



January 10, 2017

NQF Board of Directors  
National Quality Forum  
1030 15<sup>th</sup> St NW, Suite 800  
Washington, DC 20005

*Subject: Appeal of Measure 2979—Standardized Transfusion Ratio (STrR) for Dialysis Facilities*

Dear NQF Board of Directors:

Pursuant to the National Quality Forum's (NQF) Consensus Development Process (CDP), Kidney Care Partners (KCP) appeals NQF's decision to endorse NQF 2979, *Standardized Transfusion Ratio (STrR) for Dialysis Facilities*, considered within the NQF 2015-2017 Renal Project. We do so on the grounds of procedural errors reasonably likely to affect the outcome of the original endorsement decision, which directly and materially affects the interests of dialysis providers and dialysis patients and will have an adverse effect on those interests.

KCP is a coalition of members of the kidney care community that includes the full spectrum of stakeholders related to dialysis care—patient advocates, health care professionals, dialysis providers, researchers, and manufacturers and suppliers—organized to advance policies that improve the quality of care for individuals with both chronic kidney disease and end stage renal disease (ESRD). KCP commends NQF for undertaking this important work addressing renal care. Further, we appreciate the process by which performance measures are evaluated through the CDP and recognize the need for careful and deliberate evaluation of candidate national voluntary consensus standards—particularly those proposed for use in Federal payment and penalty-based purchasing programs, as is NQF 2979.

KCP strongly agrees that transfusion avoidance is an important aspect of ESRD care, but we do not believe the Renal Standing Committee fully considered and appropriately addressed the substantial technical concerns with the STrR raised by KCP during the project comment periods or conveyed these concerns to the Consensus Standards Advisory Committee (CSAC). **We note that this failure has resulted in dissonance between the CSAC's endorsement decision and that of the Measure Applications Partnership (MAP) Hospital Workgroup, which due to the same technical concerns raised by KCP and the proposed deployment of the measure in the penalty-based ESRD Quality Incentive Program (QIP), recommended that the developer (CMS) refine and resubmit the STrR to the MAP for future consideration.**

KCP believes that the volunteer, multi-stakeholder Standing Committees are a central component to the endorsement process. We fully understand the substantial amount of work required of these Standing Committees in a relatively narrow timeframe. However, we believe the Renal Committee did not consider, nor revisit and reconsider after the comment period, the significant technical concerns identified by a broad-based multi-stakeholder coalition and longstanding NQF-member. As we describe in further detail below, we posit significant reliability and validity issues exist, and that the decision to advance the measure despite it clearly not meeting NQF's endorsement criteria was a process failure that compromises the integrity of NQF's endorsement.

## **BACKGROUND**

KCP first conveyed its concerns with the measure to the Renal Standing Committee in its [June](#)

[13 early comment letter](#) (Attachment 1). As we detail further below, the issues were raised during the in-person meeting on June 28, but we believe the Committee's conclusion to advance the measure was not well-documented nor consistent with NQF's established guidelines for evaluating measure testing.<sup>1</sup> As indicated below, this is particularly true regarding NQF's *Reliability* criterion, for which we posit the failure of the measure to meet even the minimal established endorsement criteria invalidate the measure for use in public reporting and accountability programs and should have, at a minimum, resulted in a more robust discussion by the Committee.

KCP reiterated its concerns in its [August 31 comment letter](#) (Attachment 2) and in its October 17 voting comments. While CMS did respond to some of those comments following the project's formal comment period (Attachment 3 and [excerpts] below), those responses did not resolve the measure's numerous technical flaws or alleviate the concerns with the proposed use in a penalty-based program of a performance metric that has been demonstrated through empirical testing as statistically unreliable for a large proportion of the providers for which it will be deployed. Despite the developer's failure to adequately address these underlying technical flaws, the Renal Standing Committee opted not to revisit the issues raised during the comment period, instead adopting the response proposed for it by NQF staff:<sup>2</sup> *"The Committee thoroughly reviewed the specifications, reliability, and validity of the measure during the in-person meeting and maintains that the measure meets the NQF criteria."*

We disagree with the characterization that the review was thorough, and believe the Committee failed to fully consider the technical issues of the measure—and the adverse impact those issues will have on both dialysis patients and providers—at the in-person meeting. It is the failure of this initial step in the process, in particular, upon which KCP seeks to appeal the endorsement decision.

Below we reiterate KCP's two most significant concerns with the STrR. We also footnote responses from CMS and the Renal Standing Committee and detail the procedural errors we believe were at odds with NQF's endorsement criteria and process.

### UNACCEPTABLE RELIABILITY IN SMALL AND MEDIUM FACILITIES

As KCP has noted ([Attachments 1](#) and [2](#)), the most problematic issue with the STrR is the unacceptably low reliability of the measure for small and, to a lesser degree, medium dialysis facilities. Specifically:

- An Inter-Unit Reliability (IUR) statistic of  $\geq 0.70$  is generally considered "acceptable" in the statistical literature; 0.60-0.69 is "questionable", 0.50-0.59 is "poor", and  $\leq 0.49$  is "unacceptable".<sup>3,4,5</sup>
- Empirical testing for the STrR yielded *overall* IURs *below* the "acceptable" threshold—0.60-0.66 ("questionable") across all facilities for each of the four testing years (2011, 2012, 2013, and 2014).<sup>6</sup>
- IURs were substantially lower when looking exclusively at small facilities (defined as  $\leq 46$  patients by CMS). Specifically, IURs for small facilities ranged from only 0.30-0.41 ("unacceptable"), indicating that approximately **60-70% of a small facility's score is due to random noise**.<sup>7</sup>
- Even in medium-sized facilities (47-78 patients), the STrR was found to have what is generally interpreted as "poor" reliability, with IURs of 0.50-0.56 (i.e., **50% of a medium-sized facility's score is secondary to noise**).<sup>8</sup>
- Lastly, we note that per CMS's testing data submitted to NQF,<sup>9</sup> the number of small facilities over the four years of testing averaged 1,829 per year, medium facilities 1,804

per year, and large facilities 1,756 per year—i.e., facilities for which empirical testing demonstrated the measure has “poor” or “unacceptable” reliability outnumber those for which reliability is “acceptable” by greater than two to one.

We assert that a performance metric that has been empirically demonstrated as statistically unreliable for greater than 67% of the providers for which it will be deployed very clearly does not meet NQF’s established endorsement criteria for *Reliability*. Specifically, using NQF’s *Guidance for Evaluating Reliability Algorithm* (Algorithm 2), which states that where there is “low certainty or confidence that the performance measure scores are reliable” and where no other reliability testing was reported (there was none), reliability should be rated as “**low**.”<sup>10</sup> We note that NQF endorsement of a measure that does not meet the criteria for reliability is of particular concern when that measure is proposed for use in a penalty-based program, where statistical aberrancies can translate to hardship for the many facilities already operating under significant financial strain.

While not captured in the draft or voting reports for the project, we recall and note that the meeting transcript confirms that the Renal Standing Committee did discuss the issue of reliability, though we posit that it did not fully consider all issues prior to the vote, thereby resulting in an “incomplete” consideration:<sup>11</sup>

- One Committee member expressed significant concern that empirical testing demonstrated that the measure is unreliable in small and medium facilities—together more than two-thirds of the facilities in which it will be used.
- In response, another member noted that because reporting reliability statistic variations by provider size is not specifically requested by NQF in its measure submission documents, another developer might have only provided the more acceptable (but still only “questionable,” per accepted statistical literature standards) overall averaged IURs of 0.60-0.66, and the Standing Committee would have remained unaware of the issue of poor reliability in small and medium-sized facilities. The Committee member remarked that providing “additional details that are not necessarily requested” that then raise concerns leads to “some cognitive dissonance about how to process and arrive at a conclusion.”
- Despite these comments, no further discussion was pressed and the vote, despite the empirical evidence on low reliability for >67% of facilities, proceeded (75% moderate, 25% low).

We believe the process failed at this point. The Committee should have discussed the matter to definitively arrive at a conclusion that would justify the noted “cognitive dissonance” between the empirical data and the resulting vote. That is, the Committee and/or NQF staff should have pressed for a conclusive discussion at the in-person meeting and also noted to the CSAC the magnitude of the reliability statistic disparities between facilities of varying sizes and that the measure is unreliable in more than 67% of facilities in which it will be applied. We assert that merely because another developer might not have provided the stratified reliability statistics that revealed the lack of reliability in small- and medium-sized facilities does not mean these data, once presented, should be ignored. Additionally, reliability in large facilities that is sufficiently high to raise the *overall* average IUR to a more acceptable level does not negate the fact that the STTrR is clearly unreliable in more than two-thirds of the facilities to which the measure applies. Finally, we do not believe that endorsing a measure for which 50-70% of the score is due to random noise for the vast majority of the measured entities is in accordance with NQF’s definition of the reliability criterion.

