TAB 1-B

TO: NQF Executive Committee
FR: Helen Burstin, Chief Scientific Officer
Marcia Wilson, Senior Vice President, Quality Measurement
RE: Appeal of Measures for the Readmissions 2015-2017 Project
DA: February 28, 2017

ACTION REQUIRED

The Executive Committee will discuss an appeal of the endorsement of five measures in the Readmissions 2015-2017 project and determine whether to ratify the Consensus Standards Approval Committee's (CSAC) decision to uphold endorsement. The appealed measures are:

- 0330: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following heart failure (HF) hospitalization (CMS)
- 0506: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following pneumonia hospitalization (CMS)
- 1789: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR) (CMS)
- 1891: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following chronic obstructive pulmonary disease (COPD) hospitalization (CMS)
- 2881: Excess days in acute care (EDAC) after hospitalization for acute myocardial infarction (AMI) (CMS)

BACKGROUND

In accordance with the National Quality Forum (NQF) Consensus Development Process (CDP), the measures recommended by the Admissions and Readmissions Standing Committee were released for a 30-day appeals period, which closed on January 11, 2017. The readmissions project remains under the existing appeals process. NQF received one appeal of its endorsement of the measures listed above from Adventist Health System (AHS).

- Appendix A – Appeal Letter from the (AHS)
- Appendix B – Measure Developer Response to the Appeal

APPEAL OF ENDORSEMENT

AHS raised concerns that these measures are used by the Centers for Medicare & Medicaid Services (CMS) in the Hospital Readmission Reduction Program (HRRP) (#0330, #0506, and #1891) and the Hospital Inpatient Quality Reporting (HIQR) Program (all five measures). The results of measures in HRRP are used to determine payment penalties for excess readmissions. Information from the HIQR Program is publicly reported on the Hospital Compare website.

AHS appeals the endorsement decisions on the grounds that 1) procedural errors were made that were likely to affect the outcome of the original endorsement decision and 2) new

information or evidence has become available that is reasonably likely to have affected the outcome of the original endorsement decision.

Procedurally, the appellants state that the measure did not meet NQF's standards for reliability and that the member voting did not achieve consensus. In addition, the appellants note that new information had become available following the endorsement decision that likely would have affected the endorsement decision. In December 2016, the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (ASPE) published "Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs" https://aspe.hhs.gov/system/files/pdf/253971/ASPESESRTCfull.pdf. The second item is a New England Journal of Medicine (NEJM) perspective titled "Should Medicare Value-Based Purchasing Take Social Risk into Account?" published on December 28, 2016, and attached to this memo. Both the report and the article are discussed in more detail in AHS' appeals letter on pages 5-6 and 11-12 of these materials.

NQF RESPONSE

Reliability

NQF does not maintain a specific standard for reliability. When developers use test-retest reliability to assess the Intra-class Correlation Coefficient (ICC), NQF provides information on the conventions put forth by Landis and Koch in the preliminary analysis developed for each measure. (The conventions of Landis and Koch provide guidance in interpreting statistical results.) However, the Standing Committee retains the ability to make its own assessment on the reliability of a measure.

Member Vote

Once a project standing committee has reviewed all of the comments submitted during the public and member commenting period and made any revisions to the draft report, NQF members vote on the candidate standards recommended for endorsement by the committee. All candidate consensus standards recommended for endorsement will proceed to the next step in the consensus development process: decision by the CSAC. NQF staff provides a summary of the results of the member vote to the CSAC. If the member voting does not reach consensus >60%, CSAC has the option to request a re-vote or an all-member meeting.

The memo to the CSAC on the Readmissions 2015-2017 project highlighted the member voting results. The memo noted that one of the recommended measures was approved with 67% or higher. The memo also stated that Representatives of 19 member organizations voted; no votes were received from Consumer, Supplier/Industry, or Public/Community Health Agency Councils. Detailed breakdowns of the vote on each memo were provided in an appendix.

The CSAC did not request a re-vote or an all-member meeting on the voting results of #0330, #0506, #1789, #1891, or #2881.

CSAC REVIEW

CSAC considered this measures appeal on February 14, 2017, and voted to uphold endorsement of the measures. The CSAC noted that NQF does not currently maintain set standards for reliability and that the member vote is an input into the endorsement process, but not dispositive. The CSAC determined that the ASPE report and the NEJM perspective did not introduce new evidence. The CSAC recognized the continuing concerns about the effects of social risk factors on measures of readmissions. In light of that, the CSAC previously developed language to accompany its recommendations on these measures:

At this time, the CSAC supports continued endorsement of the hospital readmission measures without SDS adjustment based on available measures and risk adjustors. The CSAC recognizes the complexity of the issue and recognizes that the issue is not resolved.

The CSAC recommends the following:

1. *SDS adjustor availability be considered as part of the annual update process;*
2. *NQF should focus efforts on the next generation of risk adjustment, including social risk as well as consideration of unmeasured clinical complexity;*
3. *Given potential, unintended effects of the readmission penalty program on patients, especially in safety net hospitals, the CSAC encourages MAP and the NQF Board to consider other approaches.*
4. *Directs the Disparities Standing Committee to address unresolved issues and concerns regarding risk adjustment approaches, including potential for adjustment at the hospital and community level.*

The CSAC reiterated this language and agreed it continues to address the committee's concerns about adjustment for social risk factors.

APPENDIX A: APPEAL LETTER FROM ADVENTIST HEALTH SYSTEM

Measure 2881 Appeal Request
Adventist Health System, Submitted January 11, 2017

Adventist Health System (AHS) wishes to appeal the decision to endorse the excess days in acute care (EDAC) after hospitalization for acute myocardial infarction (AMI) (NQF# 2881). We believe our interests will be directly and materially affected by this recently endorsed consensus standard because will be used by the Centers for Medicare and Medicaid Services (CMS) in the Hospital Inpatient Quality Reporting (IQR) Program. This program has a substantial impact on AHS facilities. HIQR measure results are publicly reported and affect public perception of AHS hospital facilities.

We wish to appeal the endorsement of this measure on grounds that 1) procedural errors were made that were likely to affect the outcome of the original endorsement decision and 2) on the grounds that new information or evidence has become available that is reasonably likely to have affected the outcome of the original endorsement decision.
It is the view of AHS that two significant procedural errors were made in the decision to endorse this measure.

