NATIONAL QUALITY FORUM
Driving measurable health improvements together

Memo

September 24, 2020

To: All-Cause Admissions & Readmissions Standing Committee

From: NQF staff

Re: Post-Comment Call to Discuss Public and Member Comments

Introduction

NQF closed the public commenting period on the measures submitted for endorsement consideration to the Spring 2020 measure review cycle on September 3, 2020.

Purpose of the Call

The All-Cause Admissions & Readmissions Standing Committee will meet via web meeting on September 24, 2020 from 1:00 pm to 3:00 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.
- Determine whether reconsideration of any measures or other courses of action is 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. 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 #: 1-800-768-2983; Access Code: 4364232
Web link: https://core.callinfo.com/callme/?ap=8007682983&ac=4364232%20&role=p&mode=ad

Background

Unplanned and potentially avoidable all-cause and condition-specific returns to the hospital, including emergency department encounters, continue to pose considerable strain on healthcare expenditure and quality of care for patients. These avoidable admissions and readmissions often represent an opportunity to improve care transitions and prevent the unnecessary exposure of patients to adverse events in an acute care setting. To drive improvement in admissions and readmissions, performance

measures have continued to be a key element of value-based purchasing programs to incentivize collaboration in the healthcare delivery system.

The twenty-two active members on the All-Cause Admissions and Readmissions Standing Committee have been charged with overseeing the NQF All-Cause Admissions and Readmission portfolio, evaluating both newly submitted and previously endorsed measures against NQF's measure evaluation criteria, identifying gaps in the measurement portfolio, providing feedback on how the portfolio should evolve, and serving on any ad hoc or expedited projects in its designated topic areas. The All-Cause Admissions and Readmissions portfolio includes measures for various care settings or points of care.

On June 22, 2020, NQF convened the All-Cause Admissions and Readmissions Standing Committee to evaluate two measures undergoing maintenance review and three new measure.

The Committee recommended four measures for endorsement:

- **NQF 1463:** Standardized Hospitalization Ratio for Dialysis Facilities (SHR) (UM Kidney Epidemiology and Cost Center/CMS)
- **NQF 3565:** Standardized Emergency Department Encounter Ratio (SEDR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)
- **NQF 3566:** Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge (ED30) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)
- **NQF 2539:** Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy (Yale CORE/CMS)

The Committee did not recommend one measure for continued endorsement:

- **NQF 2496:** Standardized Readmission Ratio (SRR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

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 May 1, 2020 to June 12, 2020 for the measures under review. A total of two pre-evaluation comments were received, the majority of which pertained to, and were not in support of, the two newly submitted measures (NQF 3566 and NQF 3565). All pre-evaluation comments were provided to the Committee prior to the June 22, 2020 web meeting.

Post-evaluation Comments

The Spring 2020 draft report went out for public and member comment August 5, 2020 to September 3, 2020. During this commenting period, NQF received eight comments from five member organizations:

Member Council	# of Member Organizations Who Commented
Health Professional	3
QMRI	1
Supplier/Industry	1

Where possible, NQF staff has proposed draft responses for the Committee to consider. Although all comments are subject to discussion, the intent is not to discuss each individual comment on the September 24 post-comment call. Instead, we will spend the majority of the time considering the themes discussed below, and the set of comments as a whole. Please note that the organization of the comments into major topic areas is not an attempt to limit Committee discussion. Additionally, please note measure developers were asked to respond where appropriate.

We have included all comments (both pre- and post-evaluation) that we received in the comment table in excel spreadsheet posted to the Committee SharePoint site. This comment table contains the commenter's name, comment, associated measure, topic (if applicable), and the developer or NQF response, where appropriate. Please review this table in advance of the call and refer to it as we 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 17-18, 2020. 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.

Comments and Their Disposition

Measure-Specific Comments

2496: Standardized Readmission Ratio (SRR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

One commenter raised concern regarding reliability, specifically, the decline in the overall IUR since its last review and an absence of reliability results stratified by facility size. Additionally, the commenter expressed concern with the PIUR methodology as being an appropriate measure of reliability for any measure in the ESRD QIP, as this program is used to distinguish performance between providers falling in the middle of the curve to determine penalties.

Concerns were also raised with the validity testing, specifically the commenter argues that while in the expected directions, the correlations with other outcomes measures were demonstrably weak. The commenter therefore agrees with the Committee's decision to no pass the measure on validity. The Commenter also raised concern about the non-discriminate c-statistic result (0.6768), arguing that 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.

Concerns were also raised regarding whether the increase in Medicare Advantage (MA) patients receiving dialysis and their geographic variation are appropriately accounted for in the measure testing and specifications. Specifically, the commenter 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. Additionally, the commenter raised concern about limiting comorbidity data to inpatient claims, suggesting that this may skew the models towards a sicker population and may reflect unfavorably on facilities that successfully keep hospitalization rates low.