We note that despite these unacceptably low reliability statistics, KCP has suggested in its comments that CMS could address this concern through the empirical identification and

application of a minimum sample above which all facilities included in the measure calculations meet acceptable reliability statistics. CMS has since responded that a requirement for a minimum of 10 patient-years-at-risk, while not indicated in the measure specifications, is applied in the current implementation of the measure for Dialysis Facility Compare and for PY2018 for the QIP, and was also applied to the analyses in the *Measure Testing Form* submitted to NQF.<sup>A,12</sup> However, we note that this indicates that **the low reliability statistics seen in small and medium facilities were generated after already excluding facilities with <10 patient-years-at-risk. Thus the minimum sample requirement as currently defined and applied clearly does not adequately address the issue at hand.** And despite the developer's comments, which we posit do not address the measure's reliability, the Renal Standing Committee did not revisit the reliability issues during the comment period.<sup>B</sup>

***Conclusion and Requested Actions:*** *As detailed above, KCP maintains that the measure does not meet NQF endorsement criteria for Reliability and that the Committee did not fully discuss and reach a definitive conclusion—a procedural error—as documented in the transcript, prior to the subsequent vote. Indeed, a Committee member comment noted a “cognitive dissonance,” but the Committee was not pressed to a conclusion. We request that the NQF Board rescind endorsement on these procedural grounds, which led to endorsement of measure which will adversely impact the 67% of dialysis providers for which the metric has been empirically demonstrated as statistically unreliable and dialysis patients seeking to make well-informed decisions about their care. Absent that, we request that the Board condition the current endorsement on the developer using its testing data to empirically determine appropriate facility-level exclusion parameters to assess the impact of a “small numbers” effect on reliability and to work with the renal community to determine the sample size that yields an acceptable reliability statistic (ideally,  $\geq 0.7$ ). We note that such a conditional endorsement was placed on NQF 2496, Standardized Readmission Ratio for Dialysis Facilities. We also note that such refinement would be in keeping with the MAP Workgroup’s recommendation to “refine and resubmit.”*

## TRANSFUSION EVENT DEFINITION AS THREAT TO VALIDITY

Also of substantial concern to KCP is CMS's approach in its revisions to the STTrR measure specifications to more “conservatively” define transfusion events. We have specifically noted the following:

- All inpatient transfusion events must now include an appropriate ICD-9 Procedure Code or Value Code to be captured in the measure; inpatient transfusion events for claims that include only transfusion revenue codes without an accompanying procedure or value code are not included in the numerator.
- There is no existing coding requirement that procedure or value codes be used; valid transfusion claims that include only revenue codes will be missed, creating a significant threat to measure validity.
- Current transfusion coding practices vary by hospital,<sup>13</sup> and hospital coding practices are beyond dialysis facilities' sphere of control. For example, hospitals that exclusively use revenue codes for transfusions will appear to have no events assigned to a dialysis facility, while hospitals that *do* use procedure and /or value codes *will* have recorded

<sup>A</sup> *CMS Response to KCP Comments:* “This requirement for minimum number of patients/patient years at risk is not part of the measure specifications, but applied in the current implementation of the measure for DFC and for PY2018 QIP. The analyses in the Testing Form applied this requirement, in order to align with current public reporting standards.”

<sup>B</sup> *Renal Standing Committee Response to KCP Comments:* “The Committee thoroughly reviewed the specifications, reliability, and validity of the measure during the in-person meeting and maintained that the measure meets the NQF criteria.”

events. Facilities within given catchment areas will thus be differentially affected by hospital coding variations, which will clearly impact STrR scoring.

These revisions will result in increased variability in performance across dialysis facilities wholly due to external factors and not performance. Simply put, some facilities will appear to have “poor” performance because of higher than expected numbers of transfusions—and will expend time and resources to improve—when in fact the score is merely a reflection of hospital coding practices.

CMS responded to this concern<sup>C</sup> by remarking that the definition of transfusion events used in the STrR is consistent with definitions used in numerous scientific publications,<sup>14,15,16,17,18,19</sup> including “several peer-reviewed publications by the research group that presented the American Society of Nephrology (ASN) poster” KCP cited in its comments.<sup>20</sup> We note, however, that the identification of transfusion events for the purposes of research need not be as precise as the process used to inform a public reporting and penalty-based purchasing program, wherein inaccuracies will lead to the improper and inappropriate penalization of dialysis facilities—many of which are already operating on narrow financial margins. Moreover, CMS remarks that it is not aware of any scientific publication demonstrating that the definition of transfusion events used in the revised measure is invalid, conflicting with its reference to the ASN poster. In fact, the ASN poster is the only of the several scientific works referenced by CMS on this issue that was undertaken with the specific purpose of evaluating how variability in hospital-level billing patterns impacts the reliability of dialysis facility-level transfusions rates and consequently, the validity of STrR estimation—which the authors note crucially depends on the accurate identification of transfusions during hospitalization. The authors clearly concluded that **between-hospital variability in billing patterns for blood is substantial and that use of specific definitions of whole blood or RBC transfusion will induce differential misclassification according to location, resulting in biased estimation of STrR.**

Although not captured in the draft or voting reports, the Renal Standing Committee did discuss the impact of the revised coding requirements on numerator (and thus measure score) variability during its in-person meeting on June 28, but again did not pursue the discussion to reach a definitive conclusion in terms of the impact of the revised coding requirements on measure validity, in particular. Specifically, prior to the discussion on validity, one Committee member pointed out that because procedure codes allow for the capture of multiple transfusion events in a single day while value codes can capture only a single event per day, the coding revisions could allow for the same clinical scenario to result in different numerator counts (and performance scores) depending on which type of codes a given hospital uses.<sup>21</sup> The developer responded that the “vast majority of transfusion events” are coded using either procedure codes alone or in combination with revenue center codes, and maintained that the revisions will minimize the number of false positive events (our characterization) for the measure. However, the developer also noted that up to 35% of transfusion events identified in the data presented for the measure last year used revenue codes without an accompanying procedure code.

**C Response to KCP Comments:** “The definition of transfusion events used in the revised STrR measure is consistent with definitions used in numerous scientific publications, including several peer review publications by the research group that presented the ASN abstract referred to by the commenters.<sup>9-14</sup> The definition is also structurally consistent with Medicare claims processing rules. By excluding transfusion events identified only through revenue codes, the false positive identification of blood transfusions should be reduced, per the Medicare claims processing rules and guidelines published by the American Red Cross and other blood banking organizations. By definition, exclusion of revenue code only transfusion events decreases variation due to hospital coding practices that may rely primarily on revenue codes. We have empirically demonstrated this revision does not substantially alter the strong relationship between recent prior achieved hemoglobin and subsequent transfusion risk, a relationship that has been previously shown in other research studies. Furthermore, we are not aware of any scientific publication demonstrating that the definition of transfusion events used in this revised measure is invalid. It should be noted that this issue was also discussed in detail during the ESRD Standing Committee’s discussion of the STrR at the June, 2016 in-person meeting, prior to the ESRD Standing Committee vote to recommend the measure for endorsement.”

Moreover, the developer indicated that there is “a lot” of variability in transfusion coding practices across states—for instance, while only 14% of Rhode Island’s transfusion events are coded using revenue codes only, up to 75% of Utah’s events are coded this way.<sup>22</sup> The discussion concluded and was not raised again by the Standing Committee when later discussing measure validity. Ultimately, the measure passed on validity with a vote of 75% moderate, 25% low.

As with the reliability discussion, KCP posits that an additional procedural error occurred at this point. According to NQF’s *Guidance for Evaluating Validity Algorithm* (Algorithm 3),<sup>23</sup> all potential threats to validity relevant to the measure in question (e.g., exclusions [2b3], need for risk adjustment [2b4], **missing data** [2b7]) must be empirically assessed; failure to do so leads to a validity rating of “insufficient.” We note that the developer did not provide an assessment of missing data (i.e., the measure’s failure to capture the 35% [per CMS]<sup>24</sup> of transfusions coded with a revenue code without an accompanying procedure code) in its measure submission documents to NQF,<sup>25</sup> and the Renal Committee did not discuss the implications of such missing events as a threat to measure validity during its deliberations, despite KCP’s pre-meeting comments to this effect.

We appreciate CMS’s desire to minimize the potential for the measure to capture “false positive” transfusion events, but we note that the proposed solution will instead result in a high proportion of “false negatives”, thereby sacrificing measure sensitivity for specificity. KCP maintains that the substantial regional variations and the exclusive use of revenue codes for more than a third of transfusion events noted by CMS during the Standing Committee’s in-person meeting<sup>26</sup> validates the fear that hospital coding deviations will materially and inappropriately impact the STrR performance scores for dialysis facilities. Again, however, the Committee did not pursue the discussion to a conclusion in terms of the implications of the impact of these coding variations on measure validity, and despite KCP specifically reiterating this concern during the comment period, the Committee opted not to subsequently revisit the issue.<sup>D</sup>

***Conclusion and Requested Actions:*** *As detailed above, KCP maintains that the measure does not meet NQF endorsement criteria for Validity and that the Committee did not fully discuss the matter of threats to validity and missing data, a procedural error as demonstrated in the transcript, prior to the subsequent vote. We request that the NQF Board rescind endorsement on these grounds. Absent that, we request that the Board condition the current endorsement on the developer working with the community to consider either alternative models or revising hospital transfusion coding rules to require that the ICD-9/ICD-10 procedure and value codes required for the validity of the proposed methodology be universally included in claims. We note that such refinement would be in keeping with the MAP Workgroup’s recommendation to “refine and resubmit.”*

KCP appreciates your consideration of this issue and attention to our request for this appeal. We urge the NQF Board to reconsider the recent decision to endorse NQF 2979, *Standardized Transfusion Ratio (STrR) for Dialysis Facilities*. Please do not hesitate to contact Lisa McGonigal, MD (lmcgon@msn.com), with any questions concerning this matter.