First, the Standing Committee should not have found that this measure meets the NQF's standard for reliability. The developer used a "test-retest" approached to assess reliability. The agreement between two RSRRs, as measured by Intra-class Correlation Coefficient (ICC), was 0.54. The measure developer, in its response to comments, cited a convention that "describes the ICC values as moderate (0.41-0.60) for this measure" (Landis JR and Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977; 33:159-174). AHS agrees with Landis and Koch [1977] that "[a]though these divisions are clearly arbitrary, they do provide useful 'benchmarks' for the discussion of [a] specific example […]" Furthermore, we agree with the developer that the ICC values of this measure could be described as "moderate" under the "benchmarks" put forward by Landis and Koch [1977]. However, AHS believes that NQF committees should only assess a measure as meeting NQF standards for reliability if that measures meets a threshold of reliability commensurate with the impact of its current or prospective use. It is our opinion that achieving a "moderate" benchmark of reliability is not sufficient for the endorsement of substantially impactful measures. We find measures that are used in public reporting or payment programs, such as the HIQR and HRRP, to be substantially impactful. Hence, AHS believes that for such measures to be awarded endorsement they should first be assessed as meeting a reliability benchmark or "strength of agreement" that is "substantial" according to Landis and Koch [1977]. Thus, we conclude that, according to the Landis and Koch [1977] convention cited by the developer, the "substantial" reliability "benchmark" for this measure would be an ICC value of 0.61-0.80. In other words, AHS believes that, by the developer's own scale, this measure should have achieved an ICC of at least 0.61 to meet the NQF's standard for reliability.

Second, this vote did not achieve consensus among the NQF member organizations that cast votes during the endorsement proceedings. Six members voted in favor of endorsement of the measure and seven members voted against endorsement of the measure. That is an approval rate of 46 percent. AHS believes that a member voting approval rate of 46 percent is insufficient for NQF endorsement. We think it is also worth pointing out that only three out of the eight measure councils had more than two members cast votes. Of these three councils, only one approved of the measure. We find it alarming that a measure can achieve NQF endorsement despite receiving more votes of disapproval than approval. It is our opinion that the NQF's status as the "gold standard" of quality measurement and as a consensus standard body (as defined by the Office of Management and Budget) could be in serious jeopardy if this trend persists.

It is also the view of AHS that two pieces of new information have become available since the CSAC made its endorsement decision that are reasonably likely to affect the outcome of the original endorsement decision.

The first item was a December 2016 report published by the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (ASPE) titled "Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs." The report concluded that "social factors are powerful determinants of health. In Medicare, beneficiaries with social risk factors have worse outcomes on many quality measures, including measures of processes of care, intermediate outcomes, outcomes, safety, and patient/consumer experience, as well as higher costs and resource use. Beneficiaries with social risk factors may have poorer outcomes due to higher levels of medical risk, worse living environments, greater challenges in adherence and lifestyle, and/or bias or discrimination. Providers serving these beneficiaries may have poorer performance due to fewer resources, more challenging clinical workloads, lower levels of community support, or worse quality."

In addition, the report recommended that "measuring and reporting quality for beneficiaries with social risk factors, setting high, fair quality standards for all beneficiaries."

The second item was a New England Journal of Medicine (NEJM) article titled "Should Medicare Value-Based Purchasing Take Social Risk into Account?" that was published on December 28, 2016. This article noted that "beneficiaries with social risk factors had worse outcomes on many quality measures, regardless of the providers they saw, and dual enrollment status was the most powerful predictor of poor outcomes." In addition, the article highlighted that "providers that disproportionately served beneficiaries with social risk factors tended to have worse performance on quality measures." The article also recommended that "we should measure and report quality of care for beneficiaries with social risk factors."

AHS believes that the HHS ASPE report and NEJM article highlight what the NQF's Readmission Committee stressed as "the high risk of unintended consequences related to adjustment of these measures for SDS factors and the need to reevaluate these measures as the field continues to move forwards." It is our view that these reports represent advancements in the

field that the committee suggested would necessitate reevaluation. Therefore, endorsement of this measure should be revoked because the information presented by these reports is reasonably likely to have affected the original endorsement decision.

Measure 0330, 0506, 1789, 1891, 2881 Appeal Request
Adventist Health System, Submitted January 20, 2017

To Whom It May Concern:

I am writing on behalf of Adventist Health System (AHS) to appeal the decision to endorse the following NQF Readmission measures:

- NQF #0330: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Heart Failure (HF) Hospitalization
- NQF #0506: Hospital 30-day, All-cause, Risk-Standardized Readmission Rate (RSRR) Following Pneumonia Hospitalization
- NQF #1789: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)
- NQF #1891: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Chronic Obstructive Pulmonary Disease (COPD) Hospitalization
- NQF #2881: Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI)

We believe our interests are directly and materially affected by these recently endorsed consensus standards because they are used or are proposed to be used by the Centers for Medicare and Medicaid Services (CMS) in the Hospital Inpatient Quality Reporting (HIQR) Program and the Hospital Readmission Reduction Program (HRRP). These federal quality measurement programs are substantially impactful. HIQR measure results are publicly reported and thereby affect public perception of AHS hospital facilities. HRRP measures results are used to adjust payments that AHS hospital facilities receive from Medicare.

A recent study, titled “Reliability of 30-Day Readmission Measures Used in the Hospital Readmission Reduction Program,” that was published in the Health Services Research journal, concluded that “[m]any of the RSRRs employed by the HRRP are unreliable” and “few hospitals have acceptable reliability on all measures for which they are assessed by HRRP.” Furthermore, Adventist Health System — NQF Readmission Measures Endorsement Appeal the study found that “one quarter of payments [penalties] for excess readmissions are associated with unreliable RSRRs.”

According to the authors, for many hospitals “[HRRP] penalties are likely the result of statistical noise and unlikely to provide constructive information about areas needing improvement.” AHS believes that one quarter of the payment penalties tied to readmissions measures is substantial and material.

We wish to appeal the endorsement of these measures on the grounds that procedural errors were made that were likely to affect the outcome of the original endorsement decision. We

also wish to appeal the endorsement of these measures on the grounds that new information or evidence has become available that is reasonably likely to have affected the outcome of the original endorsement decision.

Procedural Errors

It is the view of AHS that two procedural errors were made in the decision to endorse these measures.

First, we believe that the recommendation and subsequent endorsement of several of these measures was inconsistent with NQF's Scientific Acceptability criterion for reliability.

Second, we believe that the Consensus Standards Approval Committee (CSAC) did not appropriately consider the results of the NQF Member Voting step of the NQF Consensus Development Process (CDP) before moving forward with its recommendation to endorse these measures.

New Information or Evidence

It is also the view of AHS that two pieces of new information have become available since the CSAC made its endorsement decision that are reasonably likely to affect the outcome of the original endorsement decision.

The first item was a December 2016 report published by the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (ASPE) titled "Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs."

The second item was a New England Journal of Medicine (NEJM) article titled "Should Medicare Value-Based Purchasing Take Social Risk into Account?" that was published on December 28, 2016.

