NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel – Measure Evaluation Web Meeting

The National Quality Forum (NQF) convened the Scientific Methods Panel (SMP) on March 22-23, 2022, for a discussion of the scientific acceptability (i.e., reliability and validity) of several complex measures submitted during the spring 2022 evaluation cycle. Of the 13 measures that the SMP reviewed during this cycle, 10 measures were discussed during the meeting, including those that: (1) did not receive consensus from the SMP subgroups[1] and for which the SMP subgroups identified major areas of concern during their preliminary evaluations, (2) did not initially pass the SMP's evaluation but for which the measure developers provided additional information, or (3) were pulled for discussion by SMP subgroup members or NQF staff. This meeting summary includes brief summaries of the 10 measures, the overarching methodological issues discussed during the web meeting, and the voting results for the three measures not discussed during the web meeting.

Welcome, Review of Meeting Objectives, Introductions, and Overview of Evaluation and Voting Process

Matthew Pickering, NQF senior director, welcomed the SMP and participants to the web meeting. NQF staff reviewed the meeting objectives. The SMP members each introduced themselves and disclosed any conflicts of interest. Joseph Kunisch disclosed a conflict with NQF #0471e, NQF #0716e, and NQF #3687e; Zhenqiu Lin disclosed a conflict with NQF #2377 and NQF #3687e; Eric Weinhandl disclosed a conflict with NQF #3679, NQF #3696, and NQF #3697; and Patrick Romano and Matt Austin disclosed a conflict with NQF #2820. Joseph Kunisch was recused from NQF #0471e, NQF #0716e, and NQF #3687e due to either collaborating directly with the measure developer, with or without compensation, or working directly as a consultant on the measure. Zhenqiu Lin was recused from NQF #2377 as he is employed by the organization that developed a competing measure not under review during this cycle (NQF #3613e). He was also recused on NQF #3687e as he contributed to the statistical testing for the measure. Eric Weinhandl was recused from NQF #3679, NQF #3696, and NQF #3697 due to directly collaborating with the measure developer, with or without compensation. Patrick Romano and Matt Austin were recused from NQF #2820 because they directly collaborated with the measure developer, with or without compensation. Following the introductions, Dr. Pickering and Hannah Ingber, NQF manager, reviewed the process for the measure discussions and the measure evaluation criteria.

Some SMP members were unable to attend the entire meeting due to early departures and late arrivals. The vote totals reflect members present and eligible to vote. The required attendance to host the meeting was met and maintained for all measures for the entirety of both meetings. With the exception of NQF #2820 and NQF #3687e, attendance of 12 SMP members was required to hold the meeting. These two measures alone had more than one SMP member recusal, further reducing the denominator for required attendance. Therefore, only 11 members were required for the discussion and voting for these two measures. The quorum for voting in the two subgroups for each measure was eight SMP

[1] Subgroups are determined by the NQF SMP team with input from the SMP co-chairs. NQF assigns measures to subgroup members for evaluation based on panelists' relevant expertise, interests, measure-specific disclosures of interest, and Standing Committee membership.

members, which was also maintained for all measures for the entirety of both meetings. The voting results are provided below.

Measure Evaluation

Fifty-three measures were submitted during the spring 2022 cycle; 13 of them were complex measures that the SMP reviewed for scientific acceptability. Measure evaluations that were conducted during the meeting were based on the preliminary analyses performed by assigned SMP members. Each SMP member was assigned to one of two subgroups, and both subgroups were assigned either six or seven of the 13 measures being reviewed this cycle. The subgroup members then performed in-depth reviews and analyses of their assigned measures. The developers received these preliminary analyses prior to the web meeting and were given an opportunity to submit written responses to concerns expressed by the SMP reviewers. These responses were provided to the SMP prior to the meeting to prepare for its discussion and subsequent voting conducted by the subgroups on the measures during the meeting.

During the meeting, the SMP evaluated the reliability and/or validity of 10 measures based on its preliminary analyses and the additional information that the developers submitted for consideration. The remaining three measures that the SMP evaluated during the spring 2022 cycle were not discussed during the meeting because the subgroup members passed them on both reliability and validity, and they were not otherwise pulled for discussion. For these measures, the subgroup's preliminary analyses serve as the final SMP assessment of scientific acceptability for the Standing Committees' consideration. For each measure discussed during the meeting, NQF staff described the measure, noted the preliminary evaluation ratings of the subgroup, and highlighted the criterion (or criteria) for which there was a lack of consensus and/or major areas of concern. Drs. David Nerenz and Christie Teigland, the SMP co-chairs, facilitated the remainder of the discussion. Lead discussants from either subgroup first summarized the primary concerns of the subgroup. Other subgroup members then made additional comments. Following these concerns and comments, NQF staff invited the measure developers to provide brief responses to the concerns raised by the subgroup members and to summarize their written response(s), if provided. Next, the SMP co-chairs invited comments or additional questions from other SMP members, and the developers were invited to respond. The subgroup members then voted on the measure, producing the final votes of the SMP for the relevant criteria. These votes reflect the final assessment of scientific acceptability conducted by the respective SMP subgroups.

