelines should be just that – guidelines. The mention of "criteria," although well-intended, implies that thresholds to demarcate success from failure will be applied universally without regard for the intrinsic merit of the candidate measure.
The term "criteria" will only be used to refer to NQF criteria; otherwise "guidance" will be used
 - There are many aspects to this proposed policy that will add time and effort for the testing process. The document outlines a much more rigorous and detailed empirical scientific process than previously required/interpreted. Measurement is required to be based upon an empirical research design, which is required to be documented throughout the NQF report form/document. This is beyond what we have currently required in our testing. Measures without empirical testing of reliability and validity should be considered "untested measures" and subject to NQF's conditions for considering untested measure" for endorsement according to this new policy. The development and re-endorsement process must now demonstrate a rigorous scientific process that focuses on both reliability and validity based on scientific research design and principles. This change in focus will require a formal research process, including increased hours to meet the rigorous empirical research process and related documentation requirements on the MET, MIF and NQF forms and other reports designated. NQF still does not tell us how many cases should be tested, but they have now added language about testing each provider TYPE in a measure, which we have not

done. For example, we did not specifically test 75 Social Workers, 75 Psychologists, but rather we did a random sample to get 75 cases.

- CMS will need to make some significant changes to our testing and also revisit how our measures are written, particularly around the exclusion criteria. We have let the TEP dictate exclusions, but now there needs to be hard scientific evidence to support any exclusion.

NQF criteria for reliability and validity have not changed - this document provides more explanation of reliability and validity testing which have been employed in measurement. Measures used for public reporting do require scientific rigor; however, these recommendation do not require the rigor expected in research studies. For example, testing may be conducted on a relatively small sample compared to the number of entities expected with implementation; and reliability testing is not required for each time the measure is implemented. Additionally, the guidance provides options for meeting the moderate rating, which would be needed to pass the criterion of scientific acceptability of measure properties.

- A workflow diagram for tables 2, 3, 4 would be useful to help the reader followed the decision making process for these topic areas. Some comments directed at the guidance for the evaluating of composite measures may be useful.
- Consider adding examples relevant for administrative claims data in Tables A1 and A2.
- Consider including an Appendix that contains example submission forms by measure data source, such as administrative data claims, EHR, medical records). The examples could be accompanied by a brief narrative describing the application of the evaluation criteria and justification of the rating. This would be very informative to the measure developers.

This document is not intended to provide a step-by-step guide to measure development and submission; however, NQF will explore the development of additional resources.

- CMS finds that the guidance document does not adequately address claims-based measures
 - For example Table A-1 (pg. 29) on “Common Approaches to Testing Reliability Applied to Quality Measures” seems to be focused on chart abstracted measures, so does the next Table A-2 on “Common Approaches to Testing Validity.”
 - There is a mention of claims on page 10, line 280 - 281. But in general the document does not deal with what will be satisfactory and sufficient evidence for NQF concerning reliability and validity for claims data.
 - Technically, one could argue that claims data are abstracted data as well. Some coder has to fill out the claim.

The document applies to all measures regardless of data type. The distinguishing feature of claims data is the use of codes to represent the clinical information in the record - claims data was more explicitly identified in the tables in the Appendix.

- For home health care in general and OASIS-C in particular, it may be difficult to meet more than moderate reliability and validity standards because the research to date for OASIS-C specifically and home health care in general is so sparse. Thus, unlike other areas (hospitals, SNFs primary care), there may be substantial challenges. Similarly, we do not have data on "meaningful differences in performance" (line 787) for home health care so it would be difficult to say whether an acute care hospitalization rate of 40% is meaningfully different than one of 36% for example.

The original home health measures were among the best tested measures submitted to NQF. These are questions that testing should answer.

- This document does not address structural measures. If this type of measure will be considered for endorsement, testing requirements and discussion of how the measure can meet reliability and validity thresholds should be addressed.

Structural measures do not have different requirements for reliability and validity testing

- Regarding Table A-1:
 - The first part of the table; Reliability Testing – Data elements is very straightforward and easy to understand.
 - The second part; Reliability Testing – measure score needs further clarification. Examples would be very helpful here. Examples of the types of testing one should conduct for claims based process and outcome measures would be very helpful.
Reliability testing of measure scores does not vary by type of measure or data type. Examples are provided in the Appendix.
 - We believe that it is reasonable to assume that claims are a valid and reliable data source without requiring each submitter to provide such evidence. Therefore, testing at the data element level is not necessary for claims-based measures because of “prior evidence” (lines 280-281). We believe that measure level testing should be required for claims based measures. We believe this should be explicitly stated to avoid differing interpretations by various measure developers, steering committees and NQF staff.
 - Table A-1 suggests inter-rater reliability testing to establish reliability of “Codes that are used to represent the primary clinical data (ICD, CPT, CPT-II/G).” ICD and CPT -category I codes are primarily used for payment purposes and have a long history of use and verification. We do not believe that measure developers should be responsible for independently evaluating these codes. Codes such as CPT-II and G-codes, on the other hand, have been developed for the purpose of reporting quality information, therefore, should be included in measure testing.

The guidance allows for using prior evidence of data element reliability/validity rather than new testing and also testing of either the data elements or measure score; however reliability and validity are not assumed based on any data type.