Procedural Error — Reliability

Criterion 2 of NQF's Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement specifies that "[m]easures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria."

Subcriterion 2a2 requires that "[r]eliability testing demonstrates that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise."

A RAND Corporation Technical Report titled "The Reliability of Provider Profiling: A Tutorial," describes reliability as follows:

> Conceptually, it is a ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in

> performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance.

Using simpler terms, a study published in the Annals of Thoracic Surgery notes that "reliability of 1.8 means that 80% of the variance in outcomes is due to true differences in performance while 20% of the variance is attributable to statistical 'noise' or measurement error."

AHS is appealing the endorsement of several readmissions measures recently endorsed by NQF because we believe they were misjudged by the Standing Committee as having met Subcriterion 2a2. In particular, we find that the Committee used a minimum reliability level that is too low.

As specified in the Draft Report for Voting, the reliability of **NQF# 0330: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Heart Failure (HF) Hospitalization** was tested as follows:

> The developer's approach to assessing score-level reliability was to consider the extent to which assessments of a hospital using different but randomly-selected subsets of patients produce similar measures of hospital performance. The developers refer to this as a "test-retest" approach; it may also be called a "split-half" method. A total of 1,210,454 admissions over a 3-year period were examined, with 604,022 in one sample and 606,432 in the other randomly-selected sample. Two risk-standardized readmission rates (RSRR) were calculated for each hospital: one from each of the two separate samples. The agreement between the two RSRRs for each hospital (as measured by an intra-class correlation coefficient (ICC)) was 0.58.

As specified in the Final Report for Voting, the reliability of **NQF #1891: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Chronic Obstructive Pulmonary Disease (COPD) Hospitalization** was tested as follows:

> The developer's approach to assessing score-level reliability was to consider the extent to which assessments of a hospital using different but randomly-selected subsets of patients produce similar measures of hospital performance. The developers refer to this as a "test-retest" approach; it may also be called a "split-half" method. This is generally considered to be an appropriate method of testing reliability. A total of 925,315 admissions over a 3-year period were examined, with 461,505 in one sample and 463,810 in the other randomly-selected sample. Two risk-standardized readmission rates (RSRR) were calculated for each hospital: one from each of the two separate samples. The agreement between the two RSRRs for each hospital (as measured by an intra-class correlation coefficient (ICC)) was 0.48.

As specified in the Draft Report for Voting, the reliability of **NQF #2881: Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI)** was tested as follows:

> The developer's approach to assessing score-level reliability was to consider the extent to which assessments of a hospital using different but randomly-selected subsets of

> patients produce similar measures of hospital performance. The developers refer to this as a “test-retest” approach; it may also be called a “split-half” method. For test-retest reliability, the developer calculated the EDAC for each hospital using first the development sample, then the validation sample. Thus, each hospital twice was measured twice, each time using an entirely distinct set of patients. The developer states that the extent to which the calculated measures of these two subsets agree is evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement, the developer calculated the intra-class correlation coefficient (ICC) as defined by ICC[2,1] by Shrout and Fleiss (1979) and assessed the values according to conventional standards (Landis and Koch, 1977). A total of 496,716 admissions were examined, with 248,358 in each sample. The agreement between the two EDAC values for each hospital (as measured by an intra-class correlation coefficient (ICC)) was 0.54.

In response to AHS’ previous comments on measure #0330 the developer noted:

> We used the Inter-Class Correlation (ICC) method to establish the reliability of the measure score. Our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, then measured again using a second random subset exclusive of the first, and finally comparing the agreement between the two resulting performance measures across hospitals (Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002;21:3431-3446.). This is a purposefully conservative approach to assessing reliability and traditional thresholds for acceptability do not apply to interpreting these results. The minimally acceptable threshold noted by AHS is not appropriate for this particular analytic approach. We have cited the more appropriate convention, which describes the ICC values as moderate (0.41-0.60) for this measure (Landis JR and Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977; 33:159-174).

AHS wishes to highlight that Subcriterion 2a2 specifically requires that the results of reliability testing demonstrate that measures can reproduce “the same results a high proportion of the time when assessed in the same population in the same time period.” We find that ICC results of 0.58, 0.48, or 0.54 do not demonstrate a level or reliability or repeatability that can be accurately described as producing the “same results a high proportion of them time.” For this reason, it is our view that Measures #0330, #1891, and #2881 should not have passed Criterion 2.

According to Rousson et al., in the paper cited by the developer as informing its approach to reliability testing, “a good reliability is attained if the lower bound of the 95 per cent confidence interval is at least 0.75.”

Adams, in the previously referenced RAND report, notes that "[p]sychometricians use a rule of thumb of 90 percent for drawing conclusions about individuals [but] lower levels (70-80 percent) are considered acceptable for drawing conclusions about groups."

The National Research Council's Committee on Performance of Military Personnel has reported that for personnel performance measures "[a]ccepted standards in the field are vague and depend on the characteristic being measured: generally speaking, reliabilities of .6 to .7 are considered marginal, .7 to .8 acceptable, .8 to .9, very good, and above .9 excellent." According to Thompson et al., 0.70 is "a commonly used benchmark for acceptable reliability, […] for group-level comparisons"

Shih and Dimick note that "[a] commonly used cutoff for acceptable reliability when comparing performance of groups is 0.7."

Furthermore, "the more appropriate convention" cited by the developer was described, in the same paper, by Landis and Koch as "clearly arbitrary." Even taken at face value, the Landis and Koch benchmarks describe reliability kappas of 0.41-0.60 as "Moderate" in terms of "Strength of Agreement." AHS believes that "Moderate" reliability does not align with NQF's criteria.

We think it is clear that the reliability testing results for measures #0330, #1891, and #2881 do not demonstrate that the measures scores are "repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period."

Measure #0330's tested ICC score of 0.58 suggests that only 58 percent of the variation in hospital performance is due to true differences in quality (signal) while 42 percent of the variation is due to measure error (noise).

Measure #1891's tested ICC score of 0.48 suggests that only 48 percent of the variation in hospital performance is due to true differences in quality (signal) while 52 percent of the variation is due to measure error (noise). **AHS wishes to highlight that this reliability score would seem to indicate that this measure does not produce the same results a majority of the time, let alone a high proportion of time.**

Measure #2881's tested ICC score of 0.54 suggests that only 54 percent of the variation in hospital performance is due to true differences in quality (signal) while 48 percent of the variation is due to measure error (noise).

We believe that there may be some confusion about reliability due to a lack of guidance from NQF as to what testing results specifically demonstrate sufficient reliability. AHS believes that the Patient Safety Standing Committee may have been highlighting a similar issue when it referenced, in its most recent report, concerns "about insufficient guidance on how to assess measure reliability and validity."