A measure passes the scientific acceptability criteria when greater than 60 percent of eligible voting members select a passing vote option (High or Moderate) on both reliability and validity. A measure does not pass the SMP's review when less than 40 percent of voting members select a passing vote option on reliability and validity. The SMP has not reached consensus on the measure if between 40 and 60 percent of eligible voting members select a passing vote option on reliability and validity.

Voting Legend:

- *Scientific Acceptability: Reliability and Validity:* H – High; M – Moderate; L – Low; I – Insufficient; NA – Not Applicable

Subgroup 1

During the meeting, subgroup #1 discussed five measures (i.e., NQF #1460, NQF #0471e, NQF #0716e, NQF #3687e, and NQF #2820). This subgroup re-voted on the reliability and validity of NQF #1460, NQF #0471e, and NQF #0716e. The subgroup also discussed NQF #3687e and NQF #2820 but decided to not re-vote on the measures; therefore, the preliminary analyses decisions stood. The subgroup accepted

the preliminary analysis decision for NQF #2377 without further discussion. The final results for the six measures evaluated by subgroup 1 are presented below.

NQF #1460 Bloodstream Infection in Hemodialysis Outpatients (Centers for Disease Control and Prevention [CDC])

Measure Steward/Developer Representatives at the Meeting

- Andrea Benin
- Jonathan Edwards

SMP Votes

- **Reliability**: Total Votes-9; H-0; M-1; L-3; I-5 (1/9 – 11 percent, No Pass)
- **Validity**: Total Votes-9; H-0; M-1; L-2; I-6 (1/9 – 11 percent, No Pass)

The SMP noted that patient/encounter-level validity testing serves as reliability testing for this measure. The SMP also noted that expanded accountable-entity level empirical testing for a maintenance measure is preferred, but this is not required. There were zero votes for "high" because the highest-possible vote for validity and reliability was "moderate" as the developer only submitted patient/encounter-level validity data.

In its preliminary analyses, the SMP did not pass the measure on reliability and did not reach consensus on validity. The SMP expressed concern with the large amount of missing data. The developer noted that significant data issues emerged due to coronavirus disease 2019 (COVID-19), and the developer validated a targeted sample of underreporting facilities. The SMP acknowledged the approach to the targeted validation but explained that the large amount of missing data threatens the reliability and validity of the measure. The SMP also expressed concern that facility-level variation may be present in underreporting. Kappa values were also missing from the validation results, and the results only included raw agreement data. Overall, the SMP concluded that the testing results could not ensure the validity or reliability of the measure. The SMP re-voted on reliability following the discussion and ultimately did not pass the measure on this criterion.

The SMP also expressed concern that very minimal risk adjustment exists for the measure. The only risk factor included in the risk adjustment model is vascular access. Given the multifactorial nature of the risk of infection, the risk adjustment model lacked face validity. Furthermore, the SMP noted that the developer tested the risk adjustment model compared to no risk adjustment, which is not adequate testing. Overall, both the approach to risk adjustment and the missing data were threats to the validity of the measure. Following the discussion, the SMP re-voted on validity and ultimately did not pass the measure on this criterion.

This measure is not eligible for a revote from the Renal Standing Committee because the SMP determined that inappropriate methods were used. First, the developer did not use a kappa statistic to report their results. Second, there were significant concerns with underreporting (i.e., missing data), which varied between 15-50 percent, ultimately reducing the validity of the measure and its ability to show meaningful differences in performance.

NQF #0471e ePC-02 Cesarean Birth (The Joint Commission)

Measure Steward/Developer Representatives at the Meeting

- Chris Walas
- Stephen Schmaltz
- Elliott Main

SMP Votes

- **Reliability**: Total Votes-9; H-0; M-4; L-3; I-2 (4/9 – 44 percent, Consensus Not Reached)
- **Validity**: Total Votes-9; H-0; M-5; L-2; I-2 (5/9 – 56 percent, Consensus Not Reached)

This is an electronic clinical quality measure (eCQM) of an existing chart-based measure (i.e., NQF #0471). In its preliminary analyses, the SMP did not pass the measure on either reliability or validity. The SMP noted that patient/encounter-level validity testing serves as reliability testing for this measure. There were zero votes for "high" because the highest-possible vote for validity and reliability was "moderate" because the developer only submitted validity testing at the patient/encounter level.

During the SMP's preliminary analyses, concerns were raised regarding the failure to provide kappa results for the required data elements in the validity testing. Given that some sites had 15 percent of denominator exclusions removed due to one of the exclusions and that NQF requires all critical data elements to be validated, this was seen as inadequate. In their response, the measure developer provided an additional table to the SMP that included updated testing with kappa results, which noted that the overall agreement rate was the same as the kappa-adjusted agreement rate. However, the results were seen as implausible because the agreement rates were identical to the kappa results, which raised questions about the validity of these results. Another issue raised pertained to one of the test sites, which had major difficulties reporting the numerator for the measure. This site used the Meditech electronic health record (EHR) system, which is used by a significant portion of the hospital market, thus raising concerns about reliability and validity. In particular, if the data are not captured in structured fields in Meditech, it will be difficult to report the measure. The developer explained that a particular hospital used a stand-alone documentation system for its obstetrics (OB) documentation, and the data were not in discrete fields in Meditech. The developer also noted that the lower kappa values pertained to 10 specific data elements. These issues, from the developer's view, are now resolved. They noted that in another measure, the data elements under concern used in this measure noted excellent agreement in three EHRs (i.e., Epic, Cerner, and Meditech) with 94–98 percent agreement. It was also noted that the developer was able to fix an identified Meditech issue by placing the data elements in discrete fields. However, they were not able to submit this in time for the NQF submission. The SMP re-voted on reliability following the discussion and ultimately did not reach consensus on this criterion.