There was also concern with harmonization between the SRR and SHR measure. The commenter mentioned that measure specifications indicate the minimum data requirement for the SHR is 5 patient-years at risk, which differs from the SRR, which uses 10 patient-years at risk. Likewise, the groupings used in the risk models for the patient age and duration of ESRD variables differ between the two measures--the SHR considers age as a continuous variable while the SRR uses three distinct age groupings, and there are four SHR groupings for ESRD duration while time on dialysis is appears to be a continuous variable in the SRR model.

Measure Steward/Developer Response:

Reliability/IUR:

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.”

For each of the four measures considered here, the IUR is less than the PIUR indicating that these measures are more useful for identifying the more extreme values that are present in the data. For example, for the SHR, the PIUR of 0.75 indicates that this measure based on the single year outcomes is moderately to highly reliable with respect to identifying extreme outcomes, which tend to be important in this measure. Similarly, the PIUR of 0.89 for the SEDR indicates a very high reliability with respect to the more extreme outcomes whereas the IUR of 0.62 is more modest. SRR and ED30 have the lowest PIURs, at 0.61 and 0.57, respectively. These are still moderate values indicating that the measures have value in identifying the more extreme facilities. Note that measures with low IUR but higher PIUR are still useful in giving a ranking of facilities, but it is best to concentrate on the more extreme values and not to use the measure to distinguish among facilities whose values are in the intermediate range. This comment also applies, perhaps with less force, to measures with higher IUR due to the effect of unmeasured confounders. We have chosen the 2.5% critical point in defining the PIUR, but we could have chosen other values (e.g. 5% or 10%) and so flagged more or less facilities than we have reported.

Regarding the assertion that the Scientific Methods Panel found the PIUR to be an inappropriate measure of reliability, we respectfully disagree with that interpretation of the panel's discussions. The developer was asked to present information regarding the PIUR to the Admissions/Readmissions committee, and it was met with interest and a constructive discussion.

Decisions on whether or not to include a measure in a particular quality reporting program is beyond the scope of measure endorsement. Therefore, the concerns regarding using the PIUR to determine whether a measure is sufficiently reliable for the ESRD QIP is beyond the scope of the endorsement process.

Validity

See request for reconsideration, submitted 9/3/2020

Medicare Advantage

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.

Risk Models

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.

Developer Request for Reconsideration:

The developer is requesting reconsideration of the Standardized Readmission Ratio (SRR) on the basis that the measure evaluation criteria were not applied appropriately. The developer stated that the Admissions/Readmissions Committee voted 18-0 in favor of upholding the Scientific

Methods Panel (SMP) recommendation not to pass the measure on validity because of inadequate demonstration of measure score validity based on correlations with other outcome measures. The developer contends that the results from validity testing are sufficient for achieving a moderate score on validity. They respectfully request reconsideration from the committee on this criterion. The developer's request and rationale can be found in Appendix B.

Proposed Committee Response:

Thank you for your comments and for the developer's request for consideration for NQF 2496. The Committee will review the comments and this reconsideration request during its deliberations on the Post-Comment Call scheduled on September 24, 2020. The Committee will also discuss any opportunity for harmonization.

Action Item:

The Committee should review the review the comments, the developer's responses, and the request for reconsideration be prepared to decide whether or not to reconsider the measure.

1463: Standardized Hospitalization Ratio (SHR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

One commenter raised concern regarding reliability, specifically, the decline (0.53-0.59 for 2015-2018) in the overall IUR since its last review (0.70 – 0.72 from 2010-2013) and an absence of reliability results stratified by facility size. Additionally, the commenter expressed concern with the PIUR methodology as being an appropriate measure of reliability for any measure in the ESRD QIP, as this program is used to distinguish performance between providers falling in the middle of the curve to determine penalties.

Concerns were also raised with the validity testing, specifically the commenter argues that multicollinearity and the non-discriminate c-statistic result. The commenter also raised concern regarding whether the increase in Medicare Advantage patients receiving dialysis and their geographic variation are appropriately accounted for in the measure testing and specifications. Specifically, the commenter 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. Additionally, the commenter raised concern about limiting comorbidity data to inpatient claims, suggesting that this may skew the models towards a sicker population and may reflect unfavorably on facilities that successfully keep hospitalization rates low. There was also concern with harmonization between the SRR and SHR measure.

Lastly, one commenter mentioned that patients residing in a nursing home is important characteristic to account for in all of the dialysis facility measures (hospitalizations, ED visits, and readmissions), but it is not clear if this was accounted for in the hospitalizations and 30-day ED encounter measures.

