NATIONAL QUALITY FORUM
Driving measurable health improvements together

Memo

September 22, 2020

To: Renal Standing Committee

From: NQF staff

Re: Post-comment web meeting to discuss public comments received and NQF member expression of support

Introduction

NQF closed the public commenting period on the measures submitted for endorsement consideration to the Spring 2020 measure review cycle on August 25, 2020.

Purpose of the Call

The Renal Standing Committee will meet via web meeting on September 22, 2020 from 12:00 pm to 2:00 pm ET. The purpose of this call is to:

- Review and discuss comments received during the post-evaluation public and member comment period;
- Provide input on proposed responses to the post-evaluation comments;
- Review and discuss NQF members' expression of support of the measures under consideration; and
- Determine whether reconsideration of any measures or other courses of action are warranted.

Standing Committee Actions

1. Review this briefing memo and draft report.
2. Review and consider the full text of all comments received and the proposed responses to the post-evaluation comments (see comment table and additional documents included with the call materials).
3. Review the NQF members' expressions of support of the submitted measures.
4. Be prepared to provide feedback and input on proposed post-evaluation comment responses.

Conference Call Information

Please use the following information to access the conference call line and webinar:

Speaker dial-in #: 800-768-2983 (Access Code: 7445915*)*
Web link: https://core.callinfo.com/callme/?ap=8007682983&ac=7445915&role=p&mode=ad

Background

Renal disease is a leading cause of morbidity and mortality in the United States. More than 20 million adults in the United States (10 percent of the population) have chronic kidney disease (CKD), which is associated with premature mortality, decreased quality of life, and increased healthcare costs. Risk factors for CKD include cardiovascular disease, diabetes, hypertension, and obesity. Untreated CKD can

result in end-stage renal disease (ESRD). Currently, over half a million people in the United States have received a diagnosis of ESRD.

This project sought to identify and endorse performance measures for accountability and quality improvement that address conditions, treatments, interventions, or procedures relating to kidney disease.

On June 16, 2020 and June 18, 2020, NQF convened a multistakeholder Standing Committee composed of 24 individuals to evaluate three measures undergoing maintenance review. The Committee recommended two measures for continued endorsement and did not reach consensus on one measure.

Comments Received

NQF solicits comments on measures undergoing review in various ways and at various times throughout the evaluation process. First, NQF solicits comments on endorsed measures on an ongoing basis through the Quality Positioning System (QPS). Second, NQF solicits member and public comments during a 16-week comment period via an online tool on the project webpage.

Pre-evaluation Comments

NQF solicits comments prior to the evaluation of the measures via an online tool on the project webpage. For this evaluation cycle, the pre-evaluation comment period was open from April 28, 2020-June 5, 2020 for the measures under review. No comments were received during this period.

Post-evaluation Comments

The draft report was posted on the project webpage for public and NQF member comment on July 27 for 30 calendar days. During this commenting period, NQF received nine comments from four member organizations:

Member Council	# of Member Organizations Who Commented
Health Professional	2
Public/Community Health Agency	1
QMRI	1

We have included all comments that we received (both pre- and post-evaluation) in the comment table (excel spreadsheet) posted to the Committee SharePoint site. This comment table contains the commenter's name, comment, associated measure, topic (if applicable), and—for the post-evaluation comments—draft responses (including measure steward/developer responses) for the Committee's consideration. Please review this table in advance of the meeting and consider the individual comments received and the proposed responses to each. The Standing Committee's recommendations will be reviewed by the Consensus Standards Approval Committee (CSAC) on November 18. The CSAC will determine whether or not to uphold the Standing Committee's recommendation for each measure submitted for endorsement consideration. All committee members are encouraged to attend the CSAC meeting to listen to the discussion.

Additional Comments not included in the comment table were submitted by the Kidney Care Partners. The comment is included in Appendix B. Developer responses to the comment from Kidney Care Partners is included in Appendix C.

Additionally, please note measure stewards/developers were asked to respond where appropriate. Where possible, NQF staff has proposed draft responses for the Committee to consider.

Comments and Their Disposition

Measure-Specific Comments

2977: Hemodialysis Vascular Access: Standardized Fistula Rate

Commenters suggested excluding from the denominator patients evaluated by vascular surgery but not eligible for arteriovenous fistula (AVF) due to either being a poor surgical candidate (e.g., having a lack of vessels amenable to fistula creation) and patients who refuse AVF creation. Since this measure did not reach consensus for the Evidence criterion, commenters encouraged NQF to consider the "Insufficient Evidence with Exception" pathway towards endorsement.

Measure Steward/Developer Response:

The measure already excludes pediatric patients and patients on hospice.

The TEP that originally developed the two vascular access measures spent a significant amount of time discussing the proposed exclusion for patients who have exhausted anatomic options for permanent access. The TEP agreed that this was an important exclusion, but they also recognized that it would be difficult to implement. A major concern was also that there are not currently data sources or infrastructure in place that would allow identification of patients who have no further surgical options for vascular access. There would also need to be strong consensus on what determines whether patients do not meet criteria for successful fistula placement. We will continue to evaluate this criterion and data availability to determine feasibility of adding this exclusion in a future iteration of this measure.

In response to feedback that the measure should account for patient choice in selecting a vascular access, we acknowledge that it is important that patient preference is taken into account in treatment plans and a patient's "ESKD life plan", as referenced in the updated 2019 KDOQI vascular access guidelines. Patient choice or preference is ostensibly a patient reported outcome however at this time there are no standard criteria for how to validate informed choice, such as a patient's preference to have a catheter (or arteriovenous graft) versus an AV fistula. Check-boxes indicating "received educational materials" or attested to selecting their preferred access type may not be sufficient for determining whether an informed and express choice was made by the patient. Doing so would require ascertaining and measuring 1) what constitutes adequate vascular access education by a nephrologist from the patient's perspective and 2) how that in turn resulted in an informed choice. Accurately capturing informed patient choice may be a particular concern for vulnerable patients.