The SMP members stated that outcome measures without risk adjustment raise concerns for them generally, but the developer did provide a rationale for the lack of risk adjustment. While the developer did attempt to create a homogenous population by using exclusions, there were concerns that the population was not homogeneous enough to perform this method. Some variables such as advanced maternal age, high body mass index (BMI), and others that may be important variables for risk adjustment were not excluded. One SMP member noted that exclusions were being used to identify a low-risk population rather than include actual risk factors. The developer replied that adding these conditions that may impact the risk of cesarean delivery/birth is not what the measure is intended to do, as advised by their technical expert panel (TEP) of clinicians who helped design the measure. The clinical intent is to assess and improve the rate of low-risk cesarean deliveries/births so that it is more appropriate rather than drive down the rate of all cesarean deliveries/births, which may be appropriate for higher-risk pregnancies, for example. The developer discussed several studies that justified the lack of risk adjustment, noting minimal changes in the measure results, both with and without risk adjustment. The developer also noted that adding risk factors would add complexity to the measure and increase the burden on sites for measure reporting; they also noted that research has indicated that not all of the conditions that increase cesarean delivery/birth rates are coded properly. The SMP deferred the question on the appropriateness of the risk adjustment strategy to the Perinatal and Women's Health Standing Committee.

Overall, some of the SMP members noted that the approach to testing reliability and validity was correct; however, the results raised reliability and validity concerns. Following the discussion, the SMP re-voted on validity but did not reach consensus on this criterion. This measure, along with the concerns noted, will move to the Perinatal and Women's Health Standing Committee for evaluation during the spring 2022 cycle.

NQF #0716e ePC-06 Unexpected Newborn Complications in Term Newborns (The Joint Commission)

Measure Steward/Developer Representatives at the Meeting

- Chris Walas
- Stephen Schmaltz
- Elliott Main

SMP Votes

- **Reliability**: Total Votes-9; H-0; M-2; L-3; I-4 (2/9 – 22 percent, No Pass)
- **Validity**: Total Votes-9; H-0; M-2; L-6; I-1 (2/9 – 22 percent, No Pass)

This is an eCQM of an existing chart-based measure (i.e., NQF #0716). In its preliminary analyses, the SMP did not pass the measure on reliability and did not reach consensus on validity. The SMP noted that patient/encounter-level validity testing serves as reliability testing for this measure. There were zero votes for "high" because the highest-possible vote for validity and reliability was "moderate" because the developer only submitted validity testing at the patient/encounter level.

During the SMP's preliminary analyses, the first concern raised was that testing occurred at voluntary sites, where it should be assumed that one would find the most optimistic measurements. Given that there were 61 cases across 14 hospitals (i.e., an average of four cases per site), the data were insufficient to assess within- or across-hospital variability. However, the developer mentioned that this was a statistically representative sample. Another SMP member further raised the following concern: Given the low prevalence of the outcome, conducting a chart review within 61 charts made it difficult assess the reliability and validity of this measure. In addition, some SMP members pointed out that not all data elements in the numerator or denominator were tested, which is required for eCQMs per the NQF measure evaluation criteria. The developer responded to concerns about missing data elements, which were all tested at the same time during pilot testing; however, not all of the results were included in the submission. The developer clarified that all data elements were indeed tested and that the agreement rates for each data element were excellent. While the developer did agree that this was a small sample with a low prevalence of the outcome, they did show that the numerator cases were being captured correctly. Regarding the sampling methodology, the developer clarified that a stratified random sample was present across the hospitals with an oversampling of numerator cases. The developer noted that 30 out of 61 cases had the outcome, and overall, the kappa was calculated to be 0.955 for the numerator's data element testing. An SMP member asked the developer whether the reliability testing was stratified by a data source (i.e., International Classification of Diseases, 10th Revision [ICD-10] versus Systemized Nomenclature of Medicine [SNOMED]). The developer clarified that this process is robust because SNOMED is matched to ICD-10. The SMP re-voted and did not pass the measure on reliability following the discussion.