Procedural Error — Voting

It is the view of AHS that the following measures did not achieve consensus during the NQF Member Voting step of the NQF Consensus Development Process (CDP). According to the memo that asked the Executive Committee to ratify the CSAC's recommendation to endorse all 16 measures of the All-Cause Admissions and Readmissions Project 2015-2017, only "[o]ne of the recommended measures was approved, with 67 percent approval or higher by the councils." There was no discussion about why the CSAC chose to recommend all 16 measures despite the fact that only one of the measures achieved greater than 67 percent approval of NQF members. Highlighted below are five substantially impactful measures that did not achieve a simple majority approval rate among NQF members.

- NQF# 0330: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Heart Failure (HF) Hospitalization
    - Approval Rate = 45 Percent
- NQF #0506: Hospital 30-day, All-cause, Risk-Standardized Readmission Rate (RSRR) Following Pneumonia Hospitalization
    - Approval Rate = 50 Percent
- NQF #1789: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)
    - Approval Rate = 50 Percent
- NQF #1891: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Chronic Obstructive Pulmonary Disease (COPD) Hospitalization
    - Approval Rate = 45 Percent
- NQF #2881: Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI)
    - Approval Rate = 46 Percent

AHS believes that a member voting approval rate of 60 percent or less is insufficient for NQF endorsement. We think it is also worth pointing out that for all five of the above measures only three out of the eight measure councils had more than two members cast votes. AHS questions how this can be acceptable for endorsement. We find it alarming that a measure can receive NQF endorsement despite not achieving a majority approval rate among NQF members. It is our opinion that the NQF's status as the "gold standard" of quality measurement and as a consensus standard body, as defined by the Office of Management and Budget, could be in serious jeopardy if this trend persists.

New Information or Evidence — Social Risk Factors

AHS believes that two pieces of new information have become available since the CSAC made its endorsement decision that are reasonably likely to have affected the outcome of the original endorsement decision.

The first item was a December 2016 report published by the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (ASPE) titled

"Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs." The report included the following findings regarding the Hospital

Readmissions Reduction Program:

> Dually-enrolled beneficiaries had significantly greater odds of readmission than non-dually-enrolled beneficiaries within hospitals, an effect that was relatively similar across hospitals.
>
> There was also a significant hospital effect, suggesting that safety-net hospitals have other unmeasured differences in beneficiary characteristics, provide poorer-quality care to prevent readmissions, or face other barriers that might be related to the availability of resources or community supports.

In addition, the report recommended that:

> readmission rates stratified by social risk should be developed and considered for hospital preview reports and public reporting in places such as Hospital Compare, so that hospitals, health systems, policymakers, and consumers can see and address important disparities in care.

The second item was a New England Journal of Medicine (NEJM) article titled "Should Medicare Value-Based Purchasing Take Social Risk into Account?" that was published on December 28, 2016.

This article noted that:

> beneficiaries with social risk factors had worse outcomes on many quality measures, regardless of the providers they saw, and dual enrollment status was the most powerful predictor of poor outcomes.

In addition, the article highlighted that "providers that disproportionately served beneficiaries with social risk factors tended to have worse performance on quality measures."

The article also recommended that "we should measure and report quality of care for beneficiaries with social risk factors."

AHS believes that the HHS ASPE report and NEJM article highlight what the NQF's Readmission Committee stressed as "the high risk of unintended consequences related to adjustment of these measures for SDS factors and the need to reevaluate these measures as the field continues to move forwards."

It is our view that these reports represent advancements in the field that the committee suggested would necessitate reevaluation. Therefore, endorsement of these measures should be withheld until they have demonstrated sufficient risk adjustment and/or stratification for social risk factors.

In conclusion, AHS believes that the endorsement of measures #0330, #0506, #1789, #1789,

#1891 and #2881 should be withdrawn due to the procedural errors and new information cited above.

Sincerely, Richard E. Morrison Adventist Health System Rich.Morrison@ahss.org 407-357-2377

References

Adams, J. L. 2009. *The Reliability of Provider Profiling: A Tutorial*. Santa Monica, CA: RAND Corporation.

All-Cause Admission and Readmissions 2015-2017: Draft Report for Voting [monograph on the Internet] Washington, D.C.: National Quality Forum; 2016. Oct., [cited 2017 Jan 19]. Available from: Consensus Development Process [Web page on the Internet] Washington, DC: National Quality Forum; c2016 [cited 2017 Jan 19]. Available from:www.qualityforum.org/Measuring_Performance/Consensus_Development_Process.aspx

Joynt, K.E., Lew, N.D., Sheingold, S.H., Conway, P.H., Goodrich, K., and Epstein, A. M. 2016 Dec 28. "Should Medicare Value-Based Purchasing Take Social Risk into Account?" *New England Journal of Medicine* [Epub ahead of print].

Landis, J.R. and Koch, G.C. 1977. "The Measurement of Observer Agreement for Categorical Data." Biometrics 33:159-174.

Measure Evaluation Criteria [Web page on the Internet] Washington, DC: National Quality Forum; c2016 [cited 2017 Jan 19]. Available from: http://www.qualityforum.org/Measuring_Performance/Submitting_Standards/Measure_Evalu a tion_Criteria.aspx

Patient Safety 2016: Draft Report for Member Vote [monograph on the Internet] Washington, D.C.: National Quality Forum; 2016. Nov., [cited 2017 Jan 19]. Available from: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=83819

Shih, T. and Dimick, J. 2014. "Reliability of Readmission Rates as a Hospital Quality Measure in Cardiac Surgery." *Ann Thorac Surg* 97 (4): 1214-1218.

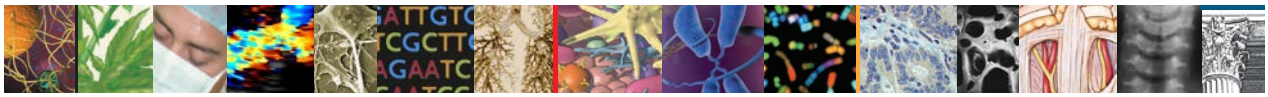
Thompson, M. P. 2016. "Reliability of 30-Day Readmission Measures Used in the Hospital Readmission Reduction Program." *Health Services Research Journal* 51 (6): 2095-2114.

Memo to Consensus Standards Approval Committee Re: All-Cause Admissions and Readmissions 2015-2017 Project [memo on the Internet] Washington, D.C.: National Quality Forum; 2016. Nov., [cited 2017 Jan 19]. Available from: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=83843 Adventist Health System — NQF Readmission Measures Endorsment Appeal Page 10 January 20, 2017

Memo to Executive Committee Re: Ratification of Measures for the All-Cause Admissions and Readmissions Project 2015-2017. [memo on the Internet] Washington, D.C.: National Quality Forum; 2016. Nov., [cited 2017 Jan 19]. Available from: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=84046

Wigdor, A.K. and Green, B. F. Editors; Committee on the Performance of Military Personnel, National Research Council. 1991. *Performance Assessment for the Workplace, Volume I.* Washington, D.C. National Academy Press.