Measure Steward/Developer Response:

All three measures (1463, 3565, 3566) are adjusted for nursing home status. We appreciate the suggestion to stratify the measure by patients living in the community versus in a nursing home – we will take that into consideration as we further refine the risk adjustment models.

Reliability/IUR:

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.”

For each of the four measures considered here, the IUR is less than the PIUR indicating that these measures are more useful for identifying the more extreme values that are present in the data. For example, for the SHR, the PIUR of 0.75 indicates that this measure based on the single year outcomes is moderately to highly reliable with respect to identifying extreme outcomes, which tend to be important in this measure. Similarly, the PIUR of 0.89 for the SEDR indicates a very high reliability with respect to the more extreme outcomes whereas the IUR of 0.62 is more modest. SRR and ED30 have the lowest PIURs, at 0.61 and 0.57, respectively. These are still moderate values indicating that the measures have value in identifying the more extreme facilities. Note that measures with low IUR but higher PIUR are still useful in giving a ranking of facilities, but it is best to concentrate on the more extreme values and not to use the measure to distinguish among facilities whose values are in the intermediate range. This comment also applies, perhaps with less force, to measures with higher IUR due to the effect of unmeasured confounders. We have chosen the 2.5% critical point in defining the PIUR, but we could have chosen other values (e.g. 5% or 10%) and so flagged more or less facilities than we have reported.

Regarding the assertion that the Scientific Methods Panel found the PIUR to be an inappropriate measure of reliability, we respectfully disagree with that interpretation of the panel’s discussions. The developer was asked to present information regarding the PIUR to the

Admissions/Readmissions committee, and it was met with interest and a constructive discussion.

Decisions on whether or not to include a measure in a particular quality reporting program is beyond the scope of measure endorsement. Therefore, the concerns regarding using the PIUR to determine whether a measure is sufficiently reliable for the ESRD QIP is beyond the scope of the endorsement process.

C-statistics (SHR)

Over the history of the development and refinement of the SRR and SHR, we evaluated multiple iterations to obtain the best model possible. Based on recent literature, the C-statistic of 0.621 indicates a model with good predictive properties and is similar in magnitude to the C-statistics of other current NQF endorsed quality measures that have been implemented by CMS. In the table below we list examples of these measures with the corresponding C-statistic.

NQF ID Name C-Statistic

- 1789 Hospital-wide Readmission Measure 0.64
- 1550 Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) 0.65
- 0173 Emergency Department Use without Hospitalization During the First 60 Days of Home Health 0.63
- 2539 Facility 7-Day Risk Standardized Hospital Visit Rate after Outpatient Colonoscopy 0.68

Prevalent Comorbidities in the SHR

Based on input from a TEP, we have selected comorbidities that were expected to be associated with hospitalization rates and not influenced by the quality of facility care. This has the advantage that the model is specified independently of the data and so should be more, not less, generalizable to other situations. A feature of this approach, however, is that collinearities can result in estimates that seem counter intuitive since the change that is being measured is that due to the variables with all other covariates fixed, and not easily interpreted. This is generally the case with models including many covariates as are commonly used in risk adjustment. As part of measure maintenance, we routinely review comorbidity selection in the model to ensure that the constellation of adjustors remains robust over time.

Response re Medicare Advantage: 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.

Response re: Risk Model: 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.

Proposed Committee Response:

Thank you for the comments. During the Spring 2020 measure evaluation call on June 22, 2020, the Committee discussed several topics related to the scientific acceptability of the measure, including the PIUR methodology. The Committee determined that this method was appropriate and passed the measure on reliability. For validity, the Committee discussed validity testing and risk adjustment approaches for this measure. The Committee considered input from the Scientific Methods Panel and those form an NQF-convened renal Technical Expert Panel (TEP), which generally agreed that the correlations for validity testing and the risk adjustment model testing were appropriate. Additionally, during the Spring 2020 measure evaluation meeting, the Committee sought clarification on the use of inpatient claims only for Medicare Advantage (MA) beneficiaries. The Committee discussed that the use of inpatient claims for MA beneficiaries was because the outpatient claims are not available for most qualifying patients. The Committee ultimately passed the measure on validity. Further, the Committee will also discuss any opportunity for harmonization.

Action Item:

The Committee should review the comments and the developer's response and determine if they approve the proposed response.

2539 Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy ((Yale Center for Outcomes Research and Evaluation (CORE)/Centers for Medicare and Medicaid Services (CMS))

Commenters raised concerns with the adequacy of the social risk factors (SRFs) inclusion analysis and the multi-step order/approach in which the social risk factors were assessed (after clinical risk factor adjustment) in the risk model. One commenter believes that the variations in the risk factor adjustment could impact how clinical or social variables perform in the model. Several commenters recommended that the developer demonstrate how the rates for facilities would shift across the three categories used for public reporting (better than the national average, no different than the nation average, or worse than the national average) prior to passing this measure on the validity criterion.