The lead discussant then described two concerns regarding validity. Although the measure was compared to chart abstraction (i.e., internal validity), the initial submission did not show testing for external validity. It was also noted that no formal testing was conducted for face validity. There were also concerns with the lack of risk adjustment. An SMP member noted that this is an actual outcome

(e.g., a complication rate) measure and that this measure should be risk-adjusted, even in a subsample of full-term nulliparous women. This SMP member was not satisfied with the rationale that the developer provided due to the absence of risk adjustment. Another SMP member noted that there were many maternal factors not examined that would be expected to be associated with newborn complication. This is an important threat to validity and was noted as an important issue that should be considered by the Standing Committee. The developer replied that the chart-based measure is endorsed by NQF and has been in use in California for six to seven years. The developer also clarified that many of the risk factors for complications were included in the exclusions. The second issue is that extracting maternal complications and linking them to the newborn data are challenging in EHRs; therefore, the measure was designed to only use the data from the newborn rather than try to achieve such a linkage. Regarding risk adjustment, the developer analyzed data in 220 hospitals, including data from about 450,000 patients, and showed that good correlation was present between the risk-adjusted and non–risk-adjusted measures, except when higher rates of the measure are reported. When this occurred, more so of a divergence occurred between the risk-adjusted and non–risk-adjusted measures. Regarding the concern for the lack of external validity testing, the developer presented correlations of the measure with several related chart-based measures in their response. The SMP re-voted and did not pass the measure on validity following the discussion.

This measure is eligible for a revote from the Standing Committee. The concerns raised focused on the results of the testing, and the methods used were appropriate. Regarding concerns with the feasibility and accuracy of data elements, Standing Committee members should weigh in on these matters as they work with them in the practice setting. Concerns related to sample size and the lack of risk adjustment need clinical consideration from Standing Committee members as well. The SMP also advised the Standing Committee to discuss the lack of accounting for the severity of the complication in the measure, which could result in facilities with severe complications looking worse than facilities with moderate complications.

NQF #3687e ePC-07 Severe Obstetric Complications (The Joint Commission)

Measure Steward/Developer Representatives at the Meeting

- Chris Walas
- Katie Balestracci
- Valery Danilack
- Lisa Suter

SMP Votes

- **Reliability**: Total Votes-10; H-4; M-5; L-1; I-0 (9/10 – 90 percent, Pass)
- **Validity**: Total Votes-10; H-2; M-6; L-0; I-2 (8/10 – 80 percent, Pass)

In its preliminary analysis, the SMP passed the measure on both reliability and validity. An SMP member pulled this measure to discuss the effect of a risk adjustment variable on the overall model, the correlation between the hospital-level rates of transfusions and the non-transfusion component of the measure, and whether the unexplained variation at the provider level in the risk adjustment model was due solely to the transfusion cases. The SMP did not re-vote on reliability or validity following its discussion. However, it was explicitly stated that the noted concerns with risk adjustment should be conveyed to the Standing Committee for further discussion.

The lead discussant noted that the measure included many complications; however, it appears that much of the variation in the measure is explained by the variation in transfusion rates, and 80 percent of the complications are transfusion related. The SMP also voiced a specific concern with one of the

variables in the risk adjustment model: housing insecurity. Housing insecurity is likely to be an unreliable data element because the prevalence is so low (0.1 percent) and does not match national data (0.6 percent) from other data sources. The underutilization of z-codes is well documented by the Centers for Medicare & Medicaid Services (CMS). The developer replied that the z-codes were used from the housing instability value set. There were 62 encounters (or 0.1 percent) with that code. Homelessness was one of the most used codes. It was noted that large changes occurred in the rate of complications following risk adjustment that excluded the transfusion cases, suggesting that the risk adjustment model may be overfitted. In other words, there may not be a lot of variation in the non-transfusion cases, and the observed differences may be entirely explained by the risk adjustment variables. However, the same did not occur in the transfusion cases. Given this difference, the observed performance rates may reflect the unadjusted differences in transfusion cases. The developer stated that they calculated relatively similar areas under the curves (AUCs) with and without transfusion-only cases, suggesting that the risk adjustment model is not overfitted.

The developer also noted that the variables included in the risk adjustment model emanated from a paper (Leonard et al.). However, an SMP member was persistent in their concern regarding the inconsistent reporting of housing instability. Therefore, it should not be included in the model for a nationally reported quality measure. The developer noted that while underreporting is present, it may compel hospitals to better document housing instability in the future. The same SMP member asked for clarification on why the housing insecurity data element was included in the model and noted that other SDOH variables were used to stratify the measure. The developer noted that race/ethnicity would be used as stratifying variables to illuminate disparities in care, and they specifically did not want to adjust for those differences due to significant national concerns regarding disparities in maternal health outcomes according to race/ethnicity.

The developer also noted that this measure is reported as out of 10,000 (i.e., a rare outcome). Because variation was not present in non-transfusion complications across the sites that were tested, the testing results may have been impacted as a result; however, once the measure is implemented and reported across a larger sample of hospitals, the developer expects variation in non-transfusion complications to increase. Based upon the above discussion, the SMP did not re-vote on reliability and validity. This measure, along with the concerns noted, will move to the Perinatal and Women's Health Standing Committee for evaluation during the spring 2022 cycle.