- Regarding Table A-2:
 - Validity Testing – Data element: Table A-2 suggests comparison to a “gold standard” to establish validity of “Codes that are used to represent the primary clinical data (ICD, CPT, CPT-II/G).” ICD and CPT-category I codes are primarily used for payment purposes and have a long history of use and verification. We do not believe that measure developers should be responsible for independently evaluating these codes. Codes such as CPT-II and G-codes, on the other hand, have been developed for the purpose of reporting quality information, therefore, should be included in measure testing. However, once the individual codes have been tested and have accepted as reliable and valid by NQF, if these same codes are used different measures submitted for endorsement in the future, testing for these elements should be “deemed”.
Reliability and validity testing applies to all measures regardless of data type; however reliability and validity are not assumed based on any data type
 - Validity Testing – Measure score. It would be helpful to give examples of types of measures for which systematic testing of face validity would be sufficient, and for which types of measures, NQF is expecting testing of the various aspects of construct validity: predictive, concurrent, convergent and discriminative.
Face validity is allowed for any measure if it meets the specific requirements identified in the guidance.

- We find that the risk-adjustment requirement for outcome measures is unproven and should therefore not be a requirement. The evidence is insufficient to specify which measures require risk adjustment and what elements should be included in models. Attention must be given to the product of risk adjustment to yield clinically meaningful and interpretable results. Perhaps one way to resolve this is to clearly state the purpose of the document, and also state what the document is NOT intended for. Risk adjustment or stratification for outcome and resource use measures is current accepted practice and is in the current evaluation criteria. The criteria allow for submitters to make a case for why risk adjustment/stratification is not needed with rationale and hopefully analysis. Risk factors should be evidence based and risk models empirically developed and tested.

Recommendation I

- Although we agree and applaud the principle of this recommendation, we are concerned that gold standards are not available for many candidate measures, and are limited to medical record abstraction for others. It must be acknowledged that medical records serve as the gold standard (as would be the case for claims, ICD diagnostic categories, and outcomes), this will be resource intensive and therefore of limited feasibility. This is also problematic for electronic record-based measures where gold standards need to be identified.
- Regarding bullet one – “evidence for reliability and validity may be accumulated over time and evaluators should remain flexible with regard to the extent of evidence submitted...”
 - CMS thinks this is ok for some types of measures (survey) but may not be acceptable for other types of measures eg. outcomes or intermediate outcome measures. The empirical evidence should be sufficient and strong enough to support P4P decisions for HHS/CMS purposes.
NQF endorses measures intended for public reporting as well as quality improvement. p4p is not a separate category of measure use.
 - Please define “relatively small scale” in regarding to the scope of testing at initial endorsement.
- Regarding bullet two- “Reliability and validity testing may be conducted on a sample of the measured entities. The analytic unit of the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing...”
 - CMS is confused by the use of the word “may”. CMS suggests this be changed to “will” and the inclusion of a sampling strategy be mandatory.
 - Some examples of adequate testing would be helpful here.

The Task Force did not think it prudent to set sample size thresholds, which depend on the question being asked and the statistical test employed.

- Steering committees and others have sometimes requested that measure developers expand the scope of measures submitted to include other populations. Based on this policy, the measure developer has to have testing done on the expanded populations. However it may be unrealistic to expect the measure developer to conduct this testing in a short time. Furthermore, the submitter may not have access to data required to test the measure in the expanded population.
 - For example, a measure submitted by CMS for the Medicare population may be applicable to other populations such as younger, commercially covered individuals; however, it was not developed for this purpose, nor was data for testing available.

Expansion of the target population depends on both clinical evidence (1c) and measure testing. Many NQF measures have artificial inclusion rules because of interest rather than evidence. If the measure developer is unable to include another patient population, then there often is no choice but to endorse multiple measures, which should be harmonized.

- Regarding bullet three: “Reliability and validity testing may be conducted for either the data elements used to calculate the measure score or the computed measure score to achieve an acceptable rating for endorsement, “ and bullet five: “Prior evidence of reliability or validity of data elements for the data source specified in the measure (e.g., hospital claims) can be used as evidence for those data elements:
 - If there is prior evidence of reliability and validity of the data elements which are derived from the EHR and it is an EHR measure than CMS agrees. If the measure uses more than one data source than all sources should be included.
 - Additional clarification on what will be required for data elements that have been tested as part of a data collection tool is requested. If the reliability and validity of these elements have already been established and the NQF steering committee disagrees with the previously conducted testing, how will this be addressed/resolved?
 - For measures based on standardized instruments, such as OASIS and MDS, the testing conducted during the development of the instrument should meet the requirements for testing at the data element level. It would be helpful to clarify that in this bullet point so that this is understood by both developers and the steering committees.
Reliability of a data element used in a measure would be the same as reliability testing of the data elements collected in an instrument if that's the data used in a measure.
 - Both the data elements and the measure calculation need to be reliable and valid; therefore, there should be evidence of both for measures to be used for accountability purposes. Commonly used types of data, such as claims, can be “deemed” by NQF to be valid and reliable at the data element level, relieving the developer of the burden of testing.
Prior testing of the data elements for the same data type can be submitted as evidence of reliability or validity of data elements; however, reliability and validity are not assumed for any data type.
 - This statement should be further expanded and incorporated into the scoring for reliability and validity. Claims are a commonly used data source for measures submitted from a variety of developers for various projects. It does not seem reasonable that measure developers should be responsible for establishing the reliability or validity of the data elements that are part of the claim and required for its primary purpose, payment. If data elements from claims are assumed by have validity and reliability, then it should be explicitly stated in the scoring criteria so that individual steering committees do not impose varying expectations of requiring data element testing. Furthermore, we believe that testing the reliability and validity of the coding of the claims on which a measure is based, will add very little to the understanding of whether or not the measure is suitable for endorsement.

