NATIONAL QUALITY FORUM
Driving measurable health improvements together

Scientific Methods Panel Discussion Guide

SPRING 2022 EVALUATION CYCLE
March 22-23, 2022

This report is funded by the Centers for Medicare & Medicaid Services under contract HHSM-500-2017-00060I Task Order HHSM-500-T0001.

Contents

Background

The Scientific Methods Panel (SMP) provides NQF Standing Committees with evaluations of submitted complex measures' Scientific Acceptability (specifically, the "must-pass" subcriteria of reliability and validity), using NQF's standard measure evaluation criteria for new and maintenance measures.

This discussion guide contains details of the complex measures submitted for evaluation during the spring 2022 measure evaluation cycle. It also contains summaries of preliminary measure analyses and responses to these analyses composed by developers. The SMP utilizes this document during measure evaluation meetings to facilitate conversations between the SMP, measure developers, and NQF staff. This cycle, the SMP evaluated 13 complex measures. Eight are up for discussion and revote. Two have been pulled by SMP members or NQF staff for further discussion, although they have passed NQF's Scientific Acceptability criterion. Vote totals in this discussion guide are the preliminary results and reflect votes the members were able to provide prior to the meeting. In this cycle, one measure's (NQF #1460) vote totals differed between reliability and validity because one member was not able to vote on reliability. Measures that are not slated for discussion will pass with preliminary votes via consent calendar by the SMP.

After the SMP reviews measures, those that pass scientific acceptability (either by consent calendar or by passing during the meeting) move on to their respective Standing Committee for measure evaluation of the remaining NQF standard measure evaluation criteria (i.e., Importance to Measure and Report, Feasibility, Usability and Use, and requirements for Related and Competing Measures). Measures that do not pass the SMP's review can be pulled by a Standing Committee member for further discussion and revote if it is an eligible measure. Please refer to *Scientific Methods Panel: Frequently Asked Questions* in NQF's standard measure evaluation criteria for details on this process.

Measures for Discussion (Brief)

Subgroup 1

- #1460 Bloodstream Infection in Hemodialysis Outpatients (Centers for Disease Control and Prevention)
    - Reliability: H-0; M-3; L-4; I-2 No pass
    - Validity: H-0; M-4; L-4; I-2 Consensus not reached
- #0471e ePC-02 Cesarean Birth (The Joint Commission)
    - Reliability: H-0; M-3; L-4; I-4 No pass
    - Validity: H-0; M-3; L-4; I-4 No pass
- #0716e ePC-06 Unexpected Newborn Complications in Term Newborns (The Joint Commission)
    - Reliability: H-1; M-2; L-4; I-4 No pass
    - Validity: H-0; M-5; L-3; I-3 Consensus not reached
- #2820 Pediatric Computed Tomography (CT) Radiation Dose (University of California, San Francisco)
    - Reliability: H-5; M-4; L-0; I-1 Pass
    - Validity: H-1; M-7; L-1; I-1 Pass
- #3687e ePC-07 Severe Obstetric Complications (The Joint Commission)
    - Reliability: H-4; M-5; L-1; I-0 Pass
    - Validity: H-2; M-6; L-0; I-2 Pass

Subgroup 2

- #3689 First Year Standardized Waitlist Ratio (FYSWR) (University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services)
    - Reliability: H-0; M-10; L-0; I-0 Pass
    - Validity: H-1; M-5; L-4; I-0 Consensus not reached
- #3694 Percentage of Prevalent Patients Waitlisted in Active Status (aPPPW) (University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services)
    - Reliability: H-5; M-3; L-0; I-2 Pass
    - Validity: H-2; M-4; L-3; I-1 Consensus not reached
- #3695 Percentage of Prevalent Patients Waitlisted (PPPW) (University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services)
    - Reliability: H-4; M-4; L-0; I-2 Pass
    - Validity: H-2; M-4; L-3; I-1 Consensus not reached
- #3679 Home Dialsys Rate (Kidney Care Quality Alliance)
    - Reliability: H-6; M-0; L-1; I-3 Consensus not reached
    - Validity: H-2; M-2; L-3; I-3 Consensus not reached

- #3697 Home Dialysis Retention (Kidney Care Quality Alliance)
    - Reliability: H-0; M-2; L-6; I-2 No pass
    - Validity: H-0; M-5; L-3; I-2 Consensus not reached

Measures That Passed (Not Pulled for Discussion) (Brief)

Subgroup 1

- #2377 Overall Defect Free Care for AMI (American College of Cardiology)
    - Reliability: H-4; M-6; L-0; I-0 Pass
    - Validity: H-2; M-7; L-0; I-1 Pass
    - Composite: H-1; M-8; L-0; I-0 Pass

Subgroup 2

- #3659 Standardized Fistula Rate for Incident Patients (University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services)
    - Reliability: H-3; M-4; L-1; I-2 Pass
    - Validity: H-1; M-7; L-2; I-0 Pass
- #3696 Standardized Modality Switch Ratio for Incident Dialysis Patients (SMoSR) (University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services)
    - Reliability: H-0; M-6; L-2; I-0 Pass
    - Validity: H-1; M-5; L-2; I-0 Pass