Measure Steward/Developer Response:

Thank you for your comment. We have responded to each of the issues you identified in your comment in our response below.

Approach to Social Risk Factor Adjustment

We appreciate the opportunity to clarify our approach for the identification and testing of social risk factors versus risk factors such as clinical comorbidity and procedural complexity.

For risk-adjusted outcome measures, CMS first considers adjustment for clinical comorbidities and frailty indicators, and then examines additional risk imparted by social risk factors after the potential for greater disease burden is included in the risk model. We believe that this is consistent with NQF current guidance, as well as the approach used by the Office of the Assistance Secretary for Planning and Evaluation (ASPE) in their recent report to Congress (ASPE 2020). It is also appropriate given the evidence cited in our submission that people who experience greater social risk are more likely to have more disease burden compared with those who have less social risk; and that this is clearly not a signal of hospital quality. In addition, according to NQF guidance, developers should assess social risk factors for their contribution of

unique variation in the outcome – that they are not redundant (NQF, 2014). Therefore, if clinical risk factors explain all or most of the patient variation in the outcome, then NQF guidance does not support adding social risk factors that do not account for variation.

There are tradeoffs inherent in adjusting for social risk factors; adjusting potentially masks disparities in care, and potentially reduces incentives to address the needs of patients with social risk factors during the provision of care. On the other hand, not adjusting for social risk factors that are related to the outcome and cannot practically be mitigated through better care has downsides, including dis-incentivizing care for patients with social risk factors. Clinical risk factors don't impose these same tradeoffs. Hence, we tested the marginal effect of social risk factors after adjusting for clinical risk factors to inform consideration of these tradeoffs by CMS, experts and stakeholders.

We considered social risk factors during measure development as an integral part of the measure development CMS's process, and with CMS reconsidered them during this application process. Our goal was to make the analytic results fully transparent and inform CMS and stakeholder decision-making as well as NQF review. As noted above, there are pros and cons of including such factors in risk adjustment, and the decision can be informed by the conceptual model defining their relationship to the outcome, stakeholder insights and preferences. We note that this measure was first endorsed by NQF in 2014 and is currently in use in the Hospital Outpatient Quality Reporting (HOQR) Program and the Ambulatory Surgery Center Quality Reporting Program (ASCQR), that CMS has decided not to add social risk factors at this time given the testing results as described in the application (section 2b3.4b.). In addition, since the submission of the NQF re-endorsement application, ASPE published its latest report to Congress which states that "quality and resource use measures should not be adjusted for social risk factors for public reporting" and that it is "important to hold providers accountable for overall results, regardless of social risk." (ASPE 2020).

The commenter also states that they "remain concerned with the lack of adequate analysis of the inclusion of social risk factors in the risk adjustment approach." We provided a thorough analysis of the relationship of social risk factors to the outcome, model performance, and the measure scores, consistent with NQF guidance. Individuals on the Scientific Methods Panel who reviewed the measure supported our analytic approach. Quoting directly from the Measure Evaluation Worksheet, Scientific Methods Panel members noted that "the analysis of social risk factors is very thoughtfully and carefully done."

Reclassification

The commenters request information on how facilities' classification would change within CMS's performance categories in relation to the national average (Better, Worse, No Different) as reported on Medicare Care Compare, if the measure were risk adjusted. This analysis is resource intensive; we are not able to currently perform such an analysis due to resources constraints and due to restrictions on staff working in the office due to COVID.

There are two important points to consider, however. First, the performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95% confidence intervals. This implementation approach is unrelated to the validity and reliability of the measure, and is not part of the NQF measure specifications. Second, we would

expect that there would be minimal impact of social risk factor adjustment on hospitals' performance category: the differences in measure scores calculated with and without social risk factors are small, and measure scores calculated for facilities with and without either social risk factor are highly correlated (correlation coefficients of 0.996-0.997 for either social risk factor, for either facility type). There is also no meaningful or systematic increase in measure scores for facilities with the highest proportion of patients with social risk factors.

References:
National Quality Forum (NQF). Risk adjustment for socioeconomic status or other sociodemographic factors: Technical report. 2014; http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx. Accessed June 16, 2020.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed September 10, 2020

The commenter request information on how facilities' classification would change within CMS's performance categories in relation to the national average (Better, Worse, No Different) as reported on Medicare Care Compare, if the measure were risk adjusted. This analysis is resource intensive; we are not able to currently perform such an analysis due to resources constraints and due to restrictions on staff working in the office due to COVID.