Recommendation II

- Table 2 – guidance for rating the level of evidence for reliability and validity which is classified as high moderate and low.

- We concur with the following, “The scope of testing may be on a relatively small scale for initial endorsement, followed by further analyses to support continued endorsement at the time of review for maintenance of endorsement.” However, it is unclear how the extent of testing will impact the “high-medium-low” ratings described in Table 2 on page 14. The testing needs to be within acceptable norms, which includes type of testing, scope, and results. Furthermore, on page 20, in the “Recommendations for Additional Testing Required for Maintenance” the second bullet point states the evidence to be presented for maintenance is dependent upon the extent of testing and evidence provided at the time of initial endorsement. We feel that the interrelationships of these three sections of the document need to be strengthened. The extent of testing should be tied to the rating; the rating should be tied to the maintenance requirements. By clearly articulating these relationships, there will be greater consistency in terms of what measures become endorsed, and what additional testing may be required at maintenance.
- The task force indicated a rating of moderate would be acceptable for endorsement. CMS doesn’t think that all types of measures, (process, outcomes, structure, and efficiency et al.) should be rated using the same rating schema. The level of confidence in concluding the measure reflects poor /good quality based on scores is critical to making the decisions regarding selection of measures included in a P4P program. CMS is concerned that a moderate rating that includes criteria such as “systematic assessment of face validity of measure score as a quality indicator...” may not be sufficient for making some of these decisions. While there needs to a balance as to what evidence can reasonably be produced, CMS recommends that measures which will have significant impact related to policy decisions be held to a higher rating on a case by case basis.
 - Suggest that the required rating for endorsement be based on the specific measure or some other type of rating plan that addresses the different types of performance measures and/or potential impact of the quality measure. Reliability and validity are measure properties that apply to all types of measures. How they are assessed may vary based on the type of data available, but the basic principles do not vary.
- More discussion of how claims based measures will be evaluated would be helpful. For example, will claims meet the definition of “commonly used data elements with little question of reliability”? We suggest that NQF clearly state what data elements it will “deem” to be valid and reliable. This is necessary to ensure that all steering committees will evaluate measures consistently. The Tables A-1 and A-2 in the appendix suggest that reliability and validity testing should be conducted on codes commonly used for claims payment purpose such as ICD and CPT codes. We recommend that NQF “deem” data elements used for the purpose of claims payment.

Recommendation III

These comments are specific to EHR measures and are based on the following assumptions which may or may not be accurate:

- Only structured data from the EHR will be collected for reporting of clinical quality measures
- CMS programs will include EHR CQMs for pay for performance now and/or in the future
- Vendors/providers will retain flexibility in determining the location of and naming of data elements within the EHR even though the EHR used is ONC certified.
- **Specific Comments:**

1. Please review definitions and revise to maximize clarity and precision using universal, gold-standard definitions.
2. Bullet one and two: “there should be a period of time when measure scores are reported only to providers... and providers should be encouraged to conduct their own internal reliability studies”
 - CMS doesn’t think these statements are relevant to this document regarding scientific acceptability of measure properties.

These recommendations are regarding implementation of EHR measures, but do not affect the rating of reliability and validity

3. Table 4:

- With regard to the statement that validity of data elements is demonstrated by analysis of agreement between data elements exported electronically and data elements abstracted by the entire EHR, CMS doesn’t understand how this demonstrates validity of the data elements in an EHR. Does this test address the threats to the reliability and validity of data elements and measure scores for EHRs which are cited in the first paragraph under the recommendation heading e.g., 3) difference in use of data fields by different users. If the EHR vendors/providers have flexibility in determining the data element fields than those fields may be defined differently depending on the user. In paper medical records the interpretation of clinical documentation is more readily apparent by the clinical word used in the content of a sentence or phrase.
- Suggest other tests for validity of EHR data elements and measure scores.
- Recommend NQF not restrict “high” rating only to “EHR measure specifications (that use only QDS elements
- Alternatively recommend that high rating be applied to other data elements with tested/proven reliability, which may be represented in a publicly available registry of metadata for data elements and quality measures such as USHIK.

Registry data are not necessarily the same as data from EHRs

4. Suggest removing language that suggests using QDS as a requirement. The QDS is an implementation issue and not a scientific concept concerning reliability and validity testing. Rather, the measure developer should submit a plan for ensuring that data elements included in the measure are comparable across and between EHRs.
 - Line 443: What does it mean if the QDS is not used? Will this negatively affect endorsement of quality measures? At this time is this only a guidance document, or do the measure stewards have to use the QDS for the measure to be considered for endorsement? Will NQF provide support if the steward has trouble comprehending the QDS?
 - Recommend that data elements with tested/proven reliability need not be submitted for addition to the QDS, but alternatively may be represented in a publicly available registry of metadata for data elements and quality measures such as USHIK (United States Health Information Knowledgebase).