NQF #2820 Pediatric Computed Tomography (CT) Radiation Dose (University of California, San Francisco [UCSF])

Measure Steward/Developer Representatives at the Meeting

- Rebecca Smith-Bindman

SMP Votes

- **Reliability**: Total Votes-10; H-5; M-4; L-0; I-1 (9/10 – 90 percent, Pass)
- **Validity**: Total Votes-10; H-1; M-7; L-1; I-1 (8/10 – 80 percent, Pass)

In its preliminary analyses, the SMP passed this measure on both reliability and validity. However, one SMP member pulled the measure for discussion based on concern about how the measure is scored and the implications for the measure's validity, specifically, the measure's ability to distinguish sites with poor or excessive radiation from those that do not. The measure has a reference distribution of radiation dose, and sites with a median radiation dose greater than the 75th percentile of the reference distribution are considered poor/excessive. The implication of this is that sites with zero to 49 percent of scans below the 75th percentile are considered acceptable, which is a large range, and up to 49 percent

of scans could be very high in the reference distribution. In addition, the mean radiation dose in acceptable sites could be higher than those in poor sites. A question was also raised about whether a poor site with a median in the 76th percentile is fundamentally different from an acceptable site with a median in the 74th percentile.

The developer noted that each reporting entity has a continuous score of between 0 and 100 percent that reflects the overall proportion of exams that are above the benchmarks of the 75th percentile; those that have more than twice the expected number of high dose exams (i.e., those with more than 50 percent) are considered excessive. However, the raw percentage on the continuous range from 0 to 100 percent allows for an understanding of where a facility lies on the spectrum. The developer also noted that the measure is not designed to identify outliers and acknowledged this as a point of interest while also noting the challenge of small sample sizes when approaching this level of detail, which would threaten the reliability. The SMP did not re-vote on the measure. This measure, along with the concerns noted, will move to the Patient Safety Standing Committee for evaluation during the spring 2022 cycle.

NQF #2377 Overall Defect-Free Care for AMI (American College of Cardiology)

SMP Votes

- **Reliability**: Total Votes-10; H-4; M-6; L-0; I-0 (10/10 – 100 percent, Pass)
- **Validity**: Total Votes-10; H-2; M-7; L-0; I-1 (9/10 – 90 percent, Pass)
- **Composite Quality Construct:** Total Votes-9; H-1; M-8; L-0; I-0 (9/9– 100 percent, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. Subgroup members found the measure to be both reliable and valid. The Primary Care and Chronic Illness Standing Committee will evaluate this measure during the spring 2022 cycle.

Subgroup 2

During the meeting, subgroup #2 discussed five measures (i.e., NQF #3689, NQF #3694, NQF #3695, NQF #3679, and NQF #3697). This subgroup re-voted on the validity of NQF #3689, NQF #3694, and NQF #3695 and on both the reliability and validity of NQF #3679 and NQF #3697. The subgroup accepted the preliminary analysis decisions for NQF #3659 and NQF #3696 without further discussion. The final results for the six measures evaluated by subgroup 2 are presented below.

NQF #3689 First Year Standardized Waitlist Ratio (FYSWR) (Centers for Medicare & Medicaid [CMS]/University of Michigan Kidney Epidemiology and Cost Center [UM-KECC])

Measure Steward/Developer Representatives at the Meeting

- Vahakn Shahinian
- Kevin He

SMP Votes

- **Reliability**: Total Votes-10; H-0; M-10; L-0; I-0 (10/10 – 100 percent, Pass)
- **Validity**: Total Votes-10; H-0; M-8; L-2; I-0 (8/10 – 80 percent, Pass)

In its preliminary analyses, the SMP passed the measure on reliability and did not reach consensus on validity. The SMP discussed the correlations between mortality and subsequent transplant. An SMP member pointed out that the transplant rate may be idiosyncratic from year to year at a dialysis center and questioned whether this is meaningfully demonstrating the quality of the transplant center or the opportunity of acquiring an organ. SMP members agreed that there are other factors that influence subsequent transplant; however, this measure is important and reflects the quality of dialysis providers.

The SMP asked the developer why a larger amount of missing data was present in this measure compared to two other similar submitted measures (i.e., NQF #3694 and NQF #3695). The developer clarified that a different approach was used to capture all dialysis patients in this measure, namely that this measure uses CMS Form #2728 for provider attribution so as not to limit the measure to a Medicare Fee-for-Service patient population.

The SMP expressed some concern with the inclusion of social risk factors but encouraged the Standing Committee to assess the appropriateness of the inclusion of these factors. The SMP was specifically concerned with including social risk factors such as Area Deprivation Index (ADI) and dual eligibility because evidence indicates that social disparities in relation to access to transplantation have a severe impact on patients with end-stage renal disease (ESRD). Therefore, these factors must be mitigated by the accountable entity rather than adjusted for in the measure. SMP members also raised questions on whether the elements in CMS Form #2728 included factors prior to the onset of care or whether there are elements that are concurrent to the care process. The developer clarified that items from CMS Form #2728 all occur prior to the start of the measurement period. The SMP re-voted following the discussion and passed the measure on validity. The Renal Standing Committee will evaluate this measure during the spring 2022 cycle.