There are two important points to consider, however. First, the performance categories are an implementation issue – CMS chooses to identify outliers based on 95% interval estimates, akin to 95% confidence intervals. This implementation approach is unrelated to the validity and reliability of the measure, and is not part of the NQF measure specifications. Second, we would expect that there would be minimal impact of social risk factor adjustment on hospitals' performance category: the differences in measure scores calculated with and without social risk factors are small, and measure scores calculated for facilities with and without either social risk factor are highly correlated (correlation coefficients of 0.996-0.997 for either social risk factor, for either facility type). There is also no meaningful or systematic increase in measure scores for facilities with the highest proportion of patients with social risk factors.

Proposed Committee Response:

Thank you for the comments. During the Spring 2020 measure evaluation call on June 22, 2020, the Committee considered and discussed several topics related to the validity of the measure, including risk adjustment meaningful differences in performance. The Committee ultimately agreed to uphold the Scientific Method Panel's rating of validity and passed the measure on this criterion.

Action Item:

The Committee should review the comments and the developer's response and determine if they approve the proposed response.

3565: Standardized Emergency Department Encounter Ratio (SEDR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

One commenter expressed concerns with the decreased reliability of the measure following the 2017 review and lack of reliability assessment across facility sizes. The commenter further expressed concern with the PIUR methodology as being an appropriate measure of reliability for any measure in the ESRD QIP, as this program is used to distinguish performance between providers falling in the middle of the curve to determine penalties.

The commenter also raised concerns with the measures ability to distinguish meaningful differences in performance, citing that the measure can only distinguish differences in performance in less than 6 percent of facilities—specifically, 2.85 percent of facilities were classified as "better than expected" and 3.05 percent as "worse than expected." The commenter mentioned that these concerns, in addition to concerns about the exclusion of Medicare Advantage patients, the all-cause construct, the lack of inclusion for urgent care center visits, and risk model fit, were previously communicated during the pre-evaluation meeting commenting period.

Measure Steward/Developer Response:

Reliability/IUR:

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."

For each of the four measures considered here, the IUR is less than the PIUR indicating that these measures are more useful for identifying the more extreme values that are present in the data. For example, for the SHR, the PIUR of 0.75 indicates that this measure based on the single year outcomes is moderately to highly reliable with respect to identifying extreme outcomes, which tend to be important in this measure. Similarly, the PIUR of 0.89 for the SEDR indicates a very high reliability with respect to the more extreme outcomes whereas the IUR of 0.62 is more modest. SRR and ED30 haves the lowest PIURs, at 0.61 and 0.57, respectively. These are still

moderate values indicating that the measures have value in identifying the more extreme facilities. Note that measures with low IUR but higher PIUR are still useful in giving a ranking of facilities, but it is best to concentrate on the more extreme values and not to use the measure to distinguish among facilities whose values are in the intermediate range. This comment also applies, perhaps with less force, to measures with higher IUR due to the effect of unmeasured confounders. We have chosen the 2.5% critical point in defining the PIUR, but we could have chosen other values (e.g. 5% or 10%) and so flagged more or less facilities than we have reported.

Regarding the assertion that the Scientific Methods Panel found the PIUR to be an inappropriate measure of reliability, we respectfully disagree with that interpretation of the panel's discussions. The developer was asked to present information regarding the PIUR to the Admissions/Readmissions committee, and it was met with interest and a constructive discussion.

Decisions on whether or not to include a measure in a particular quality reporting program is beyond the scope of measure endorsement. Therefore, the concerns regarding using the PIUR to determine whether a measure is sufficiently reliable for the ESRD QIP is beyond the scope of the endorsement process.

Meaningful differences (ED30/SEDR)

The IUR and the PIUR for the two-year ED30 (2016-2017) are 0.451 and 0.570, respectively. The IUR and the PIUR for the one year SEDR (2017) are 0.62 and 0.89, respectively. If there are no outliers, the PIUR and IUR are similar in size, but in cases where there are outliers, measures with a low IUR can have a relatively high PIUR and be useful for identifying extreme providers. The higher PIURs in both measures demonstrate that ED30 and SEDR are useful in identifying outlying facilities, the SEDR being more useful in this regard than the ED30. We used 5% as the nominal significance level and higher flagging rates would be achieved with higher levels.

Proposed Committee Response:

Thank you for your comments. During the Spring 2020 measure evaluation call on June 22, 2020, the Committee discussed several topics related to the scientific acceptability of the measure, including the PIUR methodology. The Committee determined that this method was appropriate and passed the measure on reliability.

NQF Response:

All pre-evaluation comments were provided to the Committee prior to the June 22, 2020 web meeting and taken into consideration by the Committee. The Committee also considered input from the Scientific Methods Panel and those form an NQF-convened renal Technical Expert Panel (TEP), and ultimately passed the measure.