If a QDS element necessary for a measure does not exist, it can be submitted for inclusion concurrent with a measure submission.

5. Will the QDS continue to be public open source information? If so who will maintain & how frequently update the QDS?
 - Not clear how the QDS will get updated with additional elements since the source of data elements cannot only be existing measures.

- Please explain testing, availability and maintenance of the QDS since it will be used as a standard for electronically specifying clinical quality measures.

The QDS is open source and will be maintained by NQF

6. We do not understand what will occur if electronic specifications are developed outside of NQF. Who validates whether the specifications represent the original measure? Wouldn't that be the measure steward or, if it is an implementation issue, the implementer?
7. Line 2: states "four criteria." What are the four criteria? Please list.
A hyperlink to the criteria was provided; the criteria also were listed.
8. Line 30: mention of "de novo for EHRs." Please define/explain further what this terms means. "De novo" refers to a new measure; language was revised.
9. Line 78: typo. "true score" should be "true score" Correction made.
10. Line 188: Please define "physiologic."
11. Line 193: Insert word. "EHR data are not yet available..." Added "yet" as indicated.
12. Line 196: Insert example. "EHRs...(please provide example)."
13. Line 200-201: Suggest removing the following sentence "There are no perfect quality performance measures and there will be some error in all measurement.", as the essence of the statement is captured in the second sentence "Performance measurement science is an imperfect science."
14. Line 335-336: Table should be reformatted so "Moderate" is clear and not broken up.
15. Line 370 "Legal Record" should be removed from the sentence. It does not add any value to this document and is not correct in most cases at this time.
16. Ln373-9: add another potential source of error, incorrect data entry by end-user or clinician Stated in the document.
17. Ln 414-17: "If the error source is due to clinical data entry practices.....the error would not be amenable to changes to the EHR measure specifications but may necessitate the need for further evaluation and refinement of the measure." I actually disagree and think that if the measure passed testing except for clinical data entry practices, the measure should not need evaluation and refinement, the end-users would need education about proper clinical data entry. Modification made.
18. Ln 410-17: Could more clarification be given on test/retest?
19. Add statement/table about a combination table addressing R & V of measures that may be abstracted manually and electronically; The tables in the Appendix identify examples for different data types.
20. Clarify the uses/intent of the tables;
21. Use full terminology of critical data element or quality data element through the document so there is no confusion about intent of terms.

Recommendation IV

- The recommendation "Minimum Requirements for Untested Measures under Scientific Acceptability of Measure Properties" may be the standard to abide by rather than the ideal. Also included in Table 2, so required for all measures.
- With regard to risk adjustment, specific methods employed should be tracked and evaluated in the context of outcomes. Such tracking could inform future measure adjustment appropriateness and methods.

Recommendation V

- We agree that the requirements for maintenance measure testing should be related to the availability of data from implementation and the extent of testing already completed at the time

of submission. The extent of testing should be tied to the measure rating, and in turn should drive the requirements for the extent of testing to be required at maintenance.

- At the time of endorsement, NQF should make it clear to the developer the rating of the measure and what the expected extent of testing will be required at maintenance.
- For measures being evaluated under the NQF Maintenance Process, NQF should consider the validity of the measure in actual practice to evaluate the extent to which use of process measures are associated with positive outcomes. There needs to be recognition that the use of process measures for screening may reduce the likelihood of finding effects. e.g. consistent, widespread and reliable screening for pressure ulcer risk leads to appropriate interventions that reduces pressure ulcer development to such low levels as to be difficult to statistically measure as an outcome. This is actually a positive development but may interfere with validity and reliability measurement. Thus the NQF will want to recognize this dilemma and appropriately identify reasons to maintain a measure in this kind of situation.
- Additional suggestions regarding the evidence of reliability and validity that should be required for review for maintenance of endorsement:
 - If the reliability and validity tests reported at the time of submission are only from a limited set of data used to develop the measure, then the maintenance review should require confirmation of those tests after implementation. Sampling or data analysis should then come from the entire population for which the measure has been implemented.
 - 2d Measure Exclusions: For measures implemented after initial endorsement, analysis of exclusions should be required in the maintenance submission.
 - 2f Identification of Meaningful Differences in Performance: For implemented measures, measure evaluation criteria 2f should be required. This would especially be important for measures that are “topped-off” or approaching it.

Table 6 clarified. At the time of endorsement maintenance, reliability and validity testing should be based on data from implementation and focus on the level of the measure score. Opportunity for improvement is addressed under criterion 1b and was discussed in another report on Guidance for Evaluating the Evidence Related to the focus of Quality Measurement and Importance to Measure and Report.

Recommendation VI

- We appreciate the more succinct revision.
- Suggested Edits:
 - 2b6 If disparities in care have been identified, measure specification , scoring and analysis allow for identification of disparities through stratification of results... suggest adding word after identified to clarify where the disparities in care are identified.
 - There is much valuable information provided in the footnotes. This information is in fine print and does not appear in close proximity to the passages it further defines. For these reasons it is often overlooked. We suggest that NQF develop a different format for displaying this valuable information.
Will work with communications specialists to identify the best layout.