NQF #3694 Percentage of Prevalent Patients Waitlisted in Active Status (aPPPW) (CMS/ UM-KECC)

Measure Steward/Developer Representatives at the Meeting

- Vahakn Shahinian
- Kevin He

SMP Votes

- **Reliability**: Total Votes-10; H-5; M-3; L-0; I-2 (8/10 – 80 percent, Pass)
- **Validity**: Total Votes-10; H-0; M-6; L-4; I-0 (6/10 – 60 percent, Consensus Not Reached)

In its preliminary analyses, the SMP passed the measure on reliability but did not reach consensus on validity. The SMP discussed the risk adjustment model, specifically, concurrent risk factors; transplant center characteristics; and the use of sociodemographic status (SDS) factors, such as ADI. The SMP noted the potential for adjusting away some of the transplant center effects by including transplant center characteristics in the risk adjustment model. However, the developer explained that their TEP advised that adjustment was warranted so that providers disproportionately caring for socially vulnerable patients are not unfairly penalized. The SMP also noted the lack of validation using an external data set of the risk adjustment model.

Prior to the meeting, a number of SMP members noted that this measure may be better characterized as a process measure. However, many members accepted the developer's response that being waitlisted requires the achievement and maintenance of a beneficial health status. SMP members also noted a concern: Only a small portion of facilities are categorized as "not as expected" (7.6 percent).

SMP members asked whether the score will be used as a point estimate or a differentiation between categories of provider groups (i.e., average, better than average, or worse than average). The developer replied that they will use these scores to identify those facilities that are significantly different from the average, and SMP members agreed that the testing was appropriate to detect this. SMP members were concerned that the measure may not account for the uncertainty of the estimate if point estimates are used.

NQF #3694 is distinct yet similar to NQF #3695. The SMP asked the developer to explain why they developed two different measures for waitlisting (i.e., waitlisted or waitlisted with active status). The

developer clarified that active status is a subset of waitlisting and requires active maintenance of health status. Furthermore, the developer noted that NQF #3695 is a broader measure that captures the psychological benefit of being on the waitlist. The SMP noted that the Standing Committee should consider whether both measures are clinically necessary. The SMP re-voted following the discussion and again did not reach consensus on the validity of the measure. The Renal Standing Committee will evaluate this measure during the spring 2022 cycle.

NQF #3695 Percentage of Prevalent Patients Waitlisted (PPPW) (CMS/ UM-KECC)

Measure Steward/Developer Representatives at the Meeting

- Vahakn Shahinian
- Kevin He

SMP Votes

- **Reliability**: Total Votes-10; H-4; M-4; L-0; I-2 (8/10 – 80 percent, Pass)
- **Validity**: Total Votes-9; H-0; M-5; L-4; I-0 (5/9 – 56 percent, Consensus Not Reached)

In its preliminary analyses, the SMP passed the measure on reliability but did not reach consensus on validity. Prior to the meeting, some of the SMP members expressed concerns about the nonindependence of patient-months in the measure testing. The developer explained that interdependence of the construct was accounted for using the empirical null method [Efron, B. (2004)]. SMP members also noted in their preliminary analyses that this measure may be better characterized as a process measure, but they deferred to the prior resolution on NQF #3694. SMP members also noted a concern: Only a small portion of facilities are categorized as "not as expected."

The SMP discussed the risk adjustment model, specifically the use of one-year mortality. The developer explained that patients who are at higher risk of mortality are less likely to be deemed candidates for transplantation, and so, they should be adjusted for in the measure. The SMP had similar concerns as with the prior measure with the choice of adjusting for transplant center characteristics and SDS factors, such as ADI. The SMP noted the potential of adjusting away some of the transplant center effects by including transplant center characteristics in the risk adjustment model. However, the developer explained that the attribution for this measure is primarily focused on the dialysis center, which is responsible for the care of patients, such that they may maintain an active waitlist status, not the transplant center. An SMP panel member sought clarity on whether the risk adjustment model showed observed versus predicted values at the facility level. The developer replied that they used direct standardization, which is a similar method to the indirect standardization that was requested.

SMP members asked whether the score will be used as a point estimate or a differentiation between categories of provider groups (i.e., average, better than average, or worse than average). While the developer did state that they will use these scores to identify those facilities that are significantly different from the average, SMP members were concerned that the measure may not account for the uncertainty of the estimate if point estimates are used. One SMP member noted that a better-than-expected performance band is not very good on an absolute basis. The SMP re-voted following the discussion and again did not reach consensus on the validity of the measure. The Renal Standing Committee will evaluate this measure during the fall 2022 cycle.

NQF #3679 Home Dialysis Rate (Kidney Care Quality Alliance [KCQA])

Measure Steward/Developer Representatives at the Meeting

- Lisa McGonigal
- Kathy Lester
- Craig Solid

SMP Votes

- **Reliability**: Total Votes-10; H-0; M-1; L-4; I-5 (1/10 – 10 percent, No Pass)
- **Validity**: Total Votes-10; H-0; M-5; L-2; I-3 (5/10 – 50 percent, Consensus Not Reached)

In its preliminary analyses, the SMP did not reach consensus on both reliability and validity. Regarding reliability, SMP members identified as a key concern that the unit of analysis is patient-months, and the reliability testing did not account for the interdependence of patient-months within the same patient. This interdependence violates the assumptions of the beta-binomial model, rendering it an inappropriate approach for reliability testing. A related concern was the discovery of extremely high reliability for very small sample sizes (less than 10 patient-months); these reliability values may be overestimated due to not accounting for the interdependence of patient-months. The developer appreciated the feedback and requested additional guidance on how to improve testing. The developer also noted that they did not have access to patient-level data that would allow them to address the intra-patient correlation. An additional concern focused on the level of reporting. As the SMP understood the presentation of the testing, testing was conducted at the level of individual facilities and for Hospital Referral Regions (HRRs) but not for the accountable entity of a parent organization within an HRR. An additional concern was raised: The measure is proposed as a stratified measure, but no reliability testing was conducted using stratified performance. The SMP subgroup re-voted on reliability but did not pass the measure on this criterion.