Action Item:

The Committee should review the comments and the developer's response and determine if they approve the proposed response.

3566: Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge (ED30) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

One commenter posits that ED30 is not reliable as specified and measure specifications need to either indicate minimum sample size or the measure be deemed unreliable for all facilities under the current specifications. Commenter states that the degree of reliability (as indicated by an overall IUR of 0.451) is poor, further noting that the IUR for those facilities falling within the lowest tertile (0-30.4 patient-years) was only 0.31. The commenter also expressed concern with the PIUR methodology as being an appropriate measure of reliability for any measure in the ESRD QIP, as this program is used to distinguish performance between providers falling in the middle of the curve to determine penalties.

The commenter also raised concerns with the measures ability to distinguish meaningful differences in performance, citing that the measure can only distinguish differences in performance in less than 6 percent of facilities—specifically, 2.85 percent of facilities were classified as "better than expected" and 3.05 percent as "worse than expected." The commenter mentioned that these concerns, in addition to concerns about the exclusion of Medicare Advantage patients, the all-cause construct, the lack of inclusion for urgent care center visits, and risk model fit, were previously communicated during the pre-evaluation meeting commenting period.

Measure Steward/Developer Response:

Reliability/IUR:

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."

For each of the four measures considered here, the IUR is less than the PIUR indicating that these measures are more useful for identifying the more extreme values that are present in the data. For example, for the SHR, the PIUR of 0.75 indicates that this measure based on the single year outcomes is moderately to highly reliable with respect to identifying extreme outcomes, which tend to be important in this measure. Similarly, the PIUR of 0.89 for the SEDR indicates a very high reliability with respect to the more extreme outcomes whereas the IUR of 0.62 is more modest. SRR and ED30 have the lowest PIURs, at 0.61 and 0.57, respectively. These are still moderate values indicating that the measures have value in identifying the more extreme facilities. Note that measures with low IUR but higher PIUR are still useful in giving a ranking of facilities, but it is best to concentrate on the more extreme values and not to use the measure to distinguish among facilities whose values are in the intermediate range. This comment also applies, perhaps with less force, to measures with higher IUR due to the effect of unmeasured confounders. We have chosen the 2.5% critical point in defining the PIUR, but we could have chosen other values (e.g. 5% or 10%) and so flagged more or less facilities than we have reported.

Regarding the assertion that the Scientific Methods Panel found the PIUR to be an inappropriate measure of reliability, we respectfully disagree with that interpretation of the panel's discussions. The developer was asked to present information regarding the PIUR to the Admissions/Readmissions committee, and it was met with interest and a constructive discussion.

Decisions on whether or not to include a measure in a particular quality reporting program is beyond the scope of measure endorsement. Therefore, the concerns regarding using the PIUR to determine whether a measure is sufficiently reliable for the ESRD QIP is beyond the scope of the endorsement process.

Proposed Committee Response:

Thank you for your comments. During the Spring 2020 measure evaluation call on June 22, 2020, the Committee discussed several topics related to the scientific acceptability of the measure, including the PIUR methodology. The Committee determined that this method was appropriate and passed the measure on reliability.

NQF Response:

All pre-evaluation comments were provided to the Committee prior to the June 22, 2020 web meeting and taken into consideration by the Committee. The Committee also considered input from the Scientific Methods Panel and those form an NQF-convened renal Technical Expert Panel (TEP), and ultimately passed the measure.

Action Item:

The Committee should review the comments and the developer's response and determine if they approve the proposed response.

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 expressions of non-support. See Appendix A.

Appendix A: NQF Member Expression of Support Results

Three NQF members provided their expressions of support/nonsupport. Four of the five measures under consideration received support or non-support from NQF members. Results for each measure are provided below.

NQF 2496: Standardized Readmission Ratio (SRR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

Member Council	Support	Do Not Support	Total
QMRI		1	1

NQF 3565: Standardized Emergency Department Encounter Ratio (SEDR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

Member Council	Support	Do Not Support	Total
QMRI		1	1

NQF 3566: Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge (ED30) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

Member Council	Support	Do Not Support	Total
QMRI		1	1

NQF 2539: Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy (Yale CORE/CMS)

Member Council	Support	Do Not Support	Total
Health Professional		1	1
Supplier/Industry		1	1

Appendix B: Request for Reconsideration of Standardized Readmission Ratio 2496 Standardized Readmission Ratio (SRR) for Dialysis Facilities (UM Kidney Epidemiology and Cost Center/CMS)

Submitted by the University of Michigan Kidney Epidemiology and Cost Center
September 3, 2020

Introduction

We are requesting reconsideration of the Standardized Readmission Ratio (SRR) on the basis that the measure evaluation criteria were not applied appropriately. As described below, the Admissions/Readmissions committee voted 18-0 in favor of upholding the Scientific Methods Panel (SMP) recommendation not to pass the measure on validity because of inadequate demonstration of measure score validity based on correlations with other outcome measures. We contend that the results from validity testing are sufficient for achieving a moderate score on validity, as outlined below. We respectfully request reconsideration from the committee on this criterion.