Recommendation VII

- Suggested Edits:
 - 2a24 Data Source – add electronic pharmacy data, electronic laboratory data,
Will review data type categories to allow selection of electronic types.

- 2a32 Level of Measurement/Analysis
- Suggest adding Eligible professional, eligible hospital and consumer.
The measure specifications indicate who is eligible for the denominator population.

Edward Q. Garcia III, MHS

Health Policy Analyst

Centers for Medicare & Medicaid Services

phone:410-786-6738 - fax:410-786-8532

Karen Pace, PhD, RN
Senior Program Director
National Quality Forum
601 13th Street NW, Suite 500 North
Washington DC 20005

July 13, 2010

RE: Review of "Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties"

Dear Dr. Pace,

On behalf of Boehringer Ingelheim Pharmaceuticals, Inc., I am pleased to provide comments in response to the National Quality Forum's (NQF) draft document titled "Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties." Boehringer Ingelheim (BI) supports advancement of performance measure development, implementation, and evaluation. These initiatives play an important role in improving the quality, outcomes, and delivery of patient care.

Hemal Shah, PharmD
Executive Director
Health Economics & Outcomes
Research
Boehringer Ingelheim
Pharmaceuticals, Inc.
900 Ridgebury Rd/P.O. Box 368
Ridgefield, CT 06877-0368

Given our interest in this area, we are committed to ensuring that performance measures are consistent with the most current and rigorous scientific evidence. We previously submitted comments on NQF's "Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement," in which we provided feedback on NQF's evidence grading systems, periodic evidence reviews and measure maintenance, roles of guideline and measure developers, and transparency in measure evaluation criteria and the methodologies used to calculate outcomes measure data. In this letter we focus on several areas specific to scientifically testing and evaluating measures:

- Clarify the Relative Importance of Validity to Reliability
- Further Define Testing Requirements for Measure Endorsement
- Ensure Measure Generalizability
- Define Evaluation Criteria for Registry-Based Measures
- Require Rigorous Testing in Review for Maintenance of Measure Endorsement
- Address Consequences of Adopting Unreliable or Invalid Measures for P4P and VBP
- Clarify Terms and Definitions

Clarify the Relative Importance of Validity to Reliability

We ask that NQF clarify the relationship between, and relative importance of, reliability and validity in testing measures for endorsement, as the report is inconsistent on these issues. Several sections of the report indicate that a measure's validity must only be considered after reliability has demonstrated. However, other sections suggest that reliability testing may be bypassed in some cases. For example, the "Background"

section states that “reliability is considered necessary, but not sufficient, for achieving a valid measure” (page 4); Appendix C provides a definition for “validity” specifying that “a measure cannot be valid without being reliable” (page 36); and the “Recommendations for Empirical Evidence of Reliability and Validity” section states that evidence of both reliability and validity should be expected for measures endorsed by NQF (page 8) and, further, that measures without empirical testing of both reliability and validity should be considered untested measures (page 10). However, the “Recommendations for Empirical Evidence of Reliability and Validity” section also states that “reliability testing of data elements could be bypassed if empirical validity testing of the data elements demonstrates acceptable validity” (page 10), which seems to contradict prior emphasis on first demonstrating reliability. BI supports a definition for validity that requires reliability as a baseline (as stated in Appendix C) and recommends that NQF consistently apply this definition when determining whether measures have been appropriately and rigorously tested.

Further Define Testing Requirements for Measure Endorsement

BI supports NQF’s policy of accepting measures for endorsement that have been rated as “moderate” (i.e. measures for which only the data elements or the computed measure score has been empirically demonstrated to be reliable and valid) (pages 11, 13), as we believe this level of testing is sufficiently rigorous to ensure endorsed measures are scientifically sound without placing undue burden on measure developers. This scheme also provides flexibility for endorsement of measures in therapeutic areas for which the current evidence base is relatively limited. However, we ask that the report clarify how the reliability and validity of the measure score can be analyzed without testing the elements that are “most critical” to the computed score. This guidance would encourage measure developers to focus on the most important constituent parts of the measures.

Further, NQF should consider that accepting simply “face validity” testing of the measure score (suggested as a potential strategy to minimize the burden of testing, page 10) may remove important incentives for developers to more rigorously test measure validity. BI recommends that NQF also provide more specific guidance regarding the sample size required if reliability and validity testing are conducted on a sample of the measured entities, as indicated as a potential strategy for reducing the burden of testing (page 10).

Ensure Measure Generalizability

BI requests that NQF clarify how both reliability and validity of measures will be evaluated to ensure generalizability, as measures should be reproducible and valid both in general and across diverse populations. BI encourages development of population-specific measures whenever possible. Measures that are not tailored to a specific population should be noted as such. This is a key step toward greater harmonization of the many performance measures that address certain disease areas; users of measures should be aware of which metrics are most appropriate to apply to their patients. Additionally, the report does not discriminate between internal and external validity, the latter of which concerns the extent to which the internally valid results of a study can be held true for other cases. Considering external validity is important for addressing a measure’s application across different populations (e.g. age, race, gender). Each of these issues underscores the importance of clearly defining and testing measures so they can be applied to the patients who may benefit most from their use.