U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation. 2016. Report to Congress: *Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs*. Washington, D.C.



The NEW ENGLAND JOURNAL of MEDICINE

Perspective

Should Medicare Value-Based Purchasing Take Social Risk into Account?

Karen E. Joynt, M.D., M.P.H., Nancy De Lew, M.A., Steven H. Sheingold, Ph.D., Patrick H. Conway, M.D., Kate Goodrich, M.D., and Arnold M. Epstein, M.D.

The United States is rapidly moving to a health care delivery system in which value-based payment models are the predominant way of reimbursing clinicians for care. Since caring for patients with social risk factors may cost more and make it harder to achieve high performance on quality metrics, there is long-standing concern about how these patients might fare under such systems and how the systems might affect providers who disproportionately provide care to socially at-risk populations.

In October 2014, Congress passed the Improving Medicare Post-Acute Care Transformation (IMPACT) Act, which required the Office of the Assistant Secretary for Planning and Evaluation (ASPE) of the Department of Health and Human Services to review the evidence linking social risk factors with performance under existing federal payment systems — and to suggest strategies to remedy any deficits they found. That report was sent to Congress in December 2016.[1]

Because the report focuses primarily on Medicare, the analyses centered on social risk factors covered in current Medicare data, including dual enrollment in Medicare and Medicaid as a marker for low income, residence in a low-income area, race, Hispanic ethnicity, and residence in a rural area. Disability was also examined. Medicare payment programs were analyzed if they were currently operational or defined in statute and if they incorporated quality or efficiency metrics into payment decisions. These criteria led to the inclusion of nine programs: the Hospital Readmissions Reduction Program, Hospital Value-Based Purchasing Program, Hospital-Acquired Condition Reduction Program, Medicare Advantage Quality Star Rating Program, Medicare Shared Savings Program (MSSP), Physician Value-Based Payment Modifier Program, End-Stage Renal Disease Quality Incentive Program, Skilled Nursing Facility Value-Based Purchasing Program, and Home Health Value-Based Purchasing Program.

There were two main findings. First, beneficiaries with social risk factors had worse outcomes on many quality measures, regardless of the providers they saw, and dual enrollment status was the most powerful predictor of poor outcomes. Dually enrolled beneficiaries had poorer outcomes on process measures (e.g., cancer screening), clinical outcome measures (e.g., diabetes control, readmissions), safety (e.g., infection

rates), and patient-experience measures (e.g., communication from doctors and nurses), as well as higher resource use (e.g., higher spending per hospital admission episode). These associations held even when the beneficiaries being compared were in the same hospital, health plan, accountable care organization (ACO), physician group, or facility. These findings generally persisted after risk adjustment, across care settings, measure types, and programs and were moderate in size.

Second, in every type of care setting examined, providers that disproportionately served beneficiaries with social risk factors tended to have worse performance on quality measures. Some of the performance differences were driven by beneficiary mix, but part of the difference persisted even after adjustment for beneficiary characteristics. As a result, safety-net providers were more likely to face financial penalties in most of the value-based purchasing programs in which penalties are currently assessed, though some of the differences were small because of the methods applied in calculating such penalties. The single exception was that ACOs with a high proportion of dually enrolled beneficiaries were more likely to share in savings under the MSSP, despite slightly worse quality scores, because their cost performance was better.

However, in every setting, there were some providers serving a high proportion of beneficiaries with social risk factors that achieved high performance levels — indicating that high performance is feasible with the right strategies and supports.

These findings underscore several challenges. How do we ensure that we can monitor quality of care for different social groups? How do we judge performance fairly across providers that serve beneficiaries with a different mix of social risk factors? And how do we ensure that payment reflects the resources required to provide high-quality care while also providing incentives to ameliorate existing disparities in care?

We suggest three general strategies (see table). The first strategy is foundational: we should measure and report quality of care for beneficiaries with social risk factors. For that to happen, data collection will need to be enhanced and statistical techniques developed to allow measurement and reporting of performance on key quality and resource-use measures for such subgroups.

Another important component of this strategy is to measure equity itself. Health equity measures or domains should be developed and introduced into existing payment programs to measure disparities and provide incentives for reducing them. The final component of this strategy is to monitor the financial impact of Medicare payment programs on providers that disproportionately serve beneficiaries with social risk factors. For example, as the Merit-Based Incentive Payment System is implemented, it will be important to ensure that providers caring for large numbers of socially at-risk beneficiaries are not themselves put at risk.

The second strategy is to set and maintain high, fair quality standards for the care of all beneficiaries. That does not mean that all measures should be adjusted for social risk, nor that no measures should be so adjusted. Rather, measures should be individually examined to determine whether adjustment for social risk factors is appropriate to make them as equitable as possible. This determination will depend on the measure and its empirical relationship to social risk factors. For example, for process measures, an important concern is whether the process being measured is entirely under the provider's control. Some observers argue that adherence to annual mammography or periodic colonoscopy, for instance, is influenced not only by provider recommendations but also by patient preferences and other related factors. By contrast, provision of aspirin to patients with acute myocardial infarction is more directly under the control of hospital-based providers.

For outcome measures, determining whether or not a measure should be adjusted may depend on the pathway by which the social risk factor is related to worse outcomes. For example, dual enrollment status may be associated with a higher risk of frailty, worse functional status, and lower levels of social support and education, all of which may affect readmission rates, diabetes control, and other outcome measures. Such associations might make adjustment more appropriate. In addition, research should be conducted to determine whether better ascertainment of these unmeasured medical and social factors and their use in statistical adjustment might improve the ability to delineate true differences in performance between providers.

The third strategy recognizes that regardless of whether or not measures are adjusted for social risk, we need to make strides in addressing the underlying issues themselves, and we can leverage

Strategies for Monitoring Quality of Care for Socially At-Risk Beneficiaries and Providing Incentives to Reduce Disparities in Care.