Regarding accountable-entity empirical validity testing, SMP members again expressed concern that empirical validity testing was conducted at the level of the overall HRR and not the accountable entity of the parent organization within the HRR. SMP members also questioned the comparator used for empirical validity testing with concern that it was essentially an earlier version of the proposed measure; therefore, it is not a useful indicator of validity. Regarding face validity, some members expressed concerns with the composition of the expert panel voting on face validity, including the lack of a patient/caregiver representative and the inclusion of members who were associated with the measure developer. The developer noted that there are patients involved in other workgroups within the Kidney Care Quality Alliance (KCQA); however, when they sent out the call for participation in the expert panel voting on face validity, no patient representatives responded. The developer also noted that they broadly engage a wide range of professionals and organizations in the kidney care community, and it would be difficult to exclude all of them from participating in face validity assessments and still have appropriate expertise. NQF staff clarified that NQF is not prescriptive about which stakeholder groups should be included in face validity assessments. SMP members also had concerns about the lack of risk adjustment. The developer noted that they carefully considered risk adjustment methodologically by following both NQF guidance as well as implications for practice with the primary user of the measure (i.e., CMS). Methodologically, they found that adjustment had little overall effect on the performance scores, yet stratification did enable the identification of disparities. For implementation, there is a preference to not risk-adjust in order to avoid masking disparities and to use stratification instead. The SMP subgroup re-voted on validity but did not reach consensus on this criterion.

This measure is not eligible for a re-vote from the Renal Standing Committee because the SMP determined that inappropriate methodology was used. Specifically, the nonindependence of patient-months is not accounted for and violates the assumptions of the beta-binomial mode. Furthermore, it is unclear whether the level of analysis is at the HRR or the facility.

NQF #3697 Home Dialysis Retention (KCQA)

Measure Steward/Developer Representatives at the Meeting

- Lisa McGonigal

- Kathy Lester
- Craig Solid

SMP Votes

- **Reliability**: Total Votes-10; H-0; M-0; L-9; I-1 (0/10 – 0 percent, No Pass)
- **Validity**: Total Votes-9; H-0; M-2; L-5; I-2 (2/9 – 22 percent, No Pass)

In its preliminary analyses, the SMP did not pass the measure on reliability and did not reach consensus on validity. This measure is paired with NQF #3679 above. Similar concerns were raised regarding this measure as with NQF #3679. Consequently, the discussion of this measure focused on distinct concerns. Regarding reliability, SMP members noted the large number of facilities with denominators less than 10 as well as the overall low level of reliability attained. The SMP subgroup re-voted on reliability but did not pass the measure on this criterion.

Validity testing focused on face validity using the same processes and questions for NQF #3679. Separate risk adjustment was not conducted for NQF #3697; the risk adjustment analyses from NQF #3679 were used as the basis for the determination to stratify the measure rather than risk-adjust it. The SMP subgroup re-voted on validity and did not pass the measure on this criterion. There were zero votes for "high" because the highest-possible vote for validity was "moderate" because the developer only submitted face validity testing.

This measure is not eligible for a revote from the Renal Standing Committee because the SMP determined that inappropriate methodology was used. Specifically, the beta-binomial model was not appropriately applied, and it was unclear whether the level of analysis is at the HRR or the facility. Additionally, the SMP stated that the small testing sample sizes yielded unreliable results.

NQF #3659 Standardized Fistula Rate for Incident Patients (CMS/UM-KECC)

SMP Votes

- **Reliability**: Total Votes-10; H-3; M-4; L-1; I-2 (7/10 – 70 percent, Pass)
- **Validity**: Total Votes-10; H-1; M-7; L-2; I-0 (8/10 – 80 percent, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. The Renal Standing Committee will evaluate this measure during the spring 2022 cycle.

NQF #3696 Standardized Modality Switch Ratio for Incident Dialysis Patients (SMoSR) (CMS/ UM-KECC)

SMP Votes

- **Reliability**: Total Votes-8; H-0; M-6; L-2; I-0 (6/8 – 75 percent, Pass)
- **Validity**: Total Votes-8; H-1; M-5; L-2; I-0 (6/8– 75 percent, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. The Renal Standing Committee will evaluate this measure during the spring 2022 cycle.

Discussion of Overarching Methodological Issues Identified During Measure Evaluation

Risk Adjustment

The issue of assessing risk adjustment as a threat to validity has been a recurring topic within the SMP for some time. Specifically, the SMP has had difficulty with evaluation questions of how and when risk

adjustment is appropriate as well as the inclusion of certain factors for risk adjustment. It is NQF's policy that the SMP cannot vote "low" or "insufficient" on a measure simply due to the inclusion or exclusion of certain risk factors; however, it can if the methodology used in either assessing risk adjustment or building a risk adjustment model is incorrect or flawed. Many SMP members have consistently struggled with this subtle line due to some instances in which the methods are impacted by the risk factors chosen.