Measures for Discussion (Detailed)

Subgroup 1

Measure #1460 Bloodstream Infection in Hemodialysis Outpatients

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This annual measure provides the standardized infection ratio (SIR) of bloodstream infections (BSIs) among patients receiving maintenance hemodialysis at outpatient hemodialysis facilities. BSIs are defined as positive blood cultures for hemodialysis patients, which are reported monthly by participating facilities. The SIR is reported for a yearly period (calendar year) and is calculated by dividing the number of observed BSIs into the number of predicted BSIs during the year.
- **Type of measure:** Outcome
- **Data source:** Paper Medical Records
- **Level of analysis:** Population: Regional and State
- **Risk stratification:** risk category - vascular access type
- **Sampling allowed**: No sampling methodology is used to calculate this metric.
- **Ratings for reliability:** 0 high 3 moderate 4 low and 2 insufficient → Measure does not pass with LOW rating
    - The developer's validity testing serves as a demonstration of data element reliability.

    - Regarding specifications, SMP members asked for clarification on the definition of transient patients (fewer than 30 days or 13 treatments) and if they have to be present on the first two working days of the month.
- SMP comments on reliability:
    - SMP members expressed concern about the rationale for using the highest risk vascular access type rather than the one that is currently in use for dialysis. They also expressed concerns that this may be prone to error with manual reporting.
- **Ratings for validity:** 0 high 4 moderate 4 low and 2 insufficient → Measure does not pass with LOW rating
    - Validity testing was conducted at the patient/encounter level using inter-abstractor reliability.
    - The developer calculated the percent of Blood Stream Infection (BSI) under-reporting over multiple time periods for a national sample (2015-2016, 2017-2018, and 2019-2020). They also reported state level data for Tennessee (2014), Georgia (2015), and Colorado (2017).
    - The developer also validated vascular access type (fistula, graft, tunneled central line, non-tunneled central line, and other access types).
    - The developer calculated BSI under-reporting of 33.3 percent (2015), 16.7 percent (2016), 52.2 percent (2017), 16.2 percent (2018), 14.3 percent (2019) and 33.9 percent (2020). At the state level, under-reporting of BSI was 58 percent (2014), 29 percent (2015) and 22 percent (2017).
    - Concordance with vascular access type reported was 80 percent (Fistula), 86.3 percent (graft), 93.3 percent (tunneled central line), 96.5 percent (non-tunneled central line), and 98 percent (other access type). Pooled sensitivity was high for fistula, graft, and tunneled central line (all>80%, range 81.2%-91.6%).
    - The developer notes overall improvement in national BSI under-reporting over time, with an exception granted for the first six months of 2020 due to COVID-19. The developer notes that state level BSI under-reporting showed improvement over time.
    - The developer notes that all access types had at least an 80 percent match, demonstrating high concordance.
- SMP comments on validity:
    - One SMP member noted the only documentation of accountable entity level validity was a demonstration of declining value over time (suggests improvement in quality) and an ability to distinguish among sites. Testing of patient/encounter level validity demonstrate significant underreporting of infections, at best around 15 percent and at worst up to 50 percent.
    - Multiple SMP members noted there is no discussion of the underreporting of BSIs and how that may change the rating above or below predicated BSI for ranking of facilities. Members suggested that additional information on stability of ratios over time would be helpful.
    - SMP members noted the lack of consistent kappa agreement (Vascular access type, for example, used overall agreement).
    - One SMP member raised a concern about underreporting and the potential to incorrectly show meaningful differences in performance.

- SMP members asked about non-reporting at the facility level and noted the amount and variation of under-reporting would seem to undermine the validity of the measure. Positive Predictive Value was generally low and varied by access type.
- Several SMP members noted the lack of patient level social and/or clinical risk adjustment.The developer used vascular access type as the only method of risk adjustment.

ITEMS TO BE DISCUSSED

- Are there any concerns about the overall underreporting of Blood Stream Infections?
- Are the testing results sufficient to demonstrate reliability and validity?

Measure #0471e ePC-02 Cesarean Birth

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This measure assesses the number of nulliparous women with a term, singleton baby in a vertex position delivered by cesarean birth.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Data
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: Not applicable; this measure does not use a sample
- **Ratings for reliability:** 0 high 3 moderate 4 low and 4 insufficient → Measure does not pass with LOW/INSUFFICIENT rating
    - Reliability testing was conducted at the encounter level:
    - Six sites / 15 hospitals submitted production data for one quarter of calendar year (CY) 2020. These data were used for all of the testing provided except validity testing, which used a subset of the six sites. The developer reached out to all 15 hospitals to recruit sites willing to participate in validity testing. Two pilot sites (seven hospitals) volunteered. One site is a system representing six hospitals using Epic. The seventh hospital is a stand-alone facility using Meditech.
    - No patient-level sociodemographic variables were used in the measure. There was, however, considerable variability in the distribution of patient socio-demographic characteristics across hospitals, but the developer did not analyze differences in measure rates over these variables due to the relatively small sample size.
    - Encounter level validity testing served as reliability testing. Results and methods are noted under validity testing (see below).
- SMP comments on reliability:
    - Several SMP members had concerns about the reliability testing. They noted that not all data elements were assessed, rendering this approach incomplete. Other reviewers commented that no hospital level reliability testing data was provided.
    - Regarding the results of the reliability testing, the developer noted that there was substantial difficulty in obtaining data from a second EHR system. SMP members noted that the ability to obtain accurate data across all major EHR systems was not demonstrated and raised concerns.