Strategy	Considerations
Measure and report quality of care for beneficiaries with social risk factors	Pursue reporting for care of beneficiaries with social risk factors Develop health equity measures Prospectively monitor program impact on providers disproportionately serving beneficiaries with social risk factors
Set high, fair quality standards for care of all beneficiaries	Consider measures for adjustment on a case-by-case basis Improve risk adjustment for health status in program measures
Reward and support better outcomes for beneficiaries with social risk factors	Provide payment adjustments to reward achievement or improvement of outcomes in beneficiaries with social risk factors Use existing or new quality-improvement programs to support providers that serve such beneficiaries Encourage demonstration projects and models focusing on such beneficiaries Conduct research on the costs of caring for such beneficiaries

value-based payment programs to do so. Therefore, strategy 3 focuses on directly rewarding and supporting better outcomes for socially at-risk beneficiaries. First, whereas value-based purchasing programs reward achievement of high quality and good outcomes among all beneficiaries, we should also consider creating additional targeted financial incentives to reward achievement or improvement specifically for socially at-risk beneficiaries. Such targeted incentives could help harness the power of value-based payment to improve care and outcomes for our most vulnerable patients, and simultaneously offset any real or perceived disincentives under value-based purchasing programs to caring for these beneficiaries.

Second, we should use existing or new quality-improvement programs to provide targeted technical assistance to providers that serve beneficiaries with social risk factors, recognizing that they may face unique challenges in both participating and succeeding in new payment models.

Third, we should develop demonstrations or models focusing on care innovations that may help achieve better outcomes for beneficiaries with social risk factors but that might not be testable under current payment and delivery structures. Examples include the demonstration programs in Medicare Advantage that focus on coordinating benefits between Medicare and Medicaid and the Center for Medicare and Medicaid Innovation's Accountable Health Communities model.

Fourth, we should pursue further research to examine the costs of achieving good outcomes for beneficiaries with social risk factors and to determine whether current payments adequately account for any differences in care needs. Disproportionate Share Hospital payments are one current example of such add-on payments for social risk, and payments to Medicare Advantage contracts are higher for dually eligible beneficiaries. However, these payment adjustments are not uniform across care settings.

Social factors are powerful determinants of health. Beneficiaries with social risk factors may have poorer outcomes because of higher levels of medical risk, worse living environments, greater challenges in adherence and lifestyle, and bias or discrimination. Providers serving these beneficiaries may have poorer performance due to fewer resources, more challenging clinical workloads, lower levels of community support, or worse quality. These problems are complex and will not yield to simple fixes.

However, the strategies we propose may be an important starting point. As the scope, reach, and financial risk associated with value-based payment models grow, Medicare can use the strategies outlined above to administer fair, balanced programs that promote quality and value, provide incentives to reduce disparities, and avoid inappropriately penalizing providers that serve socially at-risk beneficiaries, ultimately helping to ensure that the best health outcomes possible can be achieved for all beneficiaries.

Disclosure forms provided by the authors are available at NEJM.org.

From the Harvard T.H. Chan School of Public Health and Brigham and Women's Hospital, Boston (K.E.J., A.M.E.); the Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, Washington, DC (K.E.J., N.D.L., S.H.S.); and the Centers for Medi-

care and Medicaid Services, Baltimore (P.H.C., K.G.).

This article was published on December 28, 2016, at NEJM.org.

1. Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation. Report to Congress: social risk factors and performance under Medicare's value-based payment programs (https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs).

DOI: 10.1056/NEJMp1616278

Appendix B: MEASURE DEVELOPER RESPONSE TO THE APPEAL

We thank NQF for the opportunity to respond to the recent appeal by Adventist Health System regarding the endorsement of measures #0330, #0506, #1789, #1891, and #2881 in the All-Cause Admissions and Readmissions project:

- NQF #0330: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Heart Failure (HF) Hospitalization
- NQF #0506: Hospital 30-day, All-cause, Risk-Standardized Readmission Rate (RSRR) Following Pneumonia Hospitalization
- NQF #1789: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)
- NQF #1891: Hospital 30-day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Chronic Obstructive Pulmonary Disease (COPD) Hospitalization
- NQF #2881: Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI)

The appeal is based on the claim of two procedural errors and the availability of new information or evidence. We address only two of the issues raised in this response below.

I. Measure Reliability

The appellant asserts that "we believe that the recommendation and subsequent endorsement of several of these measures was inconsistent with NQF's Scientific Acceptability criterion for reliability."

To support the appeal, several sources regarding reliability and standards for approaches to interpreting statistics measuring reliability are cited. Our position is that the measures under appeal meet this subcriterion, and that no procedural error occurred. We offer three points in a brief rebuttal below and a detailed explanation of our rationale on pages 4-7.

First, in their critique of the results CMS presented for measure score reliability, the appellant cites NQF's guidance on interpretation of **data element reliability**. Because these measures are calculated from claims submitted by hospitals and other providers, adjudicated by CMS, and stored electronically, the reliability of the data is extremely high. When the measures are computed on the same set of admissions, for the same providers, using the same time period, precisely the same results are obtained. That is, these are deterministic measures, reproducible by any third party, and thus demonstrably meet the standard described by NQF under item 2a2. We maintain that the NQF's measure submission forms offer no guidance on the interpretation test of measure score reliability, including the test used by CMS, the intraclass correlation coefficient, ICC[2,1]. This is a test-re-test method and NQF's guidance on interpretation of data element reliability does not apply.

Second, the appellant cites several sources that use signal-to-noise ratios to evaluate provider measures. The appellant suggests that these are suitable approaches to assessing measures

score reliability, and that the critiqued measures don't meet the standards for signal-to-noise reliability. We maintain that signal-to-noise ratio is useful for some purposes, but **signal-to-noise ratio is a provider level metric**, which assesses reliability separately for each provider's measure score. This metric is then typically averaged across all providers to create a measure reliability score. This is not consistent with standard approaches to evaluating measure reliabilities. Moreover, because signal-to-noise is the ratio of between unit variation (signal) to total between unit plus within unit variation (precision), a measure can be very imprecise at the unit level and still have a high signal-to-noise ratio, if there is large between unit variation. Conversely, a measure can be extremely precise for each unit, but have very low signal-to-noise reliability, if there is no between unit variation. For this reason, **signal-to-noise ratio is not consistent with the reliability metric we report, ICC[2,1].** In the details at the end of this memo we report on a simulated dataset with high signal-to-noise reliability and low ICC[2,1]. Thus, we maintain that the standards that are referenced for signal-to-noise ratio do not apply to ICC[2,1].

Third, since, as noted above, there is no NQF guidance for standards of test-retest reliability, and the standards cited for signal-to-noise ratio do not apply, **other guidelines or reference values for ICC[2,1] should be used.** In the absence of empirically supported standards, our position is that 'acceptability' depends on context. For simple concepts or constructs, such as a patient's weight, the expectation is that the test-retest reliability of a measure of that construct should be quite high. However, for complex constructs, such as clinical severity, patient comorbidity, or symptom profiles used to identify a condition or clinical state, reliability of measures used to define these constructs is quite a bit lower. In this memo we offer several examples of the reliability of measures of complex constructs using the ICC[2,1]. These examples provide the necessary context for interpreting the acceptability of ICC[2,1] values in the ranges found for the readmission measures. **These empirical findings indicate that our reported ICC[2,1] values are consistent with those in similar contexts**.