We agree that non-infectious complications of long-term catheter use for hemodialysis vascular access is an additional supporting argument for catheter avoidance. However, we do not utilize catheter-related complications in the measure calculation. We do agree that specifically measuring non-infectious complications of long-term catheter might lead to another measure of the negative consequences of long-term catheter use. Thank you.

Proposed Committee Response:

Thank you for your comments. The Committee will factor your perspective in to our discussion prior to revote.

Action Item:

Committee will discuss the measure to reach consensus on "Evidence" criteria and finalize the Committee response.

2978: Hemodialysis Vascular Access: Long-term Catheter Rate

Commenters suggested excluding from the denominator patients evaluated by vascular surgery but not eligible for arteriovenous fistula (AVF) due to either being a poor surgical candidate (e.g., having a lack of vessels amenable to fistula creation). They also suggested excluding patients who refuse AVF creation, as well as patients on hospice, with end-state renal disease, and pediatric populations, noting that the measure is not person-focused, as it assumes that AVF is the best option for every individual and their situation, which may not always be the case. Commenters also recommended refining the measure to address non-infectious complications.

Measure Steward/Developer Response:

Developer response is the same as the response for measure 2977.

Proposed Committee Response:

Thank you for your comments. The Committee will review and discuss them during the Post-Comment Meeting.

Action Item:

No action required

0369: Standardized Mortality Ratio for Dialysis Facilities

Commenters requested that pediatric patients and persons on hospice be removed from the measure. Concerns were also raised regarding the standardized mortality ratio's reliability, validity (risk model), specifications, and farmonization issues with CMS's other standardized measures.

Measure Steward/Developer Response:

Hospice exclusion: We have previously considered applying a hospice exclusion to this measure and there are a number of reasons why we have not implemented this exclusion in the SMR.

1. The percentage of patients who have ESRD and die in hospice is very small, and generally do not have a significant effect on measure performance at the facility level.

2. We cannot confirm that patients on hospice are not in that position because of poor dialysis outcomes potentially related to treatment by the dialysis facility.

3. Excluding hospice patients may create inappropriate incentives for "gaming" of the measure, incentivizing facilities to place their poor performing patients in hospice/encouraging withdrawal from dialysis.

End of life exclusions for SMR: We appreciate the comment on exclusions for patients with limited life expectancy (as done for measures 2978 and 2977). Those exclusions were developed specifically for those measures based on empirical analyses and TEP input. We will continue to evaluate the feasibility of such exclusions for the SMR as part of annual maintenance.

Pediatric patients: We appreciate the comments, and agree that there are specific considerations for the pediatric population that will continue to be evaluated in terms of how to strengthen the SMR model. However, we believe that the SMR is an important measure to

report for the pediatric population in order for patients and caregivers to be able to compare facilities from which they may seek treatment.

Proposed Committee Response:
Thank you for your comments. The Committee will review and discuss them during the Post-Comment Meeting.

Action Item:
No action required

NQF Member Expression of Support

Throughout the 16-week continuous public commenting period, NQF members had the opportunity to express their support ("support" or "do not support") for each measure submitted for endorsement consideration to inform the Committee's recommendations. Three NQF members provided their expressions of support: See Appendix A.

Appendix A: NQF Member Expression of Support Results

Two NQF members provided their expressions of support/nonsupport. 2 of 3 measures under consideration received support from NQF members. Results for each measure are provided below.

Measure Number: 0369 Standardized Mortality Ratio for Dialysis Facilities (University of Michigan Kidney and Epidemiology and Cost Center (UM-KECC))

Member Council	Support	Do Not Support	Total
Health Professional		1	1

Measure Number: 2977 Hemodialysis Vascular Access, Standardized Fistula Rate (University of Michigan Kidney and Epidemiology and Cost Center (UM-KECC))

Member Council	Support	Do Not Support	Total
QMRI	1		1

Measure Number: 2978 Hemodialysis Vascular Access, Long-Term Catheter Rate (University of Michigan Kidney and Epidemiology and Cost Center (UM-KECC))

Member Council	Support	Do Not Support	Total
QMRI	1		1

Appendix B: Additional Comment

August 21, 2020

National Quality Forum

1099 14th Street NW
Suite 500
Washington DC 20005

RE: NQF Renal Project Spring 2020 Cycle

Kidney Care Partners (KCP) appreciates the opportunity to comment on the measures under consideration for endorsement in the National Quality Forum's (NQF) Spring 2020 Renal 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 three measures under review.

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 SMR's reliability, validity (risk model), specifications, and has identified several harmonization issues with CMS's other standardized measures. In addition, the increasing numbers of Medicare Advantage (MA) patients in the ESRD quality programs—and the unavailability of outpatient claims data for these patients—threaten the validity of several CMS measures, including the SMR.

- **Reliability.** Based on the testing results, KCP has serious concerns about the SMR's reliability. In the most recent iteration of the measure, the overall IUR for the 4-year SMR was 0.5—a sizeable decline from the 2016 value of 0.59. That is, even with the 4-year SMR, one-half of a facility's score could be attributable to random noise and not signal. Though we recognize the characterization also depends on the analytic method, we note that statistical literature traditionally interprets a reliability statistic of 0.5-0.6 as "unacceptable."[1] Thus KCP thus believes that CMS should implement the measure adjusted to yield a reliable result (reliability statistic of 0.70 or greater), consistent with how the NQF bases its evaluation of measures and more generous than the literature.[2]

[1] Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California:RAND Corporation. TR-653-NCQA, 2009.