- Reviewers observed that the hospital level analysis did not demonstrate consistency of variations in performance across hospitals. In addition, the sensitivity of the measure numerator appeared to be a function of the testing site. One pilot site had an 88 percent sensitivity, while the second had 0 percent sensitivity. Thus, differences in performance across facilities could result from differences in coding accuracy, and not differences in true performance.
- Another comment identified substantial between-hospital variation, particularly across sites, but noted that the sample size is very small, and that the denominators varied substantially and that within-hospital variance was not estimated.
- Reviewers noted a number of outliers (without risk adjustment) and at least one hospital outside the United States (U.S.), limiting the evidence of accountable entity level reliability.
- One reviewer concluded that, "testing highlights fact that many hospitals may have systems that are not adequate for reliance on electronic transmission of data for this measure."

- **Ratings for validity:** 0 high 3 moderate 4 low and 4 insufficient → Measure does not pass with LOW/INSUFFICIENT rating
  - Validity testing conducted at the encounter level and accountable entity level:
  - A representative sample of the electronically submitted inpatient encounters was selected for re-abstraction. During the virtual visits, site staff shared their screen, navigated through the electronic health records (EHRs) of the sampled patients while Joint Commission staff manually re-abstracted each data element. To determine validity, re-abstraction findings were compared with the original electronic data submission and any disagreements were adjudicated with reasons for discrepancies noted.
  - The testing methodology included the following: (1) all clinical data elements and all editable demographic elements were scored; (2) all measure data were re-abstracted with original data having been blinded so that the re-abstraction was not biased; (3) Re-abstracted data were compared with original data for each data element to identify missing or erroneous data; (4) overall performance measure outcome rates were calculated on all cases submitted by each site.
  - Next, performance measure outcome rates were calculated on the adjudicated data for the sampled cases to demonstrate accountable entity level validity. The performance measure outcome rates were compared and agreement rates were corrected for chance variation with the kappa statistic.
  - Specificity was high for both sites and sensitivity was high for Site 1, but low for Site 2. Specificity was 96.5 percent for Site 1 and 100 percent for Site 2, and 97.7 percent overall.
  - The sensitivity was 87.5 percent for Site 1, 0 percent for Site 2 and 73.7 percent overall. The developer explained Site 2's low sensitivity by noting that cases did not qualify for the initial population as time of delivery were missing or gravida/para/term/preterm were incorrect. Site 2 uses a standalone OB documentation system that does not interface completely with Meditech. OB documentation is present in Meditech in non-discrete fields in a .pdf format. A mitigation plan was put into place for Site 2.

- Measure Outcome Agreement Rates (Table 2b.03.03) were 89.9 percent for Site 1 (Kappa 0.831), and 67.7 percent (Kappa 0.477) for Site 2.
- Data Element Agreement Rates were 92 percent overall, and were 100 percent for several variables (e.g. demographics); however, some were lower. For example, many of the "Delivery Details" variables had agreement in 80s-90s percent range.
- The C-Section Rate across hospitals varied from 0 – 72 percent.
- In the funnel plot, the developer concluded that even with the relatively small numbers of hospitals and denominator sizes in the pilot, there were high outliers identified and there was significant variation in the measure rates.
- Regarding missing data, there is variation in data completeness between Site 1 and 2 with results of 96.5 percent and 78.9 percent, respectively. Site 2 has engaged in mitigation plans to improve upon the number of missing data elements. Because Pilot Site 2 uses a stand-alone OB documentation system, data elements are not in discrete fields. Most mismatches were in the Delivery Date/Time, Estimated Gestational Age, Gravida, Para, Preterm or Term Birth fields. Of the mismatches, 57 percent were due to missing data. In comparison, Site 1 had no mismatches due to missing data.
- In the exclusions analysis, it was found that exclusions had an appreciable impact on measure rates; without excluding these cases measure rates increase overall by 17 percent, or 4.7 percentage points. Exclusion rates ranged from 0-16 percent, indicating variability over sites.
- The developer gives the following rationale for not risk-adjusting the measure: exclusion criteria were chosen to ensure that the target population would be women with nulliparous, term, singleton, vertex (NTSV) pregnancies, who have a lower risk of maternal morbidity and mortality during a vaginal birth delivery than do women who have undergone a previous C-section (American College of Obstetricians and Gynecologists [ACOG], 2014). Therefore, the population of women in the denominator as a result of the exclusions, allow the measure to focus on a more