NATIONAL QUALITY FORUM

# Memo

**TO**: Consensus Standards Approval Committee (CSAC)

**FR**: Margaret Terry, Senior Director, Wunmi Isijola, Administrative Director, Christy Skipper, Project Manager, and Yetunde Ogungbemi, Project Analyst

**RE**: Neurology Project – Measure #2876

**DA**: August 31, 2016

The CSAC will review one recommendation from the Neurology Standing Committee.

Measure Not Recommended for Endorsement:

- 2876, Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity

**MEASURE BACKGROUND:** During the August 9th conference call, the CSAC reviewed measures from the Neurology Standing Committee. One measure, *#2876 Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity* was pulled from discussion because the developer submitted a request for reconsideration. CMS/Yale CORE submitted a reconsideration request on the grounds that the Committee "did not receive appropriate guidance on the application of NQF's measure evaluation criteria" and the developer "did not feel that we had adequate opportunity to respond to these issues during the committee's discussion" (see Appendix A for the complete reconsideration request).

The reconsideration request was reviewed by the CSAC Co-Chairs and the following section includes the Co-Chairs' decision and rationale. In addition, the memo includes additional issues raised by the measure developers in the reconsideration request and the responses of the Neurology Standing Committee that occurred when the Committee was evaluating the measure.

**CSAC RECONSIDERATION DECISION AND RATIONALE – NQF MEASURE #2876:** After reviewing the reconsideration request for Measure *#2876 Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity,* the CSAC Co-Chairs did not support the request for reconsideration. The measure developers raised several issues regarding the CDP process in the request for reconsideration. While NQF recognizes that not every Neurology Standing Committee member was on the post-comment call, there was quorum on the call with 87% of the Committee members voting. (The vote for validity during the in-person meeting was H-0; M-13; L-8; I-1 and the vote for validity during the post-comment call was H-0; M-3; L-12; I-5.) During the in-person meeting, a vote for the Overall Suitability of the measure was taken. Given that the Committee was unable to reach consensus on the must-pass validity criteria, the vote for Overall Suitability should not have been taken and NQF will make sure that this process is observed going forward. Both CSAC Co-Chairs agreed that the CDP process had been followed.

The CSAC Co-Chairs agreed with the Neurology Standing Committee's decision to not recommend the measure for endorsement based on the discussion of validity at the in-person meeting and during the post-comment call. The CSAC Co-Chairs supported the concern of the Committee regarding the measure developer's inability to test the measure using ICD-10 codes since the codes will not be implemented until October 2016. While the measure developer provided risk-standardized mortality rates using data from Medicare administrative claims and data from the Get with the Guidelines-Stroke Registry, the Committee noted the measure developer could not validate the National Institutes of Health Stroke Scale (NIHSS) against ICD-10 codes at this time. The CSAC Co-Chairs also acknowledged that, while the Committee discussed the issues of missing data and patient preference versus quality of care at length, the primary reason for upholding the Committee's decision was based on the lack of testing using ICD-10 codes.

**Issues raised by the developer in the reconsideration request:**

1. The Impact of Missing NIHSS scores: The Committee erroneously believed that the impact of missing NIHSS scores was not assessed. In fact, the developer used a multiple imputation technique to account for missing data and then conducted analysis that showed that there was no relationship between missing data and performance measure results.
2. The Data Source for NIHSS Scores: In their discussions, the Committee described the NIHSS scores used in the development and testing of the measure as "simulated" rather than "real world" data. However, the data used were actual NIHSS scores taken from the GWTG registry. The NIHSS scores included in the GWTG registry were extracted directly from patient medical records and reported to the registry.
3. The Use of Multiple Imputations: The Committee expressed concern that if multiple imputations are used in implementation of the measure, hospitals might be incentivized to not report actual NIHSS scores for their stroke patients. The developer has clarified that CMS is not—at this point in time—saying that imputation will be used; instead, they stated that CMS would determine the best approach for handling missing data if the measure is implemented.

**Committee discussion regarding issues raised by the developer in the reconsideration request:**

- Issue #1 –The Committee acknowledged the developer's missing-data analysis for the NIHSS score.
- Issue #2 –The Committee erroneously used the term "simulated" data; however, Committee members understood that the NIHSS scores used by the developer were actual scores from real patients.
- Issue #3 –The Committee did express concerns about potential gaming if imputation is used when the measure is implemented, but acknowledged the developer's response that CMS' intention would be to determine the best approach of handling missing data, if the measure is implemented.