[2] Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge, p. 13; DeVellis, RF. (2012). *Scale development: Theory and applications*. Los Angeles: Sage. pp. 109–110; Adams, JL. (2009). The reliability of provider profiling. RAND Health.

Moreover, CMS did not provide testing data stratified by facility size for the measure iteration currently under review by NQF because it "is not required." Yet we note that even when using the 4-year period, prior SMR testing results indicated very poor reliability for small- and medium-sized facilities, with IURs of 0.30 and 0.45, respectively. Only large facilities had a reasonable IUR of 0.73 for 2010-2013 data. Given this history, we believe it's disingenuous, at best, not to provide reliability based on facility size merely because NQF "does not require" it. KCP believes penalizing facilities for performance due to random chance is not appropriate and that it is imperative that CMS provide the most recent reliability results stratified by facility size. Absent that information, we submit that the demonstrably unreliable SMR, as currently specified, is particularly unreliable and unsuitable for use in small facilities. KCP thus believes the measure must specifically require a minimum sample as identified through the developer's empirical testing to prevent small facilities from having scores that are highly subject to random variability. To further support this recommendation, Discern Health identified an approach using publicly available data to estimate the reliability based on facility size (Attachment A). Based on this analysis, the IUR for the smallest 20 percent of facilities is 0.449—meaning the SMR is unreliable for this population, wherein only 44.9% of their score is attributable to quality signal and more than half to random noise.

Regarding the performance period timeframe, we note that prior testing results for the 1-year SMR yielded IURs of 0.26-0.32 for each of 2010, 2011, 2012, and 2013—an "unacceptable" degree of reliability[3] where only about 30% of the variation in a score can be attributed to between-facility differences—yet the specifications at the time permitted the 1-year measure. The 4-year SMR yielded an IUR of 0.59 over the same time period, illustrating why it is critical that the performance period be precisely defined—and why that timeframe should at a minimum be a 4-year period.

Finally, to assess more directly the value of SMR in identifying facilities with extreme outcomes, CMS and UM-KECC crafted an additional metric of reliability termed the Profile-IUR (PIUR).[4] Per CMS, "The PIUR indicates the presence of outliers or heavier tails among the providers, which is not captured in the IUR itself. . . . [When] there are outlier providers, even measures with a low IUR can have a relatively high PIUR and can be very useful for identifying extreme providers."[5] The PIUR for the SMR was PIUR is 0.77, which CMS interprets as demonstrating that "the SMR is effective at detecting outlier facilities and statistically meaningful differences in performance scores across dialysis facilities."[6] We note that NQF's Scientific Methods Panel (SMP), none of whom were familiar with the PIUR, disagree that it is an appropriate measure of reliability for *any* measure used in the ESRD Quality Incentive Program (QIP), which are used to distinguish performance between providers falling in the *middle* of the curve to determine penalties. The SMP concluded that the IUR is and remains the appropriate measure of reliability for this purpose.

---

[3] Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California:RAND Corporation. TR-653-NCQA, 2009.

[4] He K, Dahlerus C, Xia L, Li Y, Kalbfleisch JD. The profile inter-unit reliability. *Biometrics.* 2019 Oct 23. doi: 10.1111/biom.13167. [Epub ahead of print.]

[5] Kalbfleisch JD, He K, Xia L, Li Y. Does the inter-unit reliability (IUR) measure reliability? *Health Services and Outcomes Research Methodology.* 2018;18(3):215-225. Doi: 10.1007/s10742-018-0185-4.

[6] CMS. SMR measures submission materials to NQF.

- **Validity and Handling of Medicare Advantage Patients.** In previous comments to CMS, KCP noted that many of the prevalent comorbidities in the final model had p-values significantly greater than 0.05. CMS responded that the large number of clinical factors in the model generates multicollinearity among covariates, likely resulting in some unexpected results. However, KCP remains concerned that this strategy results in a model that will not be generalizable. In the current model, for example, asthma is associated with a higher risk of death than critical illness myopathy, and 'obesity' is protective while 'mood disorders' are harmful. We posit these nonsensical findings are a function of collinearity and coding idiosyncrasy. Again, KCP supports prevalent comorbidity 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 validity testing yielded a c-statistic for the SMR of 0.72. 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.

  Finally, data provided by CMS indicate that at the end of 2017, 27 percent of dialysis patients had MA coverage (presumably higher now), and this varied widely across states—from about 2 percent in Wyoming to 34 percent in Rhode Island, and more than 44 percent in Puerto Rico (Table 2, Attachment B). Such geographic variation compromises the validity of the measures if MA patients are not accurately accounted for in the QIP and DFC metrics. Specifically, without changes to the current specifications, the evolving patient mix will introduce significant bias into measure calculations that could affect results for facilities with either very low or high MA patient populations. Recognizing this, KCP concurs with the need to change specifications for several CMS measures to accommodate the increase in MA patients and to avoid disparities in performance due to geography. KCP strongly believes, however, that greater transparency is required by CMS as it updates the relevant measures.

  While the approach to handling MA patients varies considerably across CMS's metrics (Table 1, Attachment B), KCP recognizes the difficulty in construction and notes there appears to be a logical rationale for most of the decisions made because of the properties and intended purpose of each measure. Nevertheless, KCP strongly recommends that CMS perform a sensitivity analysis of performance with and without MA patients for each of the applicable QIP/DFC measures and ***make the results publicly available***. Such data will provide an opportunity for KCP and others to offer potential, evidence-based mitigation strategies (e.g., a model that accounts for both populations, use of risk coefficients, as necessary).