The SMP gave some specific examples of when the line is not clearly drawn between the inclusion or exclusion of factors and methodological issues. It stated that a risk adjustment model can yield high c-statistics that are biased because of the inclusion of certain risk factors. Additionally, SMP members noted other situations in which the coefficient or odds ratio of a given variable that is excluded from the model actually suggests it has an equal or greater association with the outcome of interest than the variables retained in the model. In this instance, even without clear clinical understanding, that decision remains methodologically questionable and is within the SMP's purview. Generally speaking, the SMP members contended that while they oftentimes lack the clinical expertise that the Standing Committees have, their expertise in the statistical methodology allows them to comment on specific risk factors when a risk model is not well constructed, especially due to significant influences from specific factors that threaten the validity of a measure.

During these SMP meetings, the issue of risk adjustment arose several times. The issue typically consisted of the inclusion or exclusion of certain risk factors, a lack of validation of the data set used in the risk adjustment model, and concerns with using risk stratification as opposed to risk adjustment. In most instances of risk adjustment in which the methodology was appropriate, the SMP chose to defer the conversation to the Standing Committees. Its concerns will be relayed to the Standing Committees. If the methodology was inappropriate, the SMP did not pass the measure in part because of the inherent threat to validity; however, in these cases, there were also typically other concerns with validity. Some SMP members suggested that they discuss this issue further and possibly create some guidance on when risk adjustment is an SMP issue versus a Standing Committee issue at a future SMP advisory call.

Use of Patient-Months

During this cycle, several measures utilized patient-months in their calculations. The SMP clarified that the use of patient-months in and of itself is acceptable; however, developers must account for the nonindependence of patient-months within persons in the reliability and validity calculations. If this is not done, the interdependence of months that each patient contributes to a measure will artificially elevate tests of the measure score's reliability. One developer that used patient-months provided a response to the SMP's concerns, noting that they used the empirical null method to separate the underlying variation among patient-months from variation that might be attributed to quality of care. The SMP generally accepted this method. The other developer that used patient-months did not account for this in their testing, although they explained the importance of the patient-months construct to avoid "gaming" of the measure. Because of this matter, the SMP had significant concerns that the interdependence of the patient-months construct violated the assumptions of the statistical testing. It noted that the reliability values were likely overestimated because the developer did not account for interdependence.

The conversation regarding patient-months prompted NQF staff to remind participants that technical assistance is available to developers from NQF staff for issues such as this in which explicit guidance is not available in NQF documents. NQF is also working to expand the availability and awareness of technical assistance offerings.

Patient/Encounter-Level Validity Testing Used for Patient/Encounter-Level Reliability Testing

NQF's Measure Evaluation Criteria and Guidance states that "for some measure types, separate patient/encounter-level reliability testing of the data elements is not required if patient/encounter-level empirical validity testing of the data elements is conducted (and results are adequate)" (page 19). Several measures in the spring 2022 cycle used encounter-level validity testing to demonstrate encounter-level reliability (namely, NQF #1460, NQF #0471e, and NQF #0716e). Some of the SMP members who evaluated these measures recognized that this policy is in the criteria but raised concerns, considering reliability and validity are distinct concepts. NQF staff clarified that when voting on validity after reliability, threats to validity must also be considered beyond the encounter-level validity testing. Additionally, patient/encounter-level validation may only support patient/encounter-level reliability. It does not demonstrate reliability at the accountable-entity level, which is required for some measure types. Some SMP members still raised an issue with the policy, as it may obviate the need for developers to demonstrate accountable-entity level reliability.

Public Comment

No public or NQF member comments were provided during the measure evaluation meeting.

Next Steps

Ms. Ingber reviewed the next steps and reminders for the SMP and the measures that the SMP members reviewed during this cycle. NQF staff will inform the developers and Standing Committees of the SMP's discussion and votes. Measures that passed on both reliability and validity or for which consensus was not reached will be considered by the relevant Standing Committees during the spring 2022 evaluation cycle.

According to NQF endorsement guidance, measures that did not pass the SMP's vote may be pulled for discussion by the relevant Standing Committee. However, measures are not eligible for a revote if any of the following are true: (1) An inappropriate methodology or testing approach was applied to demonstrate reliability or validity; (2) Incorrect calculations or formulas were used for testing; (3) A description of the testing approach, results, or data is insufficient for the SMP to apply the criteria; and (4) Appropriate levels of testing are not provided or otherwise did not meet NQF's minimum evaluation requirements.

NQF #0716e is eligible for discussion and a revote from the Standing Committee if it is pulled by the respective Standing Committee. NQF #3679, NQF #1460, and NQF #3697 are ineligible for a re-vote due to the inappropriate methodology used to demonstrate reliability and/or validity. Measures moving forward in the spring 2022 evaluation cycle will be reviewed by their respective Standing Committees in June or July 2022 and discussed by the Consensus Standards Approval Committee (CSAC) in November 2022.

The SMP will convene via web meeting on April 27, 2022. During this meeting, the SMP will discuss its role in and suggestions for the Consensus Development Process (CDP).