**Additional Committee concerns related to the issues raised by the developer in the reconsideration request:**

- Because the ICD-10 coding for the NIHSS score has not yet been implemented in claims data, the Committee was concerned with the implementation and accuracy of the NIHSS in claims data.

- The NIHSS score was missing from the GWTG registry for approximately 17% of the patients included in the developer's testing dataset.   It is unclear how often the NIHSS scores will be missing from the ICD-10 claims data once implemented.
- It is not clear whether the NIHSS score data are missing at random; if not, this could impact the validity of the measure.
- It is unclear whether or not imputation will be used when the measure is implemented using the ICD-10 coding for the NIHSS score.

# Appendix A. Reconsideration Request for #2876

Center for Outcomes Research and Evaluation (CORE)

1 Church Street, Suite 200

New Haven, Connecticut 06510-3330

Phone: 203-764-5700 Fax: 203-764-5653

MEMORANDUM

TO: David Knowlton, MA and David Tirschwell, MD, MSc, Committee Chairs, Neurology Project 2015-2016 Measure Endorsement Project

FROM: Karen Dorsey, MD, PhD, and Susannah Bernheim, MD, MHS, Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation

THROUGH: The Centers for Medicare and Medicaid Services, Lein Han, PhD, Helen Dollar-Maples

RE: Request for Reconsideration of Measure #2876: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity

DATE: Monday, August 8, 2016

In April 2016, the Neurology Project Measure Endorsement Committee strongly voted to recommend Measure #2876: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity for endorsement (17-Y; 5-N) but did not reach consensus on the validity criteria. Therefore, at the June 23 post-comment call the measure was discussed a second time. This time, with a more limited set of members voting, the measure did not pass the validity criteria (H-0; M-5; L-9; I-2), thus the measure was not recommended for endorsement despite the strong overall support it received in April in-person meeting.

REASON FOR REQUEST FOR RECONSIDERATION

NQF sets forth two grounds for the appeal of endorsement recommendations – the first is the inappropriate application of NQF's measure evaluation criteria; the second is NQF's consensus development process was not followed. If either of these conditions are met and if the developer chooses to exercise its discretion to file an appeal, "the developer must submit a written request to the Committee indicating the specific evaluation criteria or sub-criteria that the developer believes was not applied properly to the specific information as submitted and evaluated by the Committee" and/or "the developer must send a written request to the CSAC citing the issues with a specific CDP process step, how it was not followed properly, and how it resulted in the specific measure not being recommended."2

After careful consideration of the Committee's proceedings and review of the Neurological Conditions (2015-2016) Draft Report , we respectfully appeal the Committee's recommendation against the endorsement of Measure #2876 primarily because we feel the committee did not receive appropriate guidance on the application of NQF's measure evaluation criteria, and secondarily because we did not feel that we had adequate opportunity to respond to these issues during the committee's discussion, including the fact that a senior member of our team requested a public line to respond but could not be heard by the committee until the very end of the discussion.

Based on the NQF's guidance for evaluating validity, and our experience presenting similar empirical evidence to committees, we met criteria in steps 1, 2, 3, 6, and 7 which should have qualified the measures for either a "high" or "moderate" score on validity. Below we describe the validity issues raised and considered by the Committee  and our response to each.

A. Potential Threats to Validity Raised During Committee Discussion

- The Impact of Missing NIHSS scores: There was a 17% rate of missing NIHSS scores (National Institutes of Health Stroke Severity score) in the measure development dataset and the committee stated that developer did not test the potential impact of these missing data on measure validity.

- The Data Source for NIHSS Scores: The committee stated developer used "simulated" NIHSS score data rather than "real world" data and therefore did not test the validity of the real world NIHSS scores.

- The Use of Multiple Imputation: The developer used multiple imputation during measure development which, if used for a fully implemented measures, could incentivize hospitals to not report actual NIHSS data on patients.

Yale CORE Response:

- The Impact of Missing NIHSS Scores: The contention that we did not test the impact of the missing data is not accurate. We agree that the rate of missing NIHSS scores in the development dataset, 17%, was sufficiently high to require that we assess the potential impact of missing data on the measure and that we use appropriate statistical techniques when using this variable in the measures' risk model. Therefore, we used multiple imputation to account for missing data, and performed diagnostic analyses to evaluate any potential bias related to missing data in measure results. These analyses and their results were described in detail in Section 2b.7 in the testing attachment and demonstrated there is not a relationship between missing data and performance scores. This analytic work was not acknowledged by the Committee.