  CMS also should perform an analysis of risk model fit under the previous approach and the new in-patient-claims-only approach; currently we are unable to assess whether model fit improved or worsened with this approach. KCP is particularly concerned that limiting comorbidity data to inpatient claims might skew the models towards a sicker population, and that such a skew might reflect unfavorably on facilities that successfully keep hospitalization rates low. That is, because comorbidity adjustors developed exclusively from hospitalization data will necessarily underestimate the comorbidity profile of patients in facilities with low hospitalization rates, the "expected" hospitalization or mortality rates calculated for such facilities will be erroneously low, and the facilities' scores will be

erroneously high. Only with transparency in these matters can the community assess the impact MA patient mix has on the QIP measures.

- **Specifications and Harmonization Issues.** Prior SMR specifications indicated that "The time window can be specified from one to four years. Currently, the measure is publicly reported using four years of data." This language has now been removed from the specifications. While we have previously noted that this prior construction was imprecise and KCP believes specifications should be unambiguous, we are concerned that there is no longer any mention at all of the performance period timeframe. We believe the time period should be exact; we further believe the 1-year period is inappropriate based on the reliability testing data (see above) and, at minimum, should be a 4-year period.

  Additionally, the risk model groupings and parameters used for patient age and duration of ESRD differ among CMS's standardized measures falling under the NQF Renal Project domain. For example, the age groups for the SMR is n=3, but for Standardized Transfusion Ratio (STrR, NQF 2979) and Standardized Fistula Rate (SFR, NQF 2977) there are 5 and 4 age groupings, respectively. Similarly, the number of groups for ESRD duration for the SMR and SFR are both n=4 but the groupings differ dramatically, and both differ from the STrR (n=6). No justification or empirical analyses are offered to justify these differences.

NQF 2977: Hemodialysis Vascular Access—Standardized Fistula Rate (SFR; CMS)

The NQF Renal Standing Committee did not reach consensus for the *Evidence* criterion for the SFR secondary to concerns from some members that the measure may effectively be "topped out" at 64% for all patients for whom an arteriovenous fistula (AVF) is clinically appropriate, and because NKF's Kidney Disease Outcomes Quality Initiative downgraded the evidence supporting the relevant guideline to "expert opinion." Nevertheless, because vascular access may be the most important measure for patients making decisions about dialysis facilities, KCP believes it remains important to include a fistula measure in the ESRD QIP and we continue our general support of this measure. If the Standing Committee does not feel the empiric evidence is sufficient for this measure, we encourage them to consider NQF's "Insufficient Evidence with Exception" pathway.

We also offer the following technical considerations for the developer:

- KCP notes that the validity testing for the SFR yielded an overall c-statistic of 0.705; as previously noted, 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.
- As with the SMR, KCP strongly recommends that CMS perform a sensitivity analysis of performance with and without Medicare Advantage patients for the SFR and ***make the results publicly available*** to provide stakeholders the opportunity to offer potential mitigation strategies (e.g., a model that accounts for both populations, use of risk coefficients, as necessary).
- As described above for the SMR, the numbers and parameters for the risk model "patient age" and "duration of ESRD" groupings differ among CMS's standardized measures in the Renal Project; no justification or empirical analyses are offered to justify these differences.

NQF 2978: Hemodialysis Vascular Access—Long-Term Catheter Rate (LTCR; CMS)

KCP generally supports this measure.

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 203.530.9524).

Sincerely,

Kidney Care Partners:

Akebia Therapeutics, Inc.

American Kidney Fund, Inc.

American Nephrology Nurses Association

American Renal Associates

American Society of Nephrology

American Society of Pediatric Nephrology

Amgen, Inc.

Ardelyx

AstraZeneca

Atlantic Dialysis Management Services, LLC

Baxter International, Inc.

B. Braun Medical, Inc.

Cara Therapeutics, Inc.

Centers for Dialysis Care

DaVita, Inc.

Dialysis Patient Citizens, Inc.

DialyzeDirect

Fresenius Medical Care North America

Fresenius Medical Care Renal Therapies Group

Greenfield Health Systems

Kidney Care Council

National Kidney Foundation, Inc.

National Renal Administrators Association

Nephrology Nursing Certification Commission

Renal Physicians Association

Renal Support Network

Rockwell Medical

Rogosin Institute

Satellite Healthcare, Inc.

U.S. Renal Care, Inc.

Vertex Pharmaceuticals

Vifor Pharma Ltd.

ATTACHMENT A

Standardized Mortality Ratio Measure Inter-Unit Reliability Measure Analysis

KCP raised a concern that some quality measures that assess the care delivered by dialysis facilities may not be reliable for small facilities. Discern Health sought to test this hypothesis by performing an Inter-Unit Reliability (IUR) analysis on the quality measures of interest. This memorandum details the results of the IUR analysis on the Standardized Mortality Ratio (SMR) measure.

Discern chose the IUR metric as it is commonly used to assess the reliability of quality measures.[7] IUR evaluates the proportion of measure variability that can be attributed to between-facility variance:

$$\text{IUR} = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

A small IUR (close to 0) shows that most of the measure variation between facilities is due to random noise, whereas a large IUR (close to 1) shows that most of the variation between facilities is due to real differences.[8]

Other techniques exist to assess the reliability of quality measures.[1,9] Some of the methods provide "risk adjustment to account for differences between patients treated by different providers" and may better able to profile facilities with 'extreme outcomes.'[10] However, the alternate treatments require more data than is publicly available, and therefore these analyses were not performed.