Define Evaluation Criteria for Registry-Based Measures

BI recommends that NQF clarify the use of clinical registries as a data source to create measures. While NQF provides detailed recommendations for evaluating measures specified for and derived from electronic health records (EHRs) (pages 15 – 18), there is no mention of measures derived from clinical registries, although registries will increasingly play a prominent role in collecting data for research and quality improvement efforts. Since measures derived from registry data are also innovative, the report should clarify whether the criteria used to evaluate EHR measures will also be used to evaluate registry measures. If other criteria are used for registry-based measures, the criteria should be clearly defined.

Address Consequences of Adopting Unreliable or Invalid Measures for P4P and VBP

While BI agrees that using unreliable or invalid measures will have negative consequences for patients as described in the Background section of the report (page 5), it is also important to address the implications of using these measures for pay-for-performance (P4P) and value-based purchasing (VBP) initiatives. The public and private sectors are increasingly focused on payment reform, and Affordable Care Act established the Center for Medicare & Medicaid Innovation. If differential payment programs are based on unreliable or invalid measures, participating providers may be inappropriately assessed and subsequently incorrectly compensated. The report should discuss the importance of using evidence-based and rigorously tested measures in public reporting and value-based purchasing initiatives.

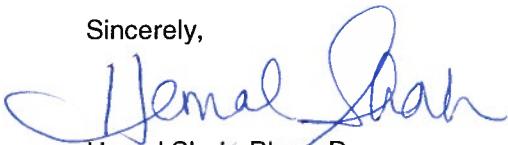
Clarify Terms and Definitions

Finally, the report introduces several key terms that should be more clearly defined. In particular, NQF recommends that “a formal and systematic testing of face validity of the measure score” be acceptable for validity as a strategy to minimize the burden of testing (page 10), but “face validity” is not defined in the report’s glossary or elsewhere in the report. Face validity pertains to the extent to which a measure *appears* to be valid to relevant experts without actual technical testing. Given the abstract nature of this concept, providing an example of the face validity testing NQF proposes may be especially useful for ensuring clarity. Additionally, although “measure score” is used throughout the report, it is never clearly defined; BI recommends that NQF clarify whether this term refers to measure ratio or calculation.

Conclusion and Next Steps

In conclusion, we look forward to working together to maximize the value of this and future initiatives for all relevant stakeholders, especially as new measures are developed.

Sincerely,



Hemal Shah, PharmD
Executive Director, Health Economics and Outcomes Research
Boehringer Ingelheim Pharmaceuticals, Inc.
900 Ridgebury Road / PO Box 368
Ridgefield, CT 06877-0368



AMERICAN
COLLEGE of
CARDIOLOGY
FOUNDATION

2400 N Street, NW
Washington, DC 20037-1153
(202) 375-6000
(800) 253-4636
Fax: (202) 375-7000



Learn and Live™

National Center
7272 Greenville Avenue
Dallas, Texas 75231-4596
Tel: 214-373-6300
Fax: 214-373-9818

July 13, 2010

Helen Burstin, MD, MPH
Senior Vice President for Performance Measures
National Quality Forum
601 Thirteenth St, NW
Suite 500 North
Washington, DC 20005

Via e-mail: performancemeasures@qualityforum.org; hburstin@qualityforum.org

Dear Dr. Burstin:

Thank you for the opportunity to comment on the draft National Quality Forum report, *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties*. On behalf of the American College of Cardiology Foundation (ACCF) and the American Heart Association (AHA) and our joint Task Force on Performance Measures, we thank you for your efforts to provide explicit guidance to Technical Advisory Panels and Steering Committees evaluating measures for endorsement. This guidance is also useful for measure developers like the ACCF/AHA, who will be required to submit data on the reliability and validity of their measures in order to obtain initial or ongoing endorsement. Although measures testing will continue to require some degree of judgment, we are hopeful that the recommendations in the draft report will promote the application of standardized criteria across measures and across projects. In general, the recommendations in the report do add clarity to the criteria that will be applied in evaluating the adequacy of testing data. We have carefully reviewed the draft report and respectfully offer the following comments for your consideration.

Validation of Administrative Data:

We are concerned that measures based upon administrative data are not held to a high enough standard. Much of the discussion regarding validation of administrative data involves only determining whether two coders arrive at the same administrative result. While such administrative vs. administrative validation is important and should be required, it misses the substantially more important issue of determining how administrative data compare with clinical data. This may not be as much of an issue where administrative data are used to ascertain outcomes like mortality or hospitalization (where administrative data are sometimes the criterion standard), but is extremely important in defining numerators, denominators, and risk adjustment variables.

Although Table A-2 refers to differences in measure scores due to multiple data sources as a potential threat to validity, we recommend that that validity assessments for all measures based on administrative data must include comparisons with clinical data. We should resist the temptation to use administrative measures, even if validated, as they are of questionable quality when compared with robust clinical measures.

Monitoring Implementation of the Recommendations:

The report notes that the proposed approach to evaluating scientific acceptability of measure properties should be monitored "to ensure it achieves the intent of endorsing reliable and valid measures and does not unduly impede endorsement of measures." The ACCF and the AHA agree that the proposed approach should be carefully monitored as it is implemented and revisited if

DM#367219

necessary. We would note that more stringent testing requirements and the contemporaneous time-limited endorsement policy, as summarized in Section IV of the report (Recommendations Related to Untested Measures), has significantly increased the burden of the process. This may pose substantial hurdles at a time when NQF is also pushing for more public reporting by measure implementers. We are concerned that this could potentially stifle public reporting efforts by registries and others. We recommend that the NQF develop an explicit process to monitor the impact of the new approach on measure development, including the impact on the proportion of measures that are based upon administrative data only.