Background

The SRR was reviewed by the SMP in early April, 2020. The validity discussion was described as follows in the summary of the SMP review meeting[1]:

> "For validity, the concerns centered on the adequacy of the correlations presented for measure score validity testing. The developers provided a detailed response to the panel's concerns. However, reviewers still found the results did not adequately demonstrate measure score validity and did not pass the measure on validity. NQF's most recent policy on measures that will be eligible for review by standing committees following SMP review states that measures that did not pass for a reason other than inappropriate methodology or inadequate testing, can be reconsidered and voted upon by the standing committee if the committee chooses to do so. Therefore, this measure will be eligible for consideration and re-vote by the Admissions and Readmissions Standing Committee in the Spring 2020 cycle."

Validity of the SRR was next discussed by the Renal TEP convened by NQF to inform the standing committee deliberations. With regards to the correlations that concerned the SMP, the TEP "agreed that the magnitudes and directions of the correlations were clinically appropriate with dialysis care, but the measure should consider excluding hospitalizations that are not dialysis-related[2]".

During the Admissions/Readmissions standing committee meeting, the SRR was reviewed by the committee but due to lack of quorum, an online vote was held at a later time after the meeting. The discussion of validity during the meeting was brief, without much substantive discussion among committee members or questions directed to UM-KECC, the developer, about specific concerns regarding the validity testing results. Per the draft report[3], the vote was 18-0 in favor of upholding the SMP recommendation, and was described as follows:

---

[1] http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=92642
[2] http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=92920
[3] http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=93502

- "For validity, the SMP concerns centered on the adequacy of the measure correlations presented for measure score validity testing. The developers provided a detailed response to the panel's concerns.
- However, the SMP still found the results did not adequately demonstrate measure score validity and did not pass the measure on validity.
- While several considerations were noted on the reliability, the Committee agreed to pass the measure on reliability. However, the Committee agreed to uphold the SMP's rating on validity (Y-18, N-0), which was to not pass the measure on validity."

Review of Validity Testing

Original validity testing results

Methods: We assessed the validity of the measure through comparisons of this measure with other quality measures in use, using Pearson correlation coefficients to examine the relationship between the SRR and other facility-level quality measures.

- Standardized Hospitalization Ratio (SHR)- We expect a fairly strong positive association with SHR since readmissions are also hospital admissions. Additionally, both hospitalization and readmission are a reflection of hospital utilization and increased comorbidity burden.

- Standardized Mortality Ratio (SMR)- We expect a positive association with SMR. Patients who require acute inpatient medical care represent an at-risk population for mortality since they likely have greater acute medical needs or complications from chronic comorbid conditions that put them at higher risk for death. Higher SMR will be positively associated with SRR.

- Vascular Access: Long-term catheter rate (catheter in use >=3 continuous months) – We expect a positive association between long-term catheter rate and SRR. Long-term catheters put patients at increased risk for infection and other complications. Additionally, a high long-term catheter rate also indicates a higher patient comorbidity burden at the facility level such that sicker patients who have a long-term catheter may also be more likely to be hospitalized and re-admitted after initial hospitalization. Higher long-term catheter rates will be positively associated with SRR.

- Vascular Access: Standardized Fistula Rate (SFR)– We expect a negative association between SFR and SRR. Successfully creating an AVF is generally seen as representing a robust process to coordinate care outside of the dialysis facility, and potentially reduces the likelihood of adverse events, like infection that can increase the risk of patient hospitalization and hospital readmission. Higher rates of the facility level SFR will be negatively associated with re-hospitalization as measured by SRR.

Results: The measure is positively correlated with the one-year Standardized Hospitalization Ratio for Admissions ($r = 0.39$, $p < 0.0001$), the Standardized Mortality Ratio ($r = 0.10$, $p < 0.0001$), and long term catheter use ($r = 0.04$, $p = 0.0006$). The SRR is negatively correlated with the rate of patients using a fistula ($r = -0.06$, $p < 0.0001$).

Interpretation: The SRR is a measure of hospital use, comprising many causes of hospitalization. The TEP

considered devising cause-specific SRRs but recommended the use of overall SRR measures due to various reasons, including the lack of clear consensus on which causes are modifiable by the dialysis facility and concerns about gaming the system if certain conditions are identified.