Sincerely,

<sup>D</sup> *Renal Standing Committee Response to KCP Comments:* “The Committee thoroughly reviewed the specifications, reliability, and validity of the measure during the in-person meeting and maintained that the measure meets the NQF criteria.”



A handwritten signature in black ink that reads "Sara Love Rawlings". The script is fluid and cursive, with the first letters of each word being capitalized and prominent.

Sara Love Rawlings  
Executive Director

- 1 NQF. *Review and Update of Guidance for Evaluating Evidence and Measure Testing*. October 8, 2013. P. 15. Available at: [http://www.qualityforum.org/Publications/2013/10/Review\\_and\\_Update\\_of\\_Guidance\\_for\\_Evaluating\\_Evidence\\_and\\_Measure\\_Testing\\_-\\_Technical\\_Report.aspx](http://www.qualityforum.org/Publications/2013/10/Review_and_Update_of_Guidance_for_Evaluating_Evidence_and_Measure_Testing_-_Technical_Report.aspx).
- 2 NQF. *Renal Comments Received and Actions Taken Spreadsheet*. Available at: <http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80747>.
- 3 George, D and Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 update (4th ed.). Boston: Allyn & Bacon.
- 4 Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge, page 13.
- 5 DeVellis, RF. (2012). *Scale development: Theory and applications*. Los Angeles: Sage. pp. 109–110.
- 6 NQF. NQF 2979 Measure Testing Form. *Section 2a2: Reliability Testing*. Table 6; page 7. Available at: <http://www.qualityforum.org/ProjectMeasures.aspx?projectID=80747>.
- 7 Ibid.
- 8 Ibid.
- 9 Ibid.
- 10 NQF. *Review and Update of Guidance for Evaluating Evidence and Measure Testing. Algorithm 2: Guidance for Evaluating Reliability*. Page 15. October 8, 2013. Available at: [http://www.qualityforum.org/Publications/2013/10/Review\\_and\\_Update\\_of\\_Guidance\\_for\\_Evaluating\\_Evidence\\_and\\_Measure\\_Testing\\_-\\_Technical\\_Report.aspx](http://www.qualityforum.org/Publications/2013/10/Review_and_Update_of_Guidance_for_Evaluating_Evidence_and_Measure_Testing_-_Technical_Report.aspx).
- 11 NQF. Renal Meeting Transcript. Pages 218-224. June 28, 2016. Available at: <http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80747>.
- 12 NQF. *Renal Comments Received and Actions Taken Spreadsheet*. Available at: <http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80747>.
- 13 Weinhandl ED, Gilbertson DT, Collins AJ. *Dialysis facility-level transfusion rates can be unreliable due to variability in hospital-level billing patterns for blood*. Chronic Disease Research Group poster, ASN. 2014.
- 14 Hirth, Turenne, Wilk, et al. Blood transfusion practices in dialysis patients in a dynamic regulatory environment. *Am J Kidney Dis*. 2014;64(4):616-621.
- 15 Gilbertson, Monda, Bradbury, and Collins. RBC transfusions among hemodialysis patients (1999-2010): Influence of hemoglobin concentrations below 10 g/dL. *Am J Kidney Dis*. 2013;62(5):919-928.
- 16 Collins et al. Effect of facility-level hemoglobin concentration on dialysis patient risk of transfusion. *Am J Kidney Dis*. 2014;63(6):997-1006.
- 17 Cappell et al. Red blood cell (RBC) transfusion rates among US chronic dialysis patients during changes to Medicare end-stage renal disease (ESRD) reimbursement systems and erythropoiesis stimulating agent (ESA) labels. *BMC Nephrology*. 2014;15:116.
- 18 Ibrahim, et al. Blood transfusions in kidney transplant candidates are common and associated with adverse outcomes. *Clin Transplant*. 2011;25:653–659.
- 19 Molony, et al. Effects of epoetin alfa titration practices, implemented after changes to product labeling, on hemoglobin levels, transfusion use, and hospitalization rates. *Am J Kidney Dis*. 2016: epub before print (published online March 12, 2016).
- 20 Weinhandl ED, Gilbertson DT, Collins AJ. *Dialysis facility-level transfusion rates can be unreliable due to variability in hospital-level billing patterns for blood*. Chronic Disease Research Group poster, ASN. 2014.
- 21 NQF. Renal Meeting Transcript. Pages 210-217. June 28, 2016. Available at: <http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80747>.
- 22 NQF. Renal Meeting Transcript. Pages 212-213. June 28, 2016. Available at: <http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80747>.
- 23 NQF. *Review and Update of Guidance for Evaluating Evidence and Measure Testing. Algorithm 3: Guidance for Evaluating Validity*. Page 16. October 8, 2013. Available at: [http://www.qualityforum.org/Publications/2013/10/Review\\_and\\_Update\\_of\\_Guidance\\_for\\_Evaluating\\_Evidence\\_and\\_Measure\\_Testing\\_-\\_Technical\\_Report.aspx](http://www.qualityforum.org/Publications/2013/10/Review_and_Update_of_Guidance_for_Evaluating_Evidence_and_Measure_Testing_-_Technical_Report.aspx).
- 24 NQF. Renal Meeting Transcript. Page 212, Line 1. June 28, 2016. Available at: <http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80747>.
- 25 NQF. NQF 2979 Measure Testing Form. *Section 2b7: Missing Data Analysis and Minimizing Bias*. Pages 26-27. Available at: <http://www.qualityforum.org/ProjectMeasures.aspx?projectID=80747>.
- 26 NQF. Renal Meeting Transcript. Pages 212-213. June 28, 2016. Available at: <http://www.qualityforum.org/ProjectMaterials.aspx?projectID=80747>.





June 13, 2016

National Quality Forum  
1030 Fifteenth Street, NW, Ste 800  
Washington, DC 20005

RE: NQF Renal Project

Kidney Care Partners (KCP) appreciates the opportunity to comment on the measures under consideration for endorsement in the National Quality Forum's (NQF) Renal Measures 2015-2017 Project. KCP is a coalition of members of the kidney care community that includes the full spectrum of stakeholders related to dialysis care – patient advocates, health care professionals, dialysis providers, researchers, and manufacturers and suppliers – organized to advance policies that improve the quality of care for individuals with both chronic kidney disease and end stage renal disease. We commend NQF for undertaking this important work and offer comment on all six measures.

**NQF 0260: Assessment of Health-Related Quality of Life (QoL) in Dialysis Patients (Witten and Associates, LLC)**

KCP recognizes the importance of assessing the health-related quality of life for individuals with ESRD. Nevertheless we have an overarching concern about the measure, as well as specific concerns about the new specifications, evidence, performance gap, and validity.

- **OVERARCHING ISSUE.** Annual administration of the KDQOL is already required by Federal regulation, the Conditions for Coverage. KCP questions how endorsement of a measure for a process that is already mandated and surveyed will further improve patient care.
- **SPECIFICATIONS.** We support the changes to the exclusions that align them with the Conditions for Coverage, but KCP opposes eliminating the exclusion for patient refusal. First, the Conditions for Coverage permit patient refusal as long as it is documented. We believe approving a measure that directly conflicts with Federal regulation is problematic. Second, not accepting patient decisionmaking ignores patient autonomy; providers should not be forced to face intruding on patient decisionmaking vs. facing a penalty for poorer performance on this measure. We further note there is no performance gap when the specifications include patient refusal.
- **EVIDENCE.** As noted, KCP recognizes the importance of assessing health-related quality of life, but questions the lack of direct evidence for the measure. The developer cites KDOQI and the Institute of Medicine on the importance of functional assessment, however no peer-reviewed, empirical evidence is provided that the specifications (i.e., annual completion rate) are associated with higher quality.
- **PERFORMANCE GAP.** Based on the updated specifications, the performance range in 2015 was 16.7%-100%, with a median of 91.8% using "KDQOL-Complete" (K-C) data. Although the performance rate at the *patient-level* with the updated exclusion criteria (i.e., refusals = fail) is 84.8% (2015), 84.7% (2014), and 84.2% (2013), the performance rate with

refusals as an exclusion (old specifications) is 100% in 2013, 2014, 2015. KCP also further examined the data and notes the refusal exclusion appears stable over this period. We posit the change in specifications creates a gap where otherwise none exists, as well as puts the measure in conflict with the Conditions for Coverage.

- **VALIDITY.** KCP has two concerns about the measure's validity: the validity testing and the lack of risk adjustment.

The developer performed validity testing on a sample that included all patients – i.e., those who refused, those who completed the survey, and those who met the exclusion criteria. It assessed association of completion with patients' KDQOL scores (linear fixed models with the score for each of the five scales as the dependent variable and facility completion rate as the main independent variable). *The models adjusted for patient-level characteristics of age, gender, race, and diabetes.* Based on this, it appears the measure was not tested as specified. First, all patients were used, even those who qualify for exclusions. Second, associations were examined, but the models were adjusted for patient-level characteristics even though the measure itself is not adjusted. Performance on the measure cannot be asserted as being associated with better quality (the five KDQOL scales) if the measure as specified is not used.

The developer also notes, "This finding [association between completion and scores] is important because it is plausible that facilities with higher rates would be obtaining completed questionnaires from sicker patients, since it has been documented that individuals completing the QoL scores tend to be younger and healthier." Again, the developer draws this conclusion from analyzing a different data set and a risk-adjusted model. The measure is not whether an all-population, risk-adjusted measure of completion validates against the scale results: Testing and demonstration of validity must be of the measure as specified.