II. New Publications Related to the Use of SES in Measure Risk-Adjustment Models

We have reviewed the two recent studies mentioned in the appeals' letter. Both the ASPE report[1] and the NEJM article[2] address the importance of social factors in quality measurement and pay-for-performance programs. We have long acknowledged and agree with the conclusion of both studies that socially disadvantaged groups, such as those earning a low-income, members of some racial or ethnic minority groups or those living with a disability, are at greater risk of poor health and health outcomes. However, we disagree that either study provided new

[1] Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (ASPE), "Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs." December 2016, https:/ /aspe.hhs.gov/ pdf-report/ report-congress-social-risk-factors-and-performance-under- medicares-value-based-purchasing-programs

[2] Joynt, Karen E., De Lew, Nancy, Sheingold, Steven H., Conway, Patrick H., Goodrich, Kate, and Epstein, Arnold M. (2017) "Should Medicare Value-Based Purchasing Take Social Risk into Account?" New England Journal of Medicine. http://www.nejm.org/doi/full/10.1056/NEJMp1616278#t=article

evidence that was meaningfully different than the evidence available to the committee during their deliberations on these measures.

The ASPE report and the NEJM article restate the NQFs recommendation that measures should be examined *individually* to determine if adjustment for social factors is appropriate (ASPE 2016: 15; NEJM 2017: 2). When determining whether adjustment is warranted, developers were instructed to consider the conceptual relationship between the SDS factor and the outcome as well as the empirical relationship. Although we found that both observed and adjusted readmission rates are higher on average for hospitals serving a large proportion of patients who were dual eligible and those living in a census block group with low AHRQ SES Index, we have also shown that many hospitals serving a high share of socially disadvantaged patients achieve high performance scores on the readmission measures (see, e.g., Bernheim et al. 2016. "Accounting for Patients' Socioeconomic Status Does Not Change Hospital Readmission Rates." Health Affairs 35(8): 1461-1470). The authors of the NEJM article also found that risk adjusting for indicators of SES or of race does not explain away performance differences between hospitals serving low- and high-proportions of beneficiaries with these indicators, which aligns with our findings presented to the Committee. Neither publication offered new relevant information that was not available to the Committee during their deliberations.

The ASPE report does not recommend risk-adjustment of readmission measures with SES risk variables. However, the report does recommend consideration of stratifying hospitals into peer groups after measure calculation for the purpose of payment calculation *rather than* adjusting measures at the patient level: "Hospitals would be judged only against their peers, and penalties would be assessed based on the average performance within each group rather than the average performance overall" (ASPE report 2016: 82). The recent 21st century CURES laws align with this recommendation and direct CMS to stratify hospitals for the purpose of determining the payment adjustment factor within the Hospital Readmission program (HRRP). This represents a change to the use of the measure within a pay-for-performance program but not a change to the measure itself.

We agree with the appellants comment that patient-level stratification of readmission rates (in contrast to stratifying hospitals into peer groups) could serve to illuminate disparities within hospitals of quality of care for beneficiaries with social risk factors. However, the NQFs guidance to measure developers for the SDS trial period was to present stratified results to the committee only for measures that included SES indicators in the measure risk model. Therefore, we did not submit stratified measure results.

We agree with the appellant's comment, the ASPE report, and the NEJM article recommendations to measure and monitor quality of care for vulnerable populations, but adding patient-level risk-adjustment to the readmission measures is not a means to do so. The rationale behind the development of equity measures is to illuminate disparities and create incentives to reduce them, improve care for vulnerable populations, and promote greater transparency for consumer choice. An example of such an initiative is the graphical tool

"Mapping Medicare Disparities" provided by the Office of Minority Health (OMH), which identifies geographical areas of disparities between subgroups of Medicare beneficiaries (e.g., dual eligible vs. non-dual eligible beneficiaries) on its webpage (https://www.cms.gov/about-cms/agency-information/omh/OMH-Mapping-Medicare-Disparities.html). CMS supports these and other initiatives to highlight disparities and promote greater equity in health care delivery and patient outcomes. CMS remains committed to developing alternative ways to measure and report disparities and to promote equity in care and outcomes among beneficiaries.

Additional Details on Measure Reliability

The appellant cites the NQF subcriterion 2a2, which states the criterion for reliability:

> *"Reliability testing demonstrates that the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise." (emphasis added).*

Notably, **this subcriterion has a footnote**:

> *Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).*

Many concerns about reliability of measures and measure attributes arise because of the multiple definitions of reliability and the multiple standards available in the literature. In this footnote to 2a2 we see a long list of somewhat exclusive types of reliability listed. Here we discuss three metrics of reliability that are relevant.

COMPUTED SCORE RELIABILITY

The appellant claims that the measures reported do not meet the standards of Subcriterion 2a2, which specifically requires that "*measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period*", a standard which according to the footnote applies to the *data elements and the computed measure score*. We will refer to this as the computed measure score reliability. This reliability can be low for measures that rely for instance on surveys (where respondents can be inconsistent with responses), data abstraction (which can introduce errors) or collecting new clinical data (which has measurement error), but is typically high for measures that rely on existing claims data. For the measures appealed, all data used to calculate the measures are derived from adjudicated and finalized Medicare claims, which are submitted and stored electronically. The reliability of such data is extremely and uniformly high. And, given the dataset of data elements of demonstrated reliability, when the measures are computed on the same set of admissions, for the same providers, using the same time period, precisely the same results are obtained. That is, these are deterministic measures, reproducible by any third party,

and demonstrably meet the standard of 2a2; they produce exactly the same results nearly 100% of the time.

SIGNAL-TO-NOISE RELIABILITY

The appellant memo cites several sources that use signal-to-noise ratios to evaluate provider measures. Signal-to-noise is a type of reliability, but it is distinct from both computed score and test-retest reliability. *This notion of reliability is not related to test-retest reliability*. It is listed in the NQF Subsection 2a2 footnote above as an example of precision, but strictly speaking, it is not a measure of precision. Rather, measures of signal-to-noise (there are several) reflect the ratio of between unit variation (signal) to total variation (between unit plus within unit, where the within unit variation reflects precision). As noted earlier, a measure can be very imprecise at the unit level and still have a high signal to noise ratio, if there is large between unit variation; conversely, a measure can be extremely precise at the unit level but have low signal-to-noise ratio, if there is no between unit variation. Moreover, it is unit level metric, calculated separately for each provider, typically averaged to create a 'measure' reliability. For both of these reasons, we do not think it is an appropriate measure of *measure* reliability; instead we use test-retest reliability.