The validity of the SRR measure is also supported by its association with other known quality measures, which include both dialysis facility outcomes and practices. Higher values of SRR are associated with higher rates of hospitalization and mortality. The SRR is also correlated with other quality measures (listed above), although the correlations are small.

Response to Scientific Methods Panel

In response to the initial comments from the SMP, we provided a written justification for the correlation results prior to the in-person meeting, which we are including here as part of the request for reconsideration.

> While the Pearson correlation coefficients were lower than in the prior submission we emphasize that the hypothesized associations (correlation coefficients) are in the expected direction and all highly significant at the $p<0.0001$ level ($p<0.0006$ for SRR and LTC). We do not consider the declines to be substantial, particularly given the many changes in the underlying data and each of the measure definitions.
>
> Data: The testing for the original SRR submission in 2014 used 2009 data which pre-dates the transition to the ICD-10 diagnoses codes (used for prior year comorbidity risk adjustment in SRR). The validity testing correlations included in the current submission uses data after the transition to ICD-10 for SRR and the measures used for correlation analysis (i.e. SMR, SHR, STrR, SFR). Additionally, the prior year comorbidities in the 2014 submission of SRR were based on the HCC groupers, while the current SRR uses the clinically derived AHRQ CCS groups.
>
> Measure changes: The 2019 SHR, SMR, and vascular access measures used in the empirical validity testing for the 2020 reevaluation were notably different with respect to risk adjustments and population being measured. The SHR and SMR used in the 2014 submission only adjusted for comorbidities at ESRD incidence while the current production version of these measures used in the 2020 testing include adjustment for 210 prevalent comorbidities; SMR in 2014 was an all–patient measures versus the current SMR which is restricted to the Medicare population. Both the vascular access measures used in the 2014 testing (unadjusted fistula rate, unadjusted catheter > 90 days rate) were claims based measures and restricted to the Medicare population, while the current LTC and SFR are CROWNWeb based measures and include all patients; SFR is also adjusted for a set of prevalent and incident comorbidities; and both SFR and LTC included exclusions for limited life expectancy. Finally, the SHR, SMR, LTC, and SFR used in the current testing with SRR are the 2019 production versions (as calculated and released on Dialysis Facility Compare) and do not yet reflect our updated method for handling of Medicare Advantage patients that was applied to the current SRR under review.
>
> In light of these changes, it is perhaps not surprising that there are relatively larger changes in the correlation coefficients observed. We maintain, however, that the expected and consistent direction of the hypothesized relationships, general magnitude of the coefficients, and the statistical significance of the associations with SRR demonstrate stability from the previous to the current empirical validation results. Therefore, we argue that the empirical validity testing

results are both stable and robust to changes and updates since 2014, which in our assessment provides validation support for SRR and its empirical association with other primary and intermediate outcomes.

Alternative validity testing

In order to support our original testing results, we have further validated the SRR by conducting the following analysis (Table 1) with the same set of quality measures used in the initial validation testing, reported above. We first stratified facilities into the 'better than', 'as expected', and 'worse than expected' categories of the SRR. Next we calculated mean performance scores for several quality measures: Standardized Mortality Ratio (SMR), Standardized Hospitalization Ratio (SHR), Standardized Transfusion Ratio (STrR), Standardized Fistula Rate (SFR), and Long-term Catheter (LTC). We then compared mean performance scores across the three strata of 'better than', 'as expected', and 'worse than expected' categories for the SRR. Statistically significant outliers (i.e., better and worse than expected) were determined using the method described in section 2b4.1 (testing form) to flag facilities as better than expected and worse than expected based on the national average, at the $p<0.05$ level.

We expect better mean performance on the above quality measures for facilities classified as 'better than/as expected' for SRR compared to facilities classified as 'worse than expected', with the exception of LTC (where lower mean performance is expected). Compared to facilities that perform 'worse than expected', facilities that perform 'better than/as expected' on SRR are likely to have more successful care coordination and other processes of care in place that may help patients avoid a readmission visit in the vulnerable period following a recent discharge.

The results in Table 1 show that for each measure, mean facility performance is consistently better for facilities classified as 'better than expected' and 'as expected' compared to 'worse than expected.'

Table 1, Related measures' mean facility scores by SRR classification, 2018 DFC Release

DFC Measure	SRR: Better than Expected	SRR: As Expected	SRR: Worse than Expected	As Hypothesized?
SMR	0.94	1.00	1.12	Yes
SHR	0.79	0.99	1.33	Yes
STrR	0.89	0.98	1.34	Yes
SFR	63.85	63.31	61.38	Yes
LTC	12.40	12.51	14.45	Yes

Summary

As outlined above, we believe the validity testing results provided for the SRR are sufficient for a moderate rating, as supported by the additional testing provided in this request. We respectfully request reconsideration from the committee on this criterion.