Finally, KCP has expressed concern about NQF 0260 in other contexts (e.g., use in CMS' Comprehensive ESRD Care Initiative) because of the lack of risk adjustment for case mix. In fact, the developer's data demonstrate that case mix impacts a facility's score. Specifically, the developer presents data on the distribution of patient characteristics and the facility-level survey completion rate; the analysis uses refusals and completions, so comports with the proposed specifications. Facilities with more males will score, on average, 0.45% lower (per 10% difference) compared to facilities that have fewer males. Conversely, facilities with higher proportions of Asians – likely to exist in certain geographic areas – will score higher. We believe the lack of adjustment for the measure presents a significant threat to validity, particularly given a median performance of 91.8% with the updated specifications.

#### **NQF 0369: Dialysis Facility Risk-Adjusted Standardized Mortality Ratio (SMR; CMS)**

KCP believes mortality is an important outcome to measure, but has concerns about the specifications, reliability, validity (risk model), and harmonization issues.

- **SPECIFICATIONS.** The specifications for the time period state "at least one year." KCP believes specifications should be unambiguous, so this construction is imprecise. We believe the time period should be an exact period, and we further believe the 1-year period is inappropriate based on the reliability testing data and, at minimum, should be a 4-year period.

As we discuss further in the following section, KCP has significant concerns about the SMR's reliability for small- and medium-sized facilities. The SMR specifications do not

address a minimum sample size by excluding facilities of “x” or fewer patients, as we are aware other measures do.

The specifications do not exclude incident hospice patients. The NQF’s Measure Applications Partnership (MAP) recently did not recommend the SMR, in part because the measure did not exclude patients who are already in hospice when they initiate dialysis. During the deliberations, it was noted that occasionally incident patients begin dialysis treatments while in hospice, but then choose to discontinue them after a period of time. KCP supports MAP’s recommendation that patients who initiate dialysis while also in hospice be excluded from the SMR. As currently constructed, such patients are attributed to the facility providing the dialysis.

The SMR documentation indicates at least three expected deaths must occur for inclusion in the SMR calculations, but no justification or empirical analyses are offered to justify this threshold – e.g., how many clinics were excluded using this approach and what is the impact on scoring because of the exclusion?

Finally, the SMR specifications indicate the measures can be expressed as a rate, but is calculated as a ratio. KCP prefers normalized rates or year-over-year improvement in rates instead of a standardized ratio. We believe comprehension, transparency, and utility to all stakeholders is superior with a scientifically valid *rate* methodology. We note that MAP also did not support the SMR because, in addition to the lack of a hospice exclusion, MAP felt “mortality rates would be more meaningful to consumers and actionable for facilities.”

- **RELIABILITY.** Based on the testing results, KCP has serious concerns about the SMR’s reliability. We note a reliability statistic of 0.70 is often considered as “good” reliability,<sup>1</sup> though we recognize the characterization also depends on the analytic method. Testing results for the 1-year SMR yielded IURs of 0.26-0.32 for each of 2010, 2011, 2012, and 2013 – a low degree of reliability, where only about 30% of the variation in a score can be attributed to between-facility differences, yet the specifications permit this 1-year measure. The 4-year SMR yielded an IUR of 0.66 for 2009-2012 and only 0.59 for 2010-2013 data. Even with the 4-year SMR, less than 60% of a facility’s score is attributable to between-facility differences for the overall sample. Moreover, 4-year SMR testing results specifically for small- and medium-sized facilities indicate very poor reliability, with IURs of 0.30 and 0.45, respectively. Only large facilities have a reasonable IUR of 0.73 for 2010-2013 data. As noted earlier, KCP thus believes the specifications must specifically require a minimum sample as identified through the developer’s empirical testing.
- **VALIDITY.** KCP has strongly advocated for the use of prevalent co-morbidities in the SMR’s risk model, and commends the developer for moving to incorporate prevalent co-morbidities in the specifications. We continue to be concerned about the validity of the Medical Evidence Form (CMS 2728) as a data source for incident co-morbidities, however, and urge that the Committee recommend that CMS assess this matter.

In previous comments to CMS, KCP noted that many of the prevalent co-morbidities in the final model had p-values significantly greater than 0.05 – e.g., paralytic ileus (p=0.5007), episodic mood disorder NOS (p=0.8254). CMS responded that these were included because: “Most of the coefficient estimates for the prevalent co-morbidities are

---

<sup>1</sup> Adams, JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009.

positive and statistically significant, but several do not obtain statistical significance. The very large number of clinical factors in the model expectedly generates multicollinearity among co-variables, likely resulting in some unexpected results in direction of coefficient sign and levels of statistical significance. Inclusion of this set of prevalent co-morbidities reflects the consensus of the TEP that adjustment for all of these prevalent co-morbidities, in addition to incident co-morbidities, is important to reflect the initial and current health condition of the patient in risk adjustment.”

We do not believe this approach is sufficient. Our conversations with TEP members for the SMR/SHR indicate they did not advocate for model building in a vacuum without accounting for the meaning of the coded co-morbid conditions, but rather for including as many co-morbid conditions as possible. This is a very different interpretation than is offered by the developer’s explanation and far more appropriate when dealing with administrative coding habits that are not static over time. It may require, for example, grouping certain individual codes together to develop a more appropriate overarching description of true co-morbidity burden.

KCP is concerned that the strategy adopted for the SMR (and SHR) results in a model that will not be generalizable. Currently, for example, having thyroid cancer is protective to the same magnitude that atrial fibrillation is harmful. This makes no sense, and we posit is a function of collinearity and coding idiosyncrasy. Similarly, in the current model, osteomyelitis NOS-ankle is associated with a lower risk of death while ulcer of lower limb NOS is harmful. In actual medical practice, osteomyelitis is far worse than an ulcer of the lower limb. In the current model, lower extremity amputation is protective while ‘status amput below knee’ is harmful. Again, KCP supports prevalent co-morbidity adjustment, but we are concerned that the proposed collection of adjusters will be less robust with each year that passes from initial model development.

KCP also notes that while the SMR applies to all patients, the current list of co-morbidities does not account for those that may be unique to pediatrics. We recommend the Standing Committee suggest to the developer that such should be considered and included when indicated.

KCP also notes that the validity testing yielded a c-statistic for the SMR of 0.724. We are concerned the model will not adequately discriminate performance – particularly that smaller units, including pediatric units, might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model’s goodness of fit and validity to represent meaningful differences among facilities and encourage continuous improvement of the model.

Information on the risk model states that determination of a prevalent co-morbidity requires at least two outpatient claims or one inpatient claim, but no justification or empirical analyses are offered to support this algorithm over other approaches. We are aware this approach has been validated for diabetes,<sup>2</sup> but we are not that it has been validated for the large number of other co-morbidities or is broadly generalizable.

Finally, the risk model includes ambiguous language. The submission indicates patient characteristics included in the stage 1 model include “nursing home status in *previous* year.” It is unclear if this means patients moving into a nursing home for the first time *during* the measurement year would not be adjusted for “nursing home status.”

---

<sup>2</sup> Hebert PL, Geiss LS, et al. Identifying persons with diabetes using Medicare claims data. *Am J Med Qual.* 1999;14(4):270-277.

Specifically, it is unclear as to whether the look-back is *one year prior to the given event* (inclusive of the data year) or if this verbiage means the look-back is *in the previous calendar year* (not inclusive of the data year). KCP believes such ambiguity should be addressed and that the current reporting year be included, not just the previous one.

- **HARMONIZATION ISSUES.** The risk models for the groupings used for patient age and duration of ESRD differ among the SMR, SHR, and STrR. For example, the age groups for the SMR is n=3, but for the SHR and STrR the age groupings are the same, but n=6. Similarly, the number of groups for ESRD duration for the SMR (n=4) differs from that for the SHR (n=6). No justification or empirical analyses are offered to justify these differences.

There also are significant inconsistencies in how facility size is defined when assessing reliability for the SMR, SHR, and STrR. Specifically, for the SMR, the definitions were <=45, 46-85, >=86 for the 1-year reliability analyses, but were <=135, 136-305, and >=306 for the 4-year analyses. For the SHR, <=50, 51-87, and >=88 were used. Finally, for STrR reliability analyses, small, medium, and large facilities were defined as <=46, 47-78, and >=79, respectively. We understand reliability for a given measure depends on sample size, but find the varying demarcations analytically troubling. We posit a more appropriate analytic approach would be to analyze reliability using consistent “bins” of size (i.e., small, medium, and large are consistently defined) and identify the facility size at which reliability for that particular measure can be confidently inferred – and then reflect the minimum size in the actual specifications.

#### **NQF 1463: Standardized Hospitalization Ratio for Admissions (SHR; CMS)**

KCP believes hospitalization is an important outcome to measure, but has concerns about the specifications, reliability, validity (risk model), and harmonization issues. Many of our comments have been articulated in the context of those we make on the SMR, but owing to the NQF’s electronic portal for measure-by-measure comments, we repeat them for the SHR.

- **SPECIFICATIONS.** KCP has strongly advocated for the use of prevalent co-morbidities in the SHR’s risk model, and commends the developer for moving to incorporate prevalent co-morbidities in the specifications. We continue to be concerned about the validity of the Medical Evidence Form (CMS 2728) as a data source for incident co-morbidities, however, and urge that the Committee recommend that CMS assess this matter.