- The Data Source for NIHSS Scores: The concern raised about the use of "simulated" data rather than real world data is not accurate. Although we used data from the Get With The Guidelines (GWTG)-Stroke Registry, the NIHSS scores in that dataset were extracted directly from patient medical records and reported to the registry. They are the gold standard scores recorded by clinicians at the time that treatment for stroke was provided. There may have been a misunderstanding about the source of the NIHSS score in our data; it is in fact abstracted from the medical record. This measure is developed for use with administrative claims data. There is an ICD-10 code for each NIHSS score (from 0 to 42). When ICD-10 codes for NIHSS are in use this fall, medical coders will only need to select the proper ICD-10 code for the NIHSS score recorded by the clinician in the medical record. Given the straightforward nature of applying these ICD 10 codes, we do not have significant concern that selection of the proper code for each NIHSS score would lead to inaccuracies, as long as it is recorded by the clinician. We will be able to test the accuracy of these ICD-10 codes once they are implemented, after October 1, 2016 and will do so before the measure is implemented. However, we disagree with the assertion that this represents a threat to validity. We used the most accurate source on NIHSS scores available (from the medical record) and the same data medical coders would use to complete a claim that includes NIHSS score. There was no "simulated" data used. Therefore, we do not believe this measure warrants a vote of "insufficient" on the basis of "simulated" data. It is not uncommon that measures are endorsed by NQF for use with a new data collection approach in the future. The validity is assessed on the development data with an understanding of ongoing reevaluation when the measure is in use. The patient-reported outcomes for instance and hybrid measures are both examples of this precedent.

If Committee members are concerned about the feasibility of collecting this data through claims in the future we believe they should be guided by NQF to express such concerns related to the feasibility, rather than validity, of the measure. We do not think this is an issue of validity of the measure.

- The Use of Multiple Imputation: The Committee's concern that our imputation methods will incentivize hospitals to skip documentation does not accurately reflect what was stated in our submission materials. The missing NIHSS values were imputed using the standard statistical method of multiple imputation based on the claims data for the development of the risk-model for the measure. Multiple imputation allowed us to develop and fully test the measure's risk model and did not threaten the validity of the risk model.

Although imputation was used to develop and test the measure, CMS is not proposing to use this approach for calculating results when the measure is implemented. We used imputation to mitigate the impact of the missing NIHSS values in the stroke registry data and to be able to include the full cohort of eligible admissions in the measure. It was our determination that imputation was the most valid way to develop and test the measure's risk model. However, in order to implement the measure hospitals would need to report the NIHSS on all or nearly all of their ischemic stroke patients. We believe this is feasible given the introduction of International Classification of Diseases 10th revision (ICD-10) codes for NIHSS scores scheduled to begin in October 2016.

It is possible that Committee members voted "insufficient" on the validity criteria due to the use of imputation for handling missing data in measure development assuming it would be used in implementation but this is not true. We stated in the in-person meeting that CMS would determine the best approach to handling missing data during implementation. As noted above, if Committee members are concerned about the feasibility of collecting this data through claims in the future we believe they should be guided by NQF to express such concerns related to the feasibility, rather than validity, of the measure. We do not think this is an issue of validity of the measure.

B. Concerns about Empirical Validity Testing Raised by the Committee

- Risk Model Validity Testing: A Committee member noted that the empiric validity testing of the measure included a comparison of two measures which both used the same information for NIHSS pulled for the claims measure (this measure) as for the registry measure (which also included variables that would not have been captured in the claims measure).

- Exclusion Due to Patient Care Preferences: The Committee weighed whether the measure was truly assessing quality if patient preferences (e.g., exclusion of patients with comfort measures only and do not resuscitate orders) had not been fully incorporated, noting that two patients with the same stroke severity may still have different care preferences. This led to a larger concern of the Committee as to whether the measure is actually measuring facility practices rather than quality of care.

- Lack of SDS Factors in Risk Adjustment: The Committee also noted that the SDS factor, race, was not included in the final risk adjustment model, although race was a negative predictor of mortality. The Committee felt that reduced mortality for African-American patients was not a reflection of the quality of care but rather that it likely reflected that African-Americans have more aggressive preferences for care.