Precision of Measurement:

The report includes limited discussion of precision of measurement in the context of reliability testing (Table A-1 and Footnote 8, lines 576-577). We suggest that the NQF provide more explicit guidance regarding the criteria applicable to the evaluation of the magnitude of changes in the measure score that are required to reflect true differences in quality.

Must-Pass Criteria:

Table 3 clearly and concisely summarizes the combinations of ratings on validity and reliability that would pass (or not pass) the criterion of *Scientific Acceptability of Measure Properties*. If criterion 1 of the measure evaluation criteria, *Importance to Measure and Report*, is not met, endorsement consideration must stop. The ACCF and AHA strongly recommend that the NQF make *Scientific Acceptability of Measure Properties* a must-pass criterion as well. If a steering committee and/or TAP determines that a measure does not pass the criterion of scientific acceptability, it should not move forward for further consideration for full endorsement.

Other specific comments on the report:

Table 4. Evaluation of Reliability and Validity of Measures Specified for EHRs: We support the emphasis on using the Quality Data Set (QDS) for measures specified for EHRs. This is important for two reasons. First, the QDS provides a framework that should allow comparability of data across settings if properly implemented. Second, requiring use of only QDS data elements as a condition for NQF endorsement will provide a strong incentive for EHR vendors to build QDS variables into what are now highly-variable products.

Section III, Recommendations for Measures Specified for EHRs (Item 3, Line377) refers to differences in use of data fields by different users or entry into the wrong EHR fields. This is a very widespread problem and may warrant further discussion of possible methods of validating comparability across users or systems for measures specified for EHRs.

Lines 551-555: Although the QDS is attempting to provide a standard for this data element for EHRs, patient preferences are not generally recorded, nor is there a standard format for recording it. Thus, there is likely to be a great deal of variability in capturing this—even with a QDS element—and is likely to be highly influenced by the treating clinician.

Footnote 9 – starting at line 578: This footnote to the validity section of the measure evaluation criteria has been revised in this report and now states: “Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator

Dr. Helen Burstin
July 13, 2010
Page 3

of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures).” This may need some revision since it seems to imply that any valid measure is necessarily redundant (needs to be highly correlated with another measure or is consistent with other conclusions). There should be room for measures on which some providers score well, even though they score poorly on others. For example, a provider could produce excellent outcomes, but be a poor communicator.

Thank you for your consideration of the comments above. We would be happy to discuss this with you at any time.

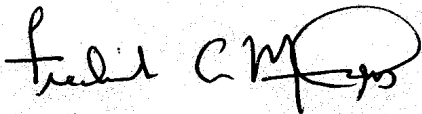
Very truly yours,



Ralph L. Sacco, MD, MS, FAHA
President, American Heart Association



Ralph W. Brindis, MD, MPH, FACC
President, American College of Cardiology



Frederick A. Masoudi, M.D, M.S.P.H., F.A.C.C., F.A.H.A.
Chair, ACCF/AHA Task Force on Performance Measures

cc: Joseph Drozda, MD, FACC
Ray Gibbons, MD, FACC, FAHA
Jack Lewin, MD
Janet Wright, MD, FACC
Rose Marie Robertson, MD, FAHA
Meighan Girgus
Gayle Whitman, PhD, RN, FAHA, FAAN
Charlene May, ACC Staff
Penelope Solis, AHA Staff

AHIP Comments on Measure Testing Report

AHIP supports the NQF's efforts to provide additional guidance to measure review panels and measure developers on the criteria for measure testing. Adding clarity around the standards used to evaluate measure testing will enable greater consistency in how measure review panels consider the reliability of proposed measures. In addition, AHIP supports NQF's subsequent revisions to their measure evaluation criteria and measure submission requirements based on the task force report's recommendations. We offer the following specific comments on the report.

General Comments

The focus of this report is testing a measure's reliability and validity. It is also important that a measure's feasibility of implementation be considered. AHIP recommends that the task force continue to review recommendations on feasibility testing.

NQF should consider the impact on measure developers in changing the evaluation criteria. Given the stringent criteria for measure testing, many measure developers may no longer be able to meet these revisions. NQF should assess the impact on the number and scope of measure developers that have the resources to meet these new criteria and submit measures to NQF. A decrease in the number of measure developers will impact the number of measures submitted for endorsement.

The NQF should provide a clear implementation plan for the changes to the measure evaluation criteria and measure submission requirements. It would be especially helpful to know how measures currently under review will be handled. We commented earlier this month on a set of outpatient measures which included fourteen measures recommended for time-limited endorsement due to lack of testing. We recommend that these measures be subject to the new measure testing requirements prior to member voting.

Recommendations for Empirical Evidence of Reliability and Validity

AHIP supports the recommendation that measures must undergo empirical testing of measure reliability and validity.