The SHR specifications for the time period also state “at least one year.” Again, as a principle, KCP believes specifications should be unambiguous. We believe the time period should be an exact period.

As we discuss in the reliability section, KCP has significant concerns about the reliability of the 1-year SHR for small and medium facilities. The SHR specifications do not address a minimum sample size by excluding facilities of “x” or fewer patients, as we are aware other measures do.

Documentation indicates the minimum data requirement for the SHR is 5 patient-years at risk, which differs from the STrR, which uses 10 patient-years at risk. No justification or empirical analyses are offered to justify the selected threshold or the difference.

Finally, the SHR specifications indicate the measure can be expressed as a rate, but is calculated as a ratio. KCP prefers normalized rates or year-over-year improvement in rates instead of a standardized ratio. We believe comprehension, transparency, and utility to all stakeholders is superior with a scientifically valid *rate* methodology.

- **RELIABILITY.** We again note a reliability statistic of 0.70 is often considered as “good” reliability, though we recognize the characterization also depends on the analytic method. Again, based on the results from the reliability testing, we have significant concerns about the reliability of the 1-year SHR for small and medium facilities (IUR range of 0.46-0.65, depending on the year). The SHR specifications do not address a minimum sample size by excluding facilities of “x” or fewer patients, as we are aware other measures do. As noted earlier, KCP thus believes the specifications must specifically require a minimum sample as identified through the developer’s empirical testing.
- **VALIDITY.** KCP has strongly advocated for the use of prevalent co-morbidities in the SHR’s risk model, and commends the developer for moving to incorporate prevalent co-morbidities in the specifications. We continue to be concerned about the validity of the 2728 as a data source for incident co-morbidities, however, and urge that the Committee recommend that CMS assess this matter.

In previous comments to CMS, KCP noted that many of the prevalent co-morbidities in the final model had p-values significantly greater than 0.05 – e.g., paralytic ileus (p=0.5007), episodic mood disorder NOS (p=0.8254). CMS responded that these were included because: “Most of the coefficient estimates for the prevalent co-morbidities are positive and statistically significant, but several do not obtain statistical significance. The very large number of clinical factors in the model expectedly generates multi-collinearity among co-variates, likely resulting in some unexpected results in direction of coefficient sign and levels of statistical significance. Inclusion of this set of prevalent co-morbidities reflects the consensus of the TEP [Technical Expert Panel] that adjustment for all of these prevalent co-morbidities, in addition to incident co-morbidities, is important to reflect the initial and current health condition of the patient in risk adjustment.”

We do not believe this approach is sufficient. Our conversations with TEP members indicate they did not advocate for model building in a vacuum without accounting for the meaning of the coded co-morbid conditions, but rather for including as many co-morbid conditions as possible. This is a very different interpretation than is offered by the developer’s explanation and more appropriate when dealing with administrative coding habits that are not static over time. It may require, for example, grouping certain individual codes together to develop an appropriate overarching description of true co-morbidity burden.

KCP is concerned the strategy adopted for the SHR (and SMR) results in a model that will not be generalizable. Currently, for example, having thyroid cancer is protective to the same magnitude that atrial fibrillation is harmful. This makes no sense, and we posit is a function of collinearity and coding idiosyncrasy. Similarly, in the current model osteomyelitis NOS-ankle is associated with a lower risk of death, while ulcer of lower limb NOS is harmful. In actual medical practice, osteomyelitis is far worse than an ulcer of the lower limb. In the current model, lower extremity amputation is protective while ‘status amput below knee’ is harmful. Again, KCP supports prevalent co-morbidity adjustment, but we are concerned that the proposed collection of adjusters will be less robust with each year that passes from initial model development.

KCP also notes that the validity testing yielded an overall c-statistic for the SHR of 0.65. We are concerned the model will not adequately discriminate performance – particularly that smaller units might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model’s goodness of fit and validity



to represent meaningful differences among facilities and encourage continuous improvement of the model.

Information on the risk model states that determination of a prevalent co-morbidity requires at least two outpatient claims or one inpatient claim, but no justification or empirical analyses are offered to support this algorithm over other approaches. As noted for the SMR, we are aware this approach has been validated for diabetes, but we are not that it has been validated for the large number of other co-morbidities or is broadly generalizable.

Finally, the risk model includes ambiguous language. The submission indicates patient characteristics included in the stage 1 model include “nursing home status in *previous* year.” It is unclear if this means patients moving into a nursing home for the first time *during* the measurement year would not be adjusted for “nursing home status.” Specifically, it is unclear as to whether the look-back is *one year prior to the given event* (inclusive of the data year) or if this verbiage means the look-back is *in the previous calendar year* (not inclusive of the data year). KCP believes such ambiguity should be addressed and that the current reporting year be included, not just the previous one.

- **HARMONIZATION ISSUES.** The risk models for the groupings used for patient age and duration of ESRD differ among the SMR, SHR, and STrR. For example, the age groups for the SMR is n=3, but for the SHR and STrR the age groupings are the same, but n=6. Similarly, the number of groups for ESRD duration for the SMR (n=4) differs from that for the SHR (n=6). No justification or empirical analyses are offered to justify these differences.

There also are significant inconsistencies in how facility size is defined when assessing reliability for the SMR, SHR, and STrR. Specifically, for the SMR, the definitions were <=45, 46-85, >=86 for the 1-year reliability analyses, but were <=135, 136-305, and >=306 for the 4-year analyses. For the SHR, <=50, 51-87, and >=88 were used. Finally, for STrR reliability analyses, small, medium, and large facilities were defined as <=46, 47-78, and >=79, respectively. We understand reliability for a given measure depends on sample size, but find the varying demarcations analytically troubling. We posit a more appropriate analytic approach would be to analyze reliability using consistent “bins” of size (i.e., small, medium, and large are consistently defined) and identify the facility size at which reliability for that particular measure can be confidently inferred – and then reflect the minimum size in the actual specifications.

#### **NQF 2977: Hemodialysis Vascular Access: Standardized Fistula Rate**

As with the catheter measure, KCP used the existing arteriovenous fistula (AVF) measure, NQF 0257, for context in our review.

- **SPECIFICATIONS.** The language in #0257 that specifically defines an autogenous AVF as using two needles has been replaced with an autogenous AVF “as the sole means of vascular access.” KCP believes the specifications are imprecise as to whether facilities would receive credit for patients using an AVF as the sole means of access, but who also have in place a graft or catheter that is no longer being used. We note patients with catheters remain at risk for infection and other adverse sequelae, so credit should not be not given when a catheter is present, even if an AVF is being used. A numerator that specifies the patient must be on maintenance hemodialysis “using an AVF *with two needles and without a dialysis catheter present*” would remove ambiguity. In contrast, removal of an AV graft is complex and not without risk of complications, so KCP

believes credit should be received for a patient who is using an AVF as the sole means of access, but who also may have a non-functioning AV graft present.

KCP notes the 90-day ESRD requirement has been removed from the denominator statement as compared to #0257, which means the “clock” for the measure starts on the first day of dialysis in a non-hospital setting – but that the permitted timeframe for catheter use in the numerator is still 90 days; we support this change. Additionally, we commend the developer for adding an exclusion for patients with limited life expectancy and for now unambiguously identifying the four subcategories, both approaches that KCP had recommended.

- **VALIDITY.** KCP believes this measure improves on #0257 and commends the developer for accepting KCP’s recommendation in previous comments to remove the covariate alcohol dependence from the model’s risk variables. We continue to believe two additional vasculature risk variables would strengthen the model: a history of multiple prior accesses and the presence of a cardiac device.

KCP notes that the validity testing yielded an overall c-statistic of 0.71. We are concerned the model will not adequately discriminate performance – particularly that smaller units might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model’s goodness of fit and validity to represent meaningful differences among facilities and encourage continuous improvement of the model.

#### **NQF 2978: Hemodialysis Vascular Access: Long-Term Catheter Rate**

As with the AVF measure, KCP used the existing catheter measure, NQF 0256, for context in our review.

- **SPECIFICATIONS.** As with the AVF measure, KCP notes the 90-day ESRD requirement has been removed from the denominator statement as compared to #0256, which means the “clock” for the measure starts on the first day of dialysis in a non-hospital setting – but that the permitted timeframe for catheter use in the numerator is still 90 days; we support this change. Additionally, we commend the developer for adding an exclusion for patients with limited life expectancy and for now unambiguously identifying the four subcategories, both approaches that KCP had recommended.

#### **NQF 2979: Standardized Transfusion Ratio for Dialysis Facilities (STrR; CMS)**

During the last project, this Standing Committee reviewed the STrR as #2699 and did not recommend it. As we discuss further in the section on Validity, we do not believe the new measure addresses the Committee’s concerns about hospital- and physician-related factors. We comment on the specifications, reliability, validity (risk model), and harmonization issues.

- **SPECIFICATIONS.** CMS has revised the measure specifications to more “conservatively” define transfusion events, such that all inpatient transfusion events must include, at a minimum, an appropriate ICD-9 Procedure Code or Value Code to be captured in the measure – inpatient transfusion events for claims that include only 038 or 039 revenue codes without an accompanying procedure or value code are not captured in the numerator. The specifications also specify a maximum of one event per day and that an event not be defined by the number of units of blood transfused.