To begin, Discern grouped dialysis facilities by size. This was done using the most recent Standardized Hospitalization Ratio (SHR) denominator, which is the "number of hospital admissions that would be expected among eligible patients at the facility during the reporting period, given the patient mix at the facility."[11] This indicator was chosen as a *proxy* measure for population size at each facility, which replicates established IUR calculation methods for SMR. Specifically, in the context of IUR for SMR, size is defined by patient time at risk for mortality, and we chose expected hospitalizations because that considers patient-months; we recognize it also incorporates risk of hospitalizations, which is not preferred, but the alternative approach considered was dialysis stations, which raises more significant issues. This variable doesn't explicitly consider patient time and could lead to facility misclassifications (consider a scenario where a facility has a high number of dialysis stations, but is only open 4 days a week). Therefore, for the IUR analyses, facilities were placed in groups of 50 expected hospital admissions (e.g. 0-50 expected admissions).

One-year IUR Calculation

---

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6455941/
[8] https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/ESRDQIP/Downloads/AnalysisoftheReliabilityoftheProposedSRRandSTrRMeasures.pdf
[9] https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR653.pdf
[10] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7318309/
[11] https://www.cms.gov/files/document/py-2023-final-technical-specifications-20191101.pdf

Then, Discern performed a one-way analysis of variance (ANOVA) on the 2018 SMR scores for the stratified groups of facilities. The ANOVA results were then used to calculate IUR. Lastly, Discern computed IUR for all facilities combined. **Table 1** displays the results of the analysis.

Table 1. 2018 SMR IUR by Facility Size

Facility Size (based on expected admissions)	Number of Facilities (%)*	IUR
0-50	1681 (25%)	0.2440
50-100	2717 (41%)	0.8544
100-150	1442 (22%)	0.9575
150+	767 (12%)	0.8623
TOTAL	6607	0.4319

*Note: Facilities were excluded from the analysis if they did not have an SMR score in the most recent Dialysis Facility Compare data.

The overall IUR for all facilities reporting SMR data in 2018 was 0.4319. IURs of at least 0.7 are required for reliable group comparisons.[12] This indicates that under half of the variation in measure performance can be attributed to differences between facilities. The IUR for facilities with fewer than 50 expected hospital admissions was 0.2440. This is significantly lower than the IUR of the other stratified groups, which all reported IURs above 0.85, indicating acceptable reliability. This analysis shows that the SMR measure does not reliably differentiate between true performance and random chance for 25 percent of all facilities.

Four-year IUR Calculation

Following the methods outlined above, Discern also calculated the 4-year IUR by averaging facility SMR scores from 2015-2018. To be conservative, facilities not reporting a SMR for any of those years were excluded from analysis. A total of 1,301 facilities, representing 20% of the population, were excluded. These excluded facilities were more likely to be newer facilities and report more volatile score. Removing them from the analysis would lead us to overestimate IUR, though our analysis confirmed that this did not introduce meaningful bias into the analysis. **Table 2** displays the results of the analysis.

[12] https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR653.pdf

Table 2. 4-Year (2018-2015) SMR IUR by Facility Size

Facility Size (based on expected admissions)	Number of Facilities (%)*	IUR
0-50	1051 (20%)	0.4490
50-100	2241 (42%)	0.8739
100-150	1304 (25%)	0.9601
150+	710 (13%)	0.8615
TOTAL	5306	0.4908

*Note: Facilities missing SMR scores in any of the Dialysis Facility Compare data years were excluded from the analysis.

The total 4-year IUR for all facilities reporting SMR data between 2018-2015 was 0.4908, with poor reliability concentrated among facilities with fewer than 50 expected admissions. Improvement in the 4-year IUR relative to the 1-year IUR is likely due to the 4-year IUR analysis requiring facilities to report SMR score for each year of data. Additionally, a longer time span of measurement will generally lead to more signal and less noise in a measure, particularly for relatively rare outcomes like mortality. However, the total IUR fails to reach the threshold of 0.7 required for reliable group comparisons. This indicates that nearly half of the variation in measure performance can be attributed to differences between facilities.

The 4-year IUR for facilities with fewer than 50 expected hospital admissions was 0.4490. This is significantly lower than the 4-year IUR of the other stratified groups, which all reported 4-year IURs above 0.85, indicating acceptable reliability. This analysis shows that the SMR measure does not reliably differentiate between true performance and random chance for 20 percent of all facilities.

ATTACHMENT B: MEDICARE ADVANTAGE ISSUES

Table 1: CMS Approach to MA Patients[13]

	NQF 0369: Standardized Mortality Ratio	NQF 1463: Standardized Hospitalization Ratio	NQF 2496: Standardized Readmission Ratio	NQF 2977: Standardized Fistula Rate	NQF 2978: Long-term Catheter Rate	NQF 2979: Standardized Transfusion Ratio	NQF 3565: Standardized ED Ratio (SEDR)	NQF 3566: ED within 30 Days (ED30)
Denominator Data Source(s)	Treatment History Files	Treatment History Files	Inpatient claims	CROWNWeb	CROWNWeb	Treatment History Files	Treatment History Files	Inpatient claims
Handling of MA Patients								
MA patients are included in all data sources, but their payment records are limited to inpatient claims; tracking by dialysis provider and treatment modality is available for all patients, including those with only partial or no Medicare coverage.	✓	✓	✓	✓	✓			
All time-at-risk for MA patients is included.	✓	✓						
Risk variable for proportion of MA months incorporated into risk model.		✓						
Risk variable to identify patients with MA coverage at time of index discharge.			✓					
Limited to Medicare patients enrolled in MA or who meet the criterion of being within 2 months after a month with either (a) $1200+ of	✓	✓						

[13] Blue cells are "not applicable" for that measure.