Yale CORE Response:

- Risk Model Validity Testing: We do not agree with the contention that the common use of the NIHSS invalidates our empiric validity testing. Our test of the validity of the risk model demonstrated that a model that includes the NIHSS score and patient comorbidities from claims data produces similar discrimination as does a model that includes NIHSS score and physiologic data (laboratory test results and vital signs) derived from the registry. The purpose of this test was to compare a model that relies on claims data with one that uses data from the medical record, which is considered the gold standard data source. The discrimination of the two models was quite good (0.821 and 0.794). Additionally, we found that the Pearson correlation coefficient, weighted by hospital volume, of the standardized rates from the administrative and registry models was 0.95647. We agree that because NIHSS score is a strong predictor of mortality, it is likely responsible for the increased discriminatory power of both models and may, in part, drive the correlation between the model results. However, this only reinforces the assumption that NIHSS score is a critical variable to include in a stroke mortality measures. We do not agree that the inclusion of the NIHSS score in both models negates their comparison as a test of validity of the claims-based risk model. In fact, as we note above the approach to coding NIHSS is likely to be identical to the approach to recording it in registry data so it is a reasonable surrogate for data that does not yet exist in claims but will in the upcoming year. There is only one NIHSS score for each index admission. Therefore, we do not agree that the use of this gold standard data source for NIHSS in both models led to an unfair or incomplete validity test. We are able to show that our surrogate for a claims measure performs similarly and has highly correlated scores with a clinical registry based measure of stroke mortality. Under prior NQF projects a similar approach to comparing a claims-based model with a clinical model has be considered to meet moderate validity.

- Exclusions Due to Patient Care Preferences: We agree with the Committee that it would be ideal to have a method or source of data that allows for accurate identification and exclusion of those patients for whom survival is not the goal of care. However, currently there is no such source of data available for Medicare beneficiaries. We do not agree that this limitation invalidates the stroke mortality measure. This is a limitation for all of our mortality measures in public reporting. The measure does exclude patients who are admitted to hospice before or on the day of admission (within the first 24 hours). In the registry data we find that approximately 3% of stroke patients elect comfort measures. Our exclusion for hospice patients captures one third of the 3%. However, most patients who elect to receive comfort measures do so after the first 24 hours of the admission. Even if the data captured this population perfectly, it is problematic to exclude these patients from the measure because we cannot know whether their decision was due to the severity of the initial stroke and low likelihood of functional recovery (which is accounted for in the risk model) or if it was due to poor quality of care delivered after they were admitted to the hospital (which we would not want to exclude).

In addition, the inclusion of the NIHSS score in the measure risk model should mitigate the impact of the unequal distribution of patients with the most severe strokes across hospitals. These patients, those with higher NIHSS, are the patients most likely to face a poor prognosis and elect to receive comfort measures. Although we agree that it would be ideal to exclude patients for whom avoidance of death is not the desired outcome, it is not feasible to do so perfectly while fully preserving the signal of quality that the measure is deigned to capture. Similar measures have been in reporting without any evidence of unintended consequences. For this measure, the addition of NIHSS better accounts for variation in the proportion of patients with severe stroke, and therefore those most likely to elect for comfort measures across hospitals. We ask that the Committee reconsider the improvement that this measure represents in properly accounting for patients with poor prognoses due to inclusion of the NIHSS score. We do not agree that the currently limitation regarding identifying patient care preferences and exclude patients who elect to receive only comfort care warrants a score of “insufficient” on the validity criteria.

- Lack of SDS Factors in Risk Adjustment: Although differences in mortality rates were observed among African-American patients compared with all other racial groups and among patients with low SES indicators compared with all others, these differences were very small in the fully risk-adjusted model. The mean absolute change in hospitals’ RSMRs when adding a dual eligibility indicator was 0.00006%. The mean absolute change in hospitals’ RSMRs when adding a low SES AHRQ indicator was 0.00009%. The mean absolute change in hospitals’ RSMRs when adding a race indicator was -0.00064%. These findings did not support including these variables in the measure’s risk model. Moreover, we would argue that we cannot be certain what contributes to the lower mortality risk for African-American patients and that assuming this is due to differences in preference may or may not be true. However, if it is true and those preferences lead to better outcomes the measure should reflect that, as it does without adjustment. We ask that the Committee consider that we tested the potential inclusion of SDS factors and reported these results in our submission materials demonstrating that inclusion of these data was not supported by the empirical evidence.

Thank you for your consideration.

Sincerely,

Karen Dorsey

Susannah Bernheim