KCP supports and appreciates the need to refine and tighten how transfusion events are counted and applauds CMS’s intent in undertaking these revisions, but we do not

believe the proposed solution is a valid representation of transfusion events. Importantly, there is no existing coding requirement that procedure or value codes be used, which means valid transfusion claims that include only revenue codes will be missed. KCP believes the proposed specification changes result in a measure with significant threats to validity.

Current transfusion coding practices clearly vary by hospital,<sup>3</sup> and hospital coding practices are beyond dialysis facilities' sphere of control. For example, we are aware of hospitals that exclusively use revenue codes and do not use the procedure or value codes. In-patients at this type of hospital will appear to have no transfusion events assigned to the dialysis facility, whereas those at a hospital that uses the procedure and/or value codes will have recorded events. Simply put, facilities within given catchment areas will be differentially affected by hospital coding variations, which clearly impact measure scoring. We are particularly concerned that the revisions, if implemented, will result in increased variability in performance across dialysis facilities wholly due to external factors and not performance. Facilities will appear to have "poor" performance because of higher than expected numbers of transfusions – and will expend time and resources to improve – when in fact the score is merely a reflection of coding practices.

Again, KCP strongly supports the need to refine how transfusion events are defined, and we urge the Standing Committee to recommend the developer continue considering alternative models to define transfusion events. Alternatively, the Committee could suggest that CMS consider revising hospital transfusion coding rules to require that the ICD-9/ICD-10 procedure and value codes necessary for the validity of the proposed methodology be universally included in claims.

Additionally, the testing documentation notes that facilities with 10 or fewer patients were excluded, but we note the specifications do not state this. Again, KCP believes that a minimum size exclusion should be indicated and, as the developer's results document, and we discuss in the following section, reliability is poor even when the facility size is significantly greater than 10 patients.

The submission also indicates the minimum data requirement for the STrR is 10 patient-years at risk, which differs from the SHR, which uses 5 patient-years at risk. No justification or empirical analyses are offered to justify the selected threshold or the difference.

Finally, the STrR specifications indicate the measure can be expressed as a rate, but is calculated as a ratio. KCP prefers normalized rates or year-over-year improvement in rates instead of a standardized ratio. We believe comprehension, transparency, and utility to all stakeholders is superior with a scientifically valid *rate* methodology.

- **RELIABILITY.** In addition to our concerns that the specifications pose a threat to the validity of the updated STrR, KCP also has concerns about the reliability testing for these revised specifications.

KCP again notes a reliability statistic of 0.70 is often considered as "good" reliability, though the characterization also depends on the analytic method. Reliability testing, overall, for the STrR yielded IURs of 0.60-0.66 across all facilities for each of 2011, 2012, 2013, and 2014. Such values indicate about 65% of the variation in a score can be

---

<sup>3</sup> Weinhandl ED, Gilbertson DT, Collins AJ. Dialysis facility-level transfusion rates can be unreliable due to variability in hospital-level billing patterns for blood. Chronic Disease Research Group poster, ASN. 2014.

attributed to between-facility differences (signal) and about 35% to within-facility differences (noise) – a moderate degree of reliability. However, when looking exclusively at small (defined as  $\leq 46$ ) and medium (47-78) facilities, the IURs are substantially lower. Specifically, the IURs ranged from 0.30-0.41 and 0.50-0.56 for small and medium facilities, respectively, over the same time period. As noted earlier, KCP thus believes the specifications must specifically require a minimum sample as identified through the developer’s empirical testing.

- **VALIDITY.** In addition to KCP’s concerns about the specifications and the threat to validity of variable capture of transfusion events depending on hospital coding practices, KCP has several concerns about the co-variates (or lack thereof) and risk model.

NQF did not endorse the STrR in 2015, in part because this Standing Committee raised concern that the measure did not adjust for hospital- and physician-related transfusion practices. Physicians independently, or following hospital protocols, make decisions about whether or not to transfuse a specific patient, so it is important to account for the variability these factors create. The revised measure does not incorporate these factors into the risk model, so KCP’s concurrence with the Committee’s original concern remains.

KCP notes that while the SMR and SHR have been revised to incorporate prevalent co-morbidities into their risk models, the STrR has not been so revised; only incident co-morbidities, derived from the Medical Evidence Form (CMS 2728), are considered. This approach means the STrR risk model only reflects those conditions present upon when the patient initiates dialysis; failure to appropriately account for prevalent co-morbidities is a threat to validity. In the harmonization section, we also note that CMS adjusts for 2728-derived co-morbidities for SHR and SMR differently than it does for the STrR. Finally, as we have noted before, we continue to be concerned about the validity of the 2728 as a data source and urge that the Committee recommend that CMS assess this matter.

KCP notes that the validity testing yielded an overall c-statistic of 0.65. We are concerned the model will not adequately discriminate performance – particularly that smaller units might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model’s goodness of fit and validity to represent meaningful differences among facilities and encourage continuous improvement of the model.

- **HARMONIZATION ISSUES.** The new SMR and SHR risk models adjust for each incident co-morbidity (from the 2728) separately, instead of using a “co-morbidity index.” The model also approaches diabetes as a single co-morbidity rather than four separate indicators (currently on insulin, on oral medications, without medications, diabetic retinopathy). The STrR has not been similarly revised. KCP believes the Standing Committee should recommend that the developer harmonize the STrR with the other measures so that each incident co-morbidity is examined separately (i.e., unbundled, as compared to the current measure) and diabetes is approached as a single co-morbidity (i.e., bundled, as compared to the current risk model).

The risk models for the groupings used for patient age and duration of ESRD differ among the SMR, SHR, and STrR. For example, the age groups for the SMR is  $n=3$ , but for the SHR and STrR the age groupings are the same, but  $n=6$ . Similarly, the number of groups for ESRD duration for the SMR ( $n=4$ ) differs from that for the SHR ( $n=6$ ). No justification or empirical analyses are offered to justify these differences.

There also are significant inconsistencies in how facility size is defined when assessing reliability for the SMR, SHR, and STrR. Specifically, for the SMR, the definitions were  $\leq 45$ , 46-85,  $\geq 86$  for the 1-year reliability analyses, but were  $\leq 135$ , 136-305, and  $\geq 306$  for the 4-year analyses. For the SHR,  $\leq 50$ , 51-87, and  $\geq 88$  were used. Finally, for STrR reliability analyses, small, medium, and large facilities were defined as  $\leq 46$ , 47-78, and  $\geq 79$ , respectively. We understand reliability for a given measure depends on sample size, but find the varying demarcations analytically troubling. We posit a more appropriate analytic approach would be to analyze reliability using consistent “bins” of size (i.e., small, medium, and large are consistently defined) and identify the facility size at which reliability for that particular measure can be confidently inferred – and then reflect the minimum size in the actual specifications.

KCP again thanks you for the opportunity to comment on this important work. If you have any questions, please do not hesitate to contact Lisa McGonigal, MD, MPH (lmcgon@msn.com or 203.298.0567).

Sincerely,

AbbVie  
Akebia  
American Kidney Fund  
American Nephrology Nurses Association  
American Renal Associates  
American Society of Nephrology  
American Society of Pediatric Nephrology  
Amgen  
Astra Zeneca  
Baxter  
Board of Nephrology Examiners Nursing Technology  
Centers for Dialysis Care  
DaVita  
Dialysis Clinic, Inc.  
Dialysis Patient Citizens  
Fresenius Medical Care  
Fresenius Medicare Care Renal Therapies  
Greenfield Health Systems  
Keryx  
Kidney Care Council  
National Renal Administrators Association  
Nephrology Nursing Certification Commission  
Northwest Kidney Centers  
NxStage Medical  
Renal Physicians Association  
Rogosin Institute  
Sanofi  
Satellite Healthcare  
U.S. Renal Care



August 31, 2016

National Quality Forum  
1030 Fifteenth Street, NW, Ste 800  
Washington, DC 20005

RE: NQF Renal Project

Kidney Care Partners (KCP) appreciates the opportunity to comment on the measures under consideration for endorsement in the National Quality Forum's (NQF) Renal Measures 2015-2017 Project. KCP is a coalition of members of the kidney care community that includes the full spectrum of stakeholders related to dialysis care – patient advocates, health care professionals, dialysis providers, researchers, and manufacturers and suppliers – organized to advance policies that improve the quality of care for individuals with both chronic kidney disease and end stage renal disease.

We commend the NQF Renal Standing Committee for its thoughtful deliberations. KCP supports and recognizes the importance and value of NQF's endorsement process to ensure the importance, reliability and validity of measures. Implementing parsimonious sets of measures that matter is becoming of increasing importance, making NQF's process more critical. We offer comment on all six measures.

**NQF 0260: Assessment of Health-Related Quality of Life (QoL) in Dialysis Patients (Witten and Associates, LLC)**

KCP supports the Committee's recommendation against endorsement.

**NQF 0369: Dialysis Facility Risk-Adjusted Standardized Mortality Ratio (SMR; CMS)**

KCP supports the Committee's recommendation against endorsement.

**NQF 1463: Standardized Hospitalization Ratio for Admissions (SHR; CMS)**

KCP believes hospitalization is an important outcome to measure, but has concerns the specifications, reliability, validity (risk model), and harmonization issues. We strongly encourage the Committee to reconsider the reliability testing data, which demonstrate significant reliability issues with the one-year SHR for small facilities, and comment specifically on the SHR's reliability for such facilities.