Medicare-paid dialysis claims or (b) at least one Medicare inpatient claim.								
Binary Medicare coverage indicator: Patient-months where patient had at least 6 Medicare-covered months or 1 or more MA-covered month(s) in past 12 months. This indicator is used to determine the presence of prevalent comorbidities from Medicare claims in prior year.	✓	✓		✓				
Excludes patients with MA coverage						✓	✓	✓
Revision(s) to Identification of Prevalent Comorbidities Process to Account for MA								
Prevalent comorbidities data limited to inpatient claims.	✓	✓						
Past-year comorbidities limited to inpatient claims.			✓					
Past-year comorbidity data obtained from multiple Part A types (inpatient, home health, hospice, SNF claims) only.	✓	✓		✓				
Uses AHRQ CCS diagnosis categories to identify patient prevalent comorbidities.	✓	✓	✓					

Table 2: Average of Dialysis Facilities' Percent MA Patients by State, 2018

State	N	Mean % (SD)
PR	44	44.2 (14.5)
RI	16	33.6 (18.5)
HI	31	27.8 (11.2)
OH	323	26.8 (11.4)
PA	307	25 (14.5)
AZ	121	24.6 (12.5)
CA	658	23.9 (16.6)
MN	119	23.5 (10.6)
OR	71	22.9 (15.3)
MI	211	22.4 (10.1)
TN	185	21 (8.9)
AL	176	19.8 (10.5)
FL	456	19.6 (10.3)
CO	125	18.7 (8.9)
WI	80	18.7 (11)
TX	675	18.6 (10.9)
NY	353	17.2 (7.6)
GA	296	17.2 (8.8)
NV	49	16.9 (9.7)
WV	45	16.6 (8.2)
KY	120	16.2 (6.7)
MO	165	15.2 (9.1)
NC	220	14.9 (8.6)
SC	150	14.4 (6.6)
IN	166	14.2 (8.1)
LA	175	14 (10)
NM	54	13.9 (12.2)
IL	317	13.2 (9.5)
MA	84	13.1 (11.8)
NJ	48	12.7 (4.9)
CT	179	12.7 (6.3)
VI	4	12.5 (25)
ID	43	12.1 (8.5)
UT	28	12.1 (8.9)
ME	17	11.6 (5.3)
WA	93	11 (8.5)
VA	189	10.9 (6.3)

State	N	Mean % (SD)
AR	70	10.8 (6.4)
KS	57	9.3 (7.5)
IA	67	8.2 (6.6)
DC	86	7.8 (6.6)
MS	90	7.8 (5.1)
OK	21	7.7 (10.1)
NE	166	7.4 (9.7)
MD	38	7.2 (7)
ND	16	6.7 (4.9)
DE	28	6.2 (4.6)
VT	8	5.5 (2.8)
SD	27	5.3 (6)
NH	19	4.8 (3.3)
MT	15	3.6 (3.7)
AK	9	2.3 (3.2)
WY	10	2.2 (3.2)

Appendix C: Developer Response to Additional Comment

SMR

Reliability. Based on the testing results, KCP has serious concerns about the SMR's reliability. In the most recent iteration of the measure, the overall IUR for the 4-year SMR was 0.5—a sizeable decline from the 2016 value of 0.59. That is, even with the 4-year SMR, one-half of a facility's score could be attributable to random noise and not signal. Though we recognize the characterization also depends on the analytic method, we note that statistical literature traditionally interprets a reliability statistic of 0.5-0.6 as "unacceptable."1 Thus KCP thus believes that CMS should implement the measure adjusted to yield a reliable result (reliability statistic of 0.70 or greater), consistent with how the NQF bases its evaluation of measures and more generous than the literature.2

Moreover, CMS did not provide testing data stratified by facility size for the measure iteration currently under review by NQF because it "is not required." Yet we note that even when using the 4-year period, prior SMR testing results indicated very poor reliability for small- and medium-sized facilities, with IURs of 0.30 and 0.45, respectively. Only large facilities had a reasonable IUR of 0.73 for 2010-2013 data. Given this history, we believe it's disingenuous, at best, not to provide reliability based on facility size merely because NQF "does not require" it. KCP believes penalizing facilities for performance due to random chance is not appropriate and that it is imperative that CMS provide the most recent reliability results stratified by facility size. Absent that information, we submit that the demonstrably unreliable SMR, as currently specified, is particularly unreliable and unsuitable for use in small facilities. KCP thus believes the measure must specifically require a minimum sample as identified through the developer's empirical testing to prevent small facilities from having scores that are highly subject to random variability. To further support this recommendation, Discern Health identified an approach using publicly available data to estimate the reliability based on facility size (Attachment A). Based on this analysis, the IUR for the smallest 20 percent of facilities is 0.449—meaning the SMR is unreliable for this population, wherein only 44.9% of their score is attributable to quality signal and more than half to random noise.

Regarding the performance period timeframe, we note that prior testing results for the 1- year SMR yielded IURs of 0.26-0.32 for each of 2010, 2011, 2012, and 2013—an "unacceptable" degree of reliability3 where only about 30% of the variation in a score can be attributed to between-facility differences—yet the specifications at the time permitted the 1- year measure. The 4-year SMR yielded an IUR of 0.59 over the same time period, illustrating why it is critical that the performance period be precisely defined—and why that timeframe should at a minimum be a 4-year period.