Recommendations for the Type of Testing and Results Needed to Demonstrate Scientific Acceptability of Measure Properties

The report includes evaluation ratings for reliability and validity testing. The recommended categories for rating reliability and validity testing appear reasonable and will assist in making more consistent recommendations. We concur that the evaluation ratings cited in the report have not been evaluated for their effect on the future pipeline for new measures and measurement endorsement. The ratings may need to be monitored and refined as they are implemented into the measure review process.

Recommendations for Measures Specified for EHRs

Reliability testing should be conducted among different EHR systems to assure that a measure's data elements can be consistently extracted.

Endorsed measures that are newly specified for EHRs should undergo validity testing that would compare measure scores produced from the new EHR specifications to the original measure specifications.

Recommendations Related to Untested Measures

We support NQF's actions to refine the criteria used to evaluate measures recommended for time-limited endorsement due to a lack of testing. The report should make it explicitly clear that ALL the criteria (e.g., incumbent measure does not address the specific topic of interest in the proposed measure; critical timeline must be met [e.g., legislative mandate]; and the measure is not complex) are required for untested measures to be considered for time-limited endorsement. This process needs to be clearly explained by an implementation plan. For example, will this only apply to new measures or measures within the review cycle?

Recommendations for Additional Testing Required for Maintenance of Endorsement

We agree with the recommendation that at the time of review for measure maintenance (i.e., every three years for endorsed measures), measures should be evaluated with the new measure evaluation criteria, including validity and reliability testing, for continued maintenance and encourage the addition of feasibility testing.

Recommendations for Modifications to the NQF Evaluation Criteria & Measure Submission Information

The modifications made to the NQF evaluation criteria and measure submission information appear to be consistent with the above recommendations.

From: Casey Jr., Donald [mailto:Don.Casey@atlanticealth.org]

Sent: Monday, July 12, 2010 7:09 AM

To: Performance Measures

Cc: Schneider, Eric; Mary Kay Crowther; Lea Anne Gardner; Steven Weinberger; Joseph Drozda; Melanie Shahriary; Helen Burstin; Maund, Christina; Meacham, Ing-Marie; Chang, Kyung; Isaac Starker; Dr Nielsen; Zampella, Edward; Dr. Robert Sussman; Casey Jr., Donald; Alice Jacobs

Subject: Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties-- Public Comment with attachment for supporting documentation

This comment is submitted to NQF on behalf of Atlantic Health, Morristown, NJ.

Please review the lead editorial in the July-August issue of the American Journal of Medical Quality, 2010; 25; 246; Donald E. Casey, Jr; **Performance Measurement 2.0: Time to Raise the Bar**; DOI: 10.1177/1062860610368483 which can be viewed at <http://ajm.sagepub.com/current.dtl>

I will also forward an electronic copy to NQF Staff.

The current methods for evaluating evidence for quality performance measures that are submitted for the NQF Consensus Development process is weak and currently not in line with better methods used by many professional medical societies who develop clinical practice guidelines and subsequent quality performance measures.

Thank you for your consideration.

Sincerely,

Don Casey

Donald E. Casey Jr., MD, MPH, MBA, FACP
Chief Medical Officer & Vice President of Quality
Chief Research Officer & Chief Academic Officer
Atlantic Health

475 South Street, PO Box 1905

Morristown, NJ 07962-1905

Associate Professor of Medicine

Mount Sinai School of Medicine, New York

973-660-3556

973-660-9116 (fax)

don.casey@atlanticealth.org

www.atlanticealth.org

Fortune Top 100 Best Companies to Work For 2009 & 2010

From: Jayne Hart Chambers [mailto:JChambers@FAH.org]
Sent: Tuesday, July 13, 2010 5:20 PM
To: National Quality Forum
Cc: Jayne Hart Chambers
Subject: Comments on the Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties

I tried to submit our comments on-line, but it wouldn't accept the comments. Please find comments below. Thank you. Please confirm that you received these comments.

The Federation of American Hospitals (FAH) appreciates the opportunity to comment on the "Review of Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties" report. The scientific acceptability of a measure endorsed by NQF is essential to both public confidence in the measure and to using the measures for quality improvement. The FAH is a strong supporter of the prioritization of the NQF four criteria for evaluating the suitability of quality measures for endorsement as voluntary consensus standards and was pleased when the National Quality Forum's Board of Directors prioritized the four criteria placing *Scientific Acceptability of Measure Properties* immediately following *Importance to Measure and Report*.

We recognize that defining *Scientific Acceptability*, the determination of reliability and validity, is not easy and commend the NQF Task Force on Measure Testing for its thoughtful report outlining elements to be considered when evaluating the scientific evidence supporting a measure. The report does an admirable job of parsing the issues and defining reliability and validity and methodologies for assessing both and recognizing that there are degrees of reliability and validity. The FAH supports the testing guidance that is defined in the report. In particular, we find Table 3, the matrix for evaluation ratings for reliability and validity of measures to be very helpful.

The FAH looks forward to further guidance on reliability and validity as more and more measures are specified for collection through electronic health records. We also look forward to similar Task Force reports on feasibility and public reporting.

Please note: FAH MOVED...

Jayne Hart Chambers
Sr. Vice President Strategic Policy &
Corporate Secretary
Federation of American Hospitals
750 9th Street, N.W.
Suite 600
Washington, DC 20001-4524

202-624-1522 (office)
202-262-1010 (mobile)