**IUR for One-year SHR, Overall and by Facility Size, 2010-2013**

	2010		2011		2012		2013	
Facility Size (Number of patients)	IUR	N	IUR	N	IUR	N	IUR	N
All	0.72	5407	0.71	5583	0.70	5709	0.70	5864
Small (<=50)	0.54	1864	0.51	1921	0.48	1977	0.46	2028
Medium (51-87)	0.65	1702	0.63	1785	0.58	1825	0.57	1930
Large (>=88)	0.81	1841	0.81	1877	0.81	1907	0.82	1906



Although the overall reliability statistic for 2013 (and previous years) is 0.7, a level generally considered the minimum by NQF, the reliability statistics for medium and small facilities fall significantly short of the 0.7 threshold. CMS's own data indicate that for facilities  $\leq 50$  patients, more than half a facility's score (54%) is due to random noise and not a signal of quality. Even for medium facilities, the IUR is significantly below the 0.7 threshold, with 43% of a facility's score attributable to random noise and not signal. We note that the intended use for the SHR will be for public reporting and the penalty-based QIP; penalizing facilities for performance due to random chance is not appropriate. Given the poor reliability testing results, KCP did not support CMS's proposal to include it in the Quality Incentive Program (QIP) for Payment Year 2020.

#### **NQF 2977: Hemodialysis Vascular Access: Standardized Fistula Rate**

KCP recommends the developer consider modifications to improve the measure going forward.

- With respect to the specifications, the language in the previously endorsed AVF measure (#0257) specifically defines an autogenous AVF as using two needles has been replaced with an autogenous AVF "as the sole means of vascular access." KCP believes the specifications for #2977 are imprecise as to whether facilities would receive credit for patients using an AVF as the sole means of access, but who also have in place a graft or catheter that is no longer being used. We note patients with catheters remain at risk for infection and other adverse sequelae, so credit should not be given when a catheter is present, even if an AVF is being used. A numerator that specifies the patient must be on maintenance hemodialysis "using an AVF *with two needles and without a dialysis catheter present*" would remove ambiguity. In contrast, removal of an AV graft is complex and not without risk of complications, so KCP believes credit should be received for a patient who is using an AVF as the sole means of access, but who also may have a non-functioning AV graft present.
- KCP believes this measure improves on #0257, but we continue to believe two additional vasculature risk variables would strengthen the model: a history of multiple prior accesses and the presence of a cardiac device. We also note that the validity testing yielded an overall c-statistic of 0.71. We are concerned the model will not adequately discriminate performance – particularly that smaller units might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model's goodness of fit and validity to represent meaningful differences among facilities and encourage continuous improvement of the model.

#### **NQF 2978: Hemodialysis Vascular Access: Long-Term Catheter Rate**

KCP supports the Committee's recommendation for endorsement.

#### **NQF 2979: Standardized Transfusion Ratio for Dialysis Facilities (STrR; CMS)**

During the last project, this Standing Committee reviewed the STrR as #2699 and did not recommend it. As we discuss further in the section on Validity, we do not believe the new measure addressed the Committee's concerns about hospital- and physician-related factors. Overall, we remain concerned about the reliability, as well as the specifications and validity. We strongly encourage the Committee to reconsider the reliability testing data, which document reliability issues with the STrR for small facilities, and comment specifically on the STrR's reliability for such facilities.

- **RELIABILITY.** KCP has significant concerns about the results from the reliability testing for the STrR. KCP notes a reliability statistic of 0.70 is often considered as "good"

reliability, though the characterization also depends on the analytic method. Reliability testing, overall, for the STrR yielded IURs of 0.60-0.66 across all facilities for each of 2011, 2012, 2013, and 2014. Such values indicate about 65% of the variation in a score can be attributed to between-facility differences (signal) and about 35% to within-facility differences (noise) – a moderate degree of reliability. However, when looking exclusively at small (defined as  $\leq 46$ ) and medium (47-78) facilities, the IURs are substantially lower. Specifically, the IURs ranged from 0.30-0.41 and 0.50-0.56 for small and medium facilities, respectively, over the same time period. In other words, approximately 60-70% of a small facility's score is due to random noise. KCP believes the specifications must specifically require a minimum sample as identified through the developer's empirical testing.

- **SPECIFICATIONS.** CMS has revised the measure specifications to more “conservatively” define transfusion events, such that all inpatient transfusion events must include, at a minimum, an appropriate ICD-9 Procedure Code or Value Code to be captured in the measure – inpatient transfusion events for claims that include only 038 or 039 revenue codes without an accompanying procedure or value code are not captured in the numerator. The specifications also specify a maximum of one event per day and that an event not be defined by the number of units of blood transfused.

KCP supports and appreciates the need to refine and tighten how transfusion events are counted and applauds CMS's intent in undertaking these revisions, but we do not believe the proposed solution is a valid representation of transfusion events. Importantly, there is no existing coding requirement that procedure or value codes be used, which means valid transfusion claims that include only revenue codes will be missed. KCP believes the proposed specification changes result in a measure with significant threats to validity.

Current transfusion coding practices clearly vary by hospital,<sup>1</sup> and hospital coding practices are beyond dialysis facilities' sphere of control. For example, we are aware of hospitals that exclusively use revenue codes and do not use the procedure or value codes. In-patients at this type of hospital will appear to have no transfusion events assigned to the dialysis facility, whereas those at a hospital that uses the procedure and/or value codes will have recorded events. Simply put, facilities within given catchment areas will be differentially affected by hospital coding variations, which clearly impact measure scoring. We are particularly concerned that the revisions, if implemented, will result in increased variability in performance across dialysis facilities wholly due to external factors and not performance. Facilities will appear to have “poor” performance because of higher than expected numbers of transfusions – and will expend time and resources to improve – when in fact the score is merely a reflection of coding practices.

Again, KCP strongly supports the need to refine how transfusion events are defined, and we urge the Standing Committee to recommend the developer continue considering alternative models to define transfusion events. Alternatively, the Committee could suggest that CMS consider revising hospital transfusion coding rules to require that the ICD-9/ICD-10 procedure and value codes necessary for the validity of the proposed methodology be universally included in claims.

---

<sup>1</sup> Weinhandl ED, Gilbertson DT, Collins AJ. Dialysis facility-level transfusion rates can be unreliable due to variability in hospital-level billing patterns for blood. Chronic Disease Research Group poster, ASN. 2014.

Additionally, the testing documentation notes that facilities with 10 or fewer patients were excluded, but we note the specifications do not state this. Again, KCP believes that a minimum size exclusion should be indicated and, as the developer's results document, and we discuss in the following section, reliability is poor even when the facility size is significantly greater than 10 patients.

The submission also indicates the minimum data requirement for the STrR is 10 patient-years at risk, which differs from the SHR, which uses 5 patient-years at risk. No justification or empirical analyses are offered to justify the selected threshold or the difference.

Finally, the STrR specifications indicate the measure can be expressed as a rate, but is calculated as a ratio. KCP prefers normalized rates or year-over-year improvement in rates instead of a standardized ratio. We believe comprehension, transparency, and utility to all stakeholders is superior with a scientifically valid *rate* methodology.

- **VALIDITY.** In addition to KCP's concerns about the specifications and the threat to validity of variable capture of transfusion events depending on hospital coding practices, KCP has several concerns about the co-variates (or lack thereof) and risk model.

NQF did not endorse the STrR in 2015, in part because this Standing Committee raised concern that the measure did not adjust for hospital- and physician-related transfusion practices. Physicians independently, or following hospital protocols, make decisions about whether or not to transfuse a specific patient, so it is important to account for the variability these factors create. The revised measure does not incorporate these factors into the risk model, so KCP's concurrence with the Committee's original concern remains.

KCP notes that while the SMR and SHR have been revised to incorporate prevalent co-morbidities into their risk models, the STrR has not been so revised; only incident co-morbidities, derived from the Medical Evidence Form (CMS 2728), are considered. This approach means the STrR risk model only reflects those conditions present upon when the patient initiates dialysis; failure to appropriately account for prevalent co-morbidities is a threat to validity. In the harmonization section, we also note that CMS adjusts for 2728-derived co-morbidities for SHR and SMR differently than it does for the STrR. Finally, as we have noted before, we continue to be concerned about the validity of the 2728 as a data source and urge that the Committee recommend that CMS assess this matter.

KCP notes that the validity testing yielded an overall c-statistic of 0.65. We are concerned the model will not adequately discriminate performance – particularly that smaller units might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model's goodness of fit and validity to represent meaningful differences among facilities and encourage continuous improvement of the model.

- **HARMONIZATION ISSUES.** The new SMR and SHR risk models adjust for each incident co-morbidity (from the 2728) separately, instead of using a "co-morbidity index." The model also approaches diabetes as a single co-morbidity rather than four separate indicators (currently on insulin, on oral medications, without medications, diabetic retinopathy). The STrR has not been similarly revised. KCP believes the Standing Committee should recommend that the developer harmonize the STrR with the other measures so that each incident co-morbidity is examined separately (i.e., unbundled, as compared to the current measure) and diabetes is approached as a single co-morbidity

(i.e., bundled, as compared to the current risk model).

The risk models for the groupings used for patient age and duration of ESRD differ among the SMR, SHR, and STTrR. For example, the age groups for the SMR is n=3, but for the SHR and STTrR the age groupings are the same, but n=6. Similarly, the number of groups for ESRD duration for the SMR (n=4) differs from that for the SHR (n=6). No justification or empirical analyses are offered to justify these differences.

There also are significant inconsistencies in how facility size is defined when assessing reliability for the SMR, SHR, and STTrR. Specifically, for the SMR, the definitions were <=45, 46-85, >=86 for the 1-year reliability analyses, but were <=135, 136-305, and >=306 for the 4-year analyses. For the SHR, <=50, 51-87, and >=88 were used. Finally, for STTrR reliability analyses, small, medium, and large facilities were defined as <=46, 47-78, and >=79, respectively. We understand reliability for a given measure depends on sample size, but find the varying demarcations analytically troubling. We posit a more appropriate analytic approach would be to analyze reliability using consistent “bins” of size (i.e., small, medium, and large are consistently defined) and identify the facility size at which reliability for that particular measure can be confidently inferred – and then reflect the minimum size in the actual specifications.

KCP again thanks you for the opportunity to comment on this important work. If you have any questions, please do not hesitate to contact Lisa McGonigal, MD, MPH (lmcgon@msn.com or 203.298.0567).

Sincerely,

/s/  
Sara Love Rawlings, JD  
Executive Director

## CMS AND STANDING COMMITTEE RESPONSES TO KCP COMMENTS

### CMS Response

We respectfully disagree that STrR, a measure of transfusion avoidance, is required to be harmonized with a measure of hospitalization. Each metric is capturing different outcomes.

As described in the STrR measure submission, the measure adjusts for each separate incident comorbidity. See S.14 in the NQF MIF (excerpt below).

"Comorbidities at ESRD incidence are determined using a selection of comorbidities reported on the CMS-2728 namely, alcohol dependence, atherosclerotic heart disease, cerebrovascular disease, chronic obstructive pulmonary disease, congestive heart failure, diabetes (includes currently on insulin, on oral medications, without medications, and diabetic retinopathy), drug dependence, inability to ambulate, inability to transfer, malignant neoplasm, cancer, other cardiac disease, peripheral vascular disease, and tobacco use (current smoker). Each comorbidity is included as a separate covariate in the model."

The categories for the Age and Duration of ESRD covariates in the risk adjustment models were empirically derived when the SMR and SHR models were first developed, and are based on model fit specific to each outcome. This accounts for the use of different groupings for each model. The STrR was developed using an adaptation of the SHR methodology, and the age groupings were left in-tact.

Regarding the definition of facility size, we will consider using consistent groupings in the future, to improve interpretation. Thank you for the feedback.

During the most recent 2016 Standing Committee review of this measure, committee members discussed the shared accountability aspect of STrR. Literature evidence supporting the strong association between prior achieved hemoglobin and subsequent transfusion risk was reviewed. In addition, the committee was presented with RBC transfusion guidelines endorsed by the American Red Cross and other national organizations describing the central role of patient hemoglobin in determining need for RBC transfusion, while considering the clinical context of the transfusion decision. In addition, the committee reviewed additional recent peer review publication evidence describing the role of facility anemia management processes of care in predicting subsequent transfusion risk. Although there is some truth to the commenter's statement the "physicians independently, or following hospital protocols, make decisions about whether or not to transfuse a specific patient", the national recommendations and the physician decisions appear to be based in large part on the patient's hemoglobin.

In addition to their responsibility for anemia management and achieved hemoglobin, with its aforementioned contribution to determination of transfusion need, both dialysis providers and the nephrologist members of their Interdisciplinary Teams have an important responsibility to educate patients, their families, and other providers involved in the care of their patients about the potential unintended consequences of RBC transfusions in transplant-eligible dialysis patients. After this evidence review and discussion, the Standing Committee did recommend endorsement of STrR in 2016, as currently specified. We believe the Standing Committee made the correct decision in 2016.

There is no published research or study demonstrating the CMS 2728 data have been shown to be invalid. We acknowledge the 2728 data have been shown to be insensitive in a few studies. Inclusion of the 2728 data is one component of a more comprehensive risk adjustment strategy. The 2012 Anemia Management Technical Expert Panel recommended development of additional risk adjustment strategies that utilized prevalent comorbidities specifically related to conditions that would impact anemia management in ESRD patients. We utilize prevalent comorbidities as exclusions rather than covariates in the risk adjustment model to minimize the risk of underestimating their impact in the care of dialysis patients.

The STrR C-statistic of 0.65 is similar in magnitude to several other current NQF endorsed quality measures that have been endorsed by NQF and implemented by CMS in ESRD quality programs, as well as for other settings. This

level is considered to be a good fit as demonstrated in peer-reviewed studies reporting similar goodness of fit statistics for outcome based models (see accompanying list of references). As we refine the risk model in the future, we will work to improve the model's ability to discriminate performance between facilities.

1. Abbott K, Trespalacios F, and Agodoa L. Arteriovenous fistula use and heart disease in long-term elderly hemodialysis patients: Analysis of United States renal data system dialysis morbidity and mortality wave II. *J NEPHROL* 2003;16: 822-830.
2. Hurst et al. Arteriovenous Fistulas among Incident Hemodialysis Patients in Department of Defense and Veterans Affairs Facilities. *J Am Soc Nephrol*. 21: 1571–1577 2010.
3. Michael P. Lily et al. 2012. Prevalence of Arteriovenous Fistulas in Incident Hemodialysis Patients: Correlation with Patient Factors That May Be Associated With Maturation Failure. *American Journal of Kidney Diseases*. April 2012 Volume 59, Issue 4, Pages 541–549.
4. Lok C et al. Risk Equation Determining Unsuccessful Cannulation Events and Failure to Maturation in Arteriovenous Fistulas (REDUCE FTM I). *Am Soc Nephrol*. 17: 3204–3212, 2006.
5. Masengu A, Maxwell A, Hanko J. AVF Failure to mature Investigating clinical predictors of arteriovenous fistula functional patency in a European cohort. *Clinical Kidney Journal*, 2016, vol. 9, no. 1, 142–147. doi: 10.1093/ckj/sfv131.
6. Nee et al. Impact of Poverty and Health Care Insurance on Arteriovenous Fistula Use among Incident Hemodialysis Patients. *Am J Nephrol*. 2015;42:328–336.

Note: the issue regarding the definition of transfusion events (first paragraph) is addressed in another comment response.

This requirement for minimum number of patients/patient years at risk is not part of the measure specifications, but applied in the current implementation of the measure for DFC and for PY2018 QIP. The analyses in the Testing Form applied this requirement, in order to align with current public reporting standards.

Thank you for the comment about standardized ratios. As you noted, the ratio can be expressed as a rate and will be displayed as such on DFC beginning October 2016.

The definition of transfusion events used in the revised STeR measure is consistent with definitions used in numerous scientific publications, including several peer review publications by the research group that presented the ASN abstract referred to by the commenters [1-6]. The definition is also structurally consistent with Medicare claims processing rules. By excluding transfusion events identified only through revenue codes, the false positive identification of blood transfusions should be reduced, per the Medicare claims processing rules and guidelines published by the American Red Cross and other blood banking organizations. By definition, exclusion of revenue code only transfusion events decreases variation due to hospital coding practices that may rely primarily on revenue codes. We have empirically demonstrated this revision does not substantially alter the strong relationship between recent prior achieved hemoglobin and subsequent transfusion risk, a relationship that has been previously shown in other research studies. Furthermore, we are not aware of any scientific publication demonstrating that the definition of transfusion events used in this revised measure is invalid. It should be noted that this issue was also discussed in detail during the ESRD Standing Committee's discussion of the STeR at the June, 2016 in-person meeting, prior to the ESRD Standing Committee vote to recommend the measure for endorsement.

1. Hirth, Turenne, Wilk et al. Blood transfusion practices in dialysis patients in a dynamic regulatory environment. *Am J Kidney Dis*. 2014 Oct;64(4):616-21. doi: 10.1053/j.ajkd.2014.01.011. Epub 2014 Feb 19.

2. Gilbertson, Monda, Bradbury & Collins. RBC Transfusions Among Hemodialysis Patients (1999-2010): Influence



of Hemoglobin Concentrations Below 10 g/dL. Am J Kidney Dis. 2013; Volume 62 , Issue 5 , 919 - 928

3. Collins et al. Effect of Facility-Level Hemoglobin Concentration on Dialysis Patient Risk of Transfusion. Am J Kidney Dis. 2014; 63(6):997-1006.

4. Cappell et al. Red blood cell (RBC) transfusion rates among US chronic dialysis patients during changes to Medicare end-stage renal disease (ESRD) reimbursement systems and erythropoiesis stimulating agent (ESA) labels. BMC Nephrology. 2014, 15:116.

5. Ibrahim, et al. Blood transfusions in kidney transplant candidates are common and associated with adverse outcomes. Clin Transplant. 2011: 25: 653–659.

6. Molony, et al. Effects of epoetin alfa titration practices, implemented after changes to product labeling, on hemoglobin levels, transfusion use, and hospitalization rates. Am J Kidney Dis. 2016: e pub before print (published online March 12, 2016).

+++++

**NQF Staff Proposed Response to KCP from Renal Standing Committee**

The Committee thoroughly reviewed the specifications, reliability, and validity of the measure during the in-person and maintains that the measure meets the NQF criteria.