Because the signal-to-noise ratio measure does not equate to the test-retest reliability measure, the same conventional thresholds do not apply to both. Thompson et al, cited by the appellant, uses 0.7 as a threshold, and justifies this with references to other authors who used it; there seems to be no empiric justification, as we provide below for test-retest reliability. To demonstrate the distinction between this approach and test-retest, we simulated a dataset (available) to demonstrate the difference between ICC[2,1] and Signal-to-Noise ratio. In this simulated dataset, which includes 100 hospitals with mean rate of 20%, and an average of 50 patients per hospital we found ICC[2,1] = 0.20 and the average Signal/Noise ratio = 0.73. This example demonstrates explicitly that the signal/noise ratio is distinct from ICC[2,1], and that the papers, standards and reports referenced by the appellant do not apply.

TEST-RETEST RELIABILITY

The measure of test-retest reliability used to assess the measures is a specific statistic known as ICC[2,1], which is analogous to the more familiar but appropriate for continuous measures. It compares two repeated measures on each provider for agreement; it is a conservative measure of test-retest reliability, because it assumes that the multiple measurements are drawn from a larger sample of tests, and that the measured providers are drawn from a larger sample of providers. This reliability does not refer to the reliability of the data elements or the precision of the estimates, the two criteria mentioned in 2a2, but rather the reliability of the risk-adjusted measure score. Note that ICC[2,1] is also distinct from the conventional "intra-class correlation", which the ratio of between unit variation to the total variation.

The appellant then references the ICC[2,1] values reported for the challenged measures. Note that these are reported as additional reliability testing, per the footnote to 2a2. No standards

are given for the types of reliability listed in the footnote. In particular, ICC[2,1] evaluates the reliability of the measure with respect to different data samples (split samples which include data from separate groups of patient admissions). Moreover, guidelines for the specific ICC[2,1] statistic are of limited availability. The appellant cites only a single source for evaluating ICC[2,1], Rousson et al, who however simply cite Lee et al; Lee et al in turn reference Burdock et al without comment. However, Burdock et al mention 0.75 without any justification, and for a different statistic:

> *"R=ut2/(ue2 + ut2) is based on the assumptions that the observers are fixed and that there is no interaction between observer and subject. Apart from considerations of the other components of the model, a minimum requirement of the instrument is that R be large, meaning that is should be as close to unity as possible. A high intraclass correlation coefficient, e.g. R ≥ .75".*

Note that Burdock et al provides no empiric justification, and moreover, are discussing a reliability metric that is not ICC[2,1], but something more similar to a signal-to-noise ratio.

If there is no evidence to support the 0.75 value, the question remains how to best determine what is an 'acceptable' level of inter-rater reliability. Some may still use the Landis & Koch (Landis, Koch 1977) convention to argue that CMS hospital measures have poor reliability, that something in the 0.61-0.80 range might be more appropriate ("substantial"), which coincides with a common instinct to think of 60% as "passing" and 80% as "above average." However, conventions are by definition flexible; to quote Landis & Koch, which NQF has mentioned as a guideline:

> *In order to maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels will be assigned to the corresponding ranges of kappa … Although these divisions are clearly arbitrary, they do provide useful "benchmarks" for the discussion of the specific example in Table 1*

Thus, even these guidelines, which have been widely adopted, were originally stated as arbitrary. Their usefulness has derived largely from their consistency with findings in a very large range of research fields over the four decades since their original publication. However, this does not make them final standards of 'acceptability'.

Therefore, our position is that 'acceptability' depends on context. For example, if we were measuring adolescent weight twice with the same scale, and assessing whether the weights were above a certain threshold, we would expect the two measurements to agree almost exactly (ICC[2,1] ~ 1); otherwise, we would discard the scale. At the other extreme, if we were measuring a latent personality trait such as a personality disorder, we would expect a much lower level of agreement. In fact, Nestadt et al assessed ICCs for several standard tools for assessing personality disorder and found test-retest reliabilities in the range of 0.06-0.27 (Nestadt 2012). (Notably, Nestadt et al conclude that these tools "may still be useful for identifying [personality disorder] constructs.")

Thus, we would argue that **one should adopt for 'acceptable' level of ICC[2,1] a standard that is consistent with that in known, familiar, and related contexts**. The current context is measuring provider quality, or specifically provider propensity to provide appropriate care as measured by subsequent outcomes. We identified several studies, which we think support the Landis & Koch guidelines when assessing test-retest reliability in the context of hospital measurement.

- Hall et al calculated test-retest reliability for determining comorbidities from chart abstraction [Hall et al]. In this study, multiple abstracters abstracted the same charts and the results were used to calculate four different common comorbidity scores. For three of the indices, test-retest reliabilities ranged from 0.59-0.68, with the fourth (the Charlson comorbidity score) achieving 0.80. We would argue that chart abstraction, with test-retest reliabilities in the 'moderate' to 'substantial' range, should be inherently more reliable than measuring hospital quality.
- Cruz et al report reliabilities for collecting risk factor information from patients presenting to an emergency department with potential acute coronary syndrome (ACS) [Cruz et al]. Each patient was queried twice, once by a clinician and once by a trained research assistant, and the reliabilities for a range of risk factors were calculated; these ranged from 0.28 (associated symptoms) to 0.69 (cardiac risk factors), with all other factors in the 0.30-0.56 range.
- Hand et al report test-retest reliabilities for bedside clinical assessment of suspected stroke [Hand et al]. Pairs of observers independently assessed suspected stroke patients; findings were recorded on a standard form to promote consistency. The reliabilities were calculated for the full range of diagnostic factors: for vascular factors reliabilities ranged from 0.47-0.69 with only four of eight above 0.6; for history they ranged from 0.37-0.65 with only five of 12 above 0.6; other categories were similar (though reliability=1 for whether the patients were conscious).

These contexts are intuitively similar to that of measuring hospital quality, and moreover suggest that the guidelines of Landis & Koch are appropriate for areas of clinical care.

SUMMARY

The appealed measures do meet the standard of high computed score reliability specified in NQF guideline section 2a2. Signal-to-noise reliability, while useful, is not a metric of scale reliability, and is distinct from test-retest reliability, and any conventional thresholds do not necessarily apply to ICC[2,1]. Accepted standards for ICC[2,1] are not available, but an examination of test-retest reliability in contexts that are intuitively similar to that of provider quality measurement finds values that are consistent with both the alternative guidelines and with CMS measures. Reliability testing in hospital quality measurement should be interpreted in context, and the evidence we present refutes that 0.7 is a minimal acceptable reliability value for test-retest reliability of complex clinical constructs such as symptomatology, health risk factors, comorbidity, or hospital performance on patient outcomes.