Finally, to assess more directly the value of SMR in identifying facilities with extreme outcomes, CMS and UM-KECC crafted an additional metric of reliability termed the Profile- IUR (PIUR).4 Per CMS, "The PIUR indicates the presence of outliers or heavier tails among the providers, which is not captured in the IUR itself. . . . [When] there are outlier providers, even measures with a low IUR can have a relatively high PIUR and can be very useful for identifying extreme providers."5 The PIUR for the SMR was PIUR is 0.77, which CMS interprets as demonstrating that "the SMR is effective at detecting outlier facilities and statistically meaningful differences in performance scores across dialysis facilities."6 We note that NQF's Scientific Methods Panel (SMP), none of whom were familiar with the PIUR, disagree that it is an appropriate measure of reliability for *any* measure used in the ESRD Quality Incentive Program (QIP), which are used to distinguish performance between providers falling in the *middle* of the curve to

determine penalties. The SMP concluded that the IUR is and remains the appropriate measure of reliability for this purpose.

Response: Regarding the time period of the SMR, our intention is that this is a four year measure, as was approved during the last maintenance cycle in 2016. We will correct the language used in the submission at the next opportunity.

Kalbfleisch, et al. (2018) explains that the interpretation of the IUR as reliability depends on the differences between providers being entirely (or mostly) due to the quality of care. In many if not most instances, however, this is not the case. There are differences in the patients treated by providers that are not accounted for in the adjustments that we are able to make. In effect, there will almost always be unmeasured confounders that are related to the outcome and also differ between facilities. For example, these include genetic differences among the patients treated that vary across facilities, or dietary differences, or differences in the level of family support, etc. We do not measure these variables, but they are undoubtedly important and they contribute to the between facility variation. Thus, one can have a high value of IUR due simply to incomplete risk adjustment and in general, adjusting for confounders can reduce the IUR. Similarly, an IUR near 0 does not mean that the measure is not useful for profiling. In fact, if most of the providers have outcomes centered very near a national average while a relatively smaller but appreciable number have outcomes that are out of line, the IUR would be near 0, and yet the measure may be very useful for identifying those extreme facilities. For these reasons, the IUR should be interpreted with care as it may not reflect the true reliability of the measure.

These considerations motivated the definition of the PIUR, which concentrates on the ability of the measure to consistently flag the same facilities. The PIUR is introduced in He et al. (2019) , where a number of examples can be found. Briefly, the PIUR is based on the estimated probability that a facility flagged as extreme on one occasion would be flagged again on a second occasion. Since we do not have two occasions to compare, the estimated reflagging probability is based on sample splitting. In many instances, one is particularly interested in identifying providers whose outcomes are extreme and the PIUR concentrates on this aspect. Note that the PIUR is very close in spirit to the definition of reliability in the testing form: ***"2a2. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise."***

The overall IUR for the four-year SMR is 0.5. The PIUR is 0.77. The value of IUR indicates a moderate degree of reliability. The higher PIUR compared to the IUR indicates the presence of outlier providers, which is not captured in the IUR itself. For instance, if there are no outliers, one should expect the PIUR to be the same as the IUR. Thus, the difference between the PIUR and the IUR demonstrates that the SMR is effective at detecting outlier facilities and statistically meaningful differences in performance scores across measured entities.

Finally, we note that the SMR is not reported in the QIP program.

Validity and Handling of Medicare Advantage Patients. In previous comments to CMS, KCP noted that many of the prevalent comorbidities in the final model had p-values significantly greater than 0.05. CMS responded that the large number of clinical factors in the model generates multicollinearity among covariates, likely resulting in some unexpected results. However, KCP remains concerned that this strategy results in a model that will not be generalizable. In the current model, for example, asthma is associated with a higher risk of death than critical illness myopathy, and 'obesity' is protective while 'mood disorders' are harmful. We posit these nonsensical findings are a function of collinearity and coding idiosyncrasy. Again, KCP supports prevalent comorbidity 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 validity testing yielded a c-statistic for the SMR of 0.72. 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.

Response: Over the history of the development and refinement of the SMR, we evaluated multiple iterations to obtain the most robust model possible. We believe that the C-statistic of 0.72 is considered to be a good fit based on recent literature and note that it is similar in magnitude to other current NQF endorsed quality measures that have been implemented by CMS. Several references are listed below from peer-reviewed literature that report C-statistics of similar magnitude to the one achieved in our model. 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

Finally, data provided by CMS indicate that at the end of 2017, 27 percent of dialysis patients had MA coverage (presumably higher now), and this varied widely across states— from about 2 percent in Wyoming to 34 percent in Rhode Island, and more than 44 percent in Puerto Rico (Table 2, Attachment B). Such geographic variation compromises the validity of the measures if MA patients are not accurately

accounted for in the QIP and DFC metrics. Specifically, without changes to the current specifications, the evolving patient mix will introduce significant bias into measure calculations that could affect results for facilities with either very low or high MA patient populations. Recognizing this, KCP concurs with the need to change specifications for several CMS measures to accommodate the increase in MA patients and to avoid disparities in performance due to geography. KCP strongly believes, however, that greater transparency is required by CMS as it updates the relevant measures.

While the approach to handling MA patients varies considerably across CMS's metrics (Table 1, Attachment B), KCP recognizes the difficulty in construction and notes there appears to be a logical rationale for most of the decisions made because of the properties and intended purpose of each measure. Nevertheless, KCP strongly recommends that CMS perform a sensitivity analysis of performance with and without MA patients for each of the applicable QIP/DFC measures and *make the results publicly available*. Such data will provide an opportunity for KCP and others to offer potential, evidence-based mitigation strategies (e.g., a model that accounts for both populations, use of risk coefficients, as necessary).

CMS also should perform an analysis of risk model fit under the previous approach and the new in-patient-claims-only approach; currently we are unable to assess whether model fit improved or worsened with this approach. KCP is particularly concerned that limiting comorbidity data to inpatient claims might skew the models towards a sicker population, and that such a skew might reflect unfavorably on facilities that successfully keep hospitalization rates low. That is, because comorbidity adjustors developed exclusively from hospitalization data will necessarily underestimate the comorbidity profile of patients in facilities with low hospitalization rates, the "expected" hospitalization or mortality rates calculated for such facilities will be erroneously low, and the facilities' scores will be erroneously high. Only with transparency in these matters can the community assess the impact MA patient mix has on the QIP measures.

> Response: We appreciate the comments, as the Medicare Advantage issue will present challenges to measure development and reporting in the years to come. However, we feel the requests from the commenter fall outside the scope of NQF measure review. The information included in our submission met the requirements laid out by NQF for maintenance review.

Specifications and Harmonization Issues. Prior SMR specifications indicated that "The time window can be specified from one to four years. Currently, the measure is publicly reported using four years of data." This language has now been removed from the specifications. While we have previously noted that this prior construction was imprecise and KCP believes specifications should be unambiguous, we are concerned that there is no longer any mention at all of the performance period timeframe. We believe the time period should be exact; we further believe the 1-year period is inappropriate based on the reliability testing data (see above) and, at minimum, should be a 4-year period.

> Response: Our intention is that this is a four year measure, as was approved during the last maintenance cycle in 2016. We will correct the language used in the submission at the next opportunity.

Additionally, the risk model groupings and parameters used for patient age and duration of ESRD differ among CMS's standardized measures falling under the NQF Renal Project domain. For example, the age

groups for the SMR is n=3, but for Standardized Transfusion Ratio (STrR, NQF 2979) and Standardized Fistula Rate (SFR, NQF 2977) there are 5 and 4 age groupings, respectively. Similarly, the number of groups for ESRD duration for the SMR and SFR are both n=4 but the groupings differ dramatically, and both differ from the STrR (n=6). No justification or empirical analyses are offered to justify these differences.

Response: The categories for the Age and Duration of ESRD covariates in the risk adjustment models were empirically derived when each model was first developed, and are based on model fit specific to each outcome. This accounts for the use of different groupings for each model.

SFR

KCP notes that the validity testing for the SFR yielded an overall c-statistic of 0.705; as previously noted, 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.

Response: We evaluated multiple iterations of our standardized fistula rate to obtain the most robust model possible. We believe that the C-statistic of 0.74 is considered to be a good fit based on recent literature and note that it is similar in magnitude to other current NQF endorsed quality measures that have been implemented by CMS. Several references are listed below from peer-reviewed literature that report C-statistics of similar magnitude to the one achieved in our model. As we refine the risk model in the future, we will work to improve the model's ability to discriminate performance between facilities. In addition, the standardized fistula model was reviewed and endorsed by the TEP, providing both face validity and an element of peer review for the measure.

7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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

As with the SMR, KCP strongly recommends that CMS perform a sensitivity analysis of performance with and without Medicare Advantage patients for the SFR and *make the results publicly available* to provide stakeholders the opportunity to offer potential mitigation strategies (e.g., a model that accounts for both populations, use of risk coefficients, as necessary).

Response: We appreciate the comments, as the Medicare Advantage issue will present challenges to measure development and reporting in the years to come. However, we feel the requests from the commenter fall outside the scope of NQF measure review. The information included in our submission met the requirements laid out by NQF for maintenance review.

As described above for the SMR, the numbers and parameters for the risk model "patient age" and "duration of ESRD" groupings differ among CMS's standardized measures in the Renal Project; no justification or empirical analyses are offered to justify these differences.

Response: The categories for the Age and Duration of ESRD covariates in the risk adjustment models were empirically derived when each model was first developed, and are based on model fit specific to each outcome. This accounts for the use of different groupings for each model.

Finally, to assess more directly the value of SMR in identifying facilities with extreme outcomes, CMS and UM-KECC crafted an additional metric of reliability termed the Profile- IUR (PIUR).4 Per CMS, "The PIUR indicates the presence of outliers or heavier tails among the providers, which is not captured in the IUR itself. . . . [When] there are outlier providers, even measures with a low IUR can have a relatively high PIUR and can be very useful for identifying extreme providers."5 The PIUR for the SMR was PIUR is 0.77, which CMS interprets as demonstrating that "the SMR is effective at detecting outlier facilities and statistically meaningful differences in performance scores across dialysis facilities."6 We note that NQF's Scientific Methods Panel (SMP), none of whom were familiar with the PIUR, disagree that it is an appropriate measure of reliability for *any* measure used in the ESRD Quality Incentive Program (QIP), which are used to distinguish performance between providers falling in the *middle* of the curve to determine penalties. The SMP concluded that the IUR is and remains the appropriate measure of reliability for this purpose.

The overall IUR for the four-year SMR is 0.5. The PIUR is 0.77. The value of IUR indicates a moderate degree of reliability. The higher PIUR compared to the IUR indicates the presence of outlier providers, which is not captured in the IUR itself. For instance, if there are no outliers, one should expect the PIUR to be the same as the IUR. Thus, the difference between the PIUR and the IUR demonstrates that the SMR is effective at detecting outlier facilities and statistically meaningful differences in performance scores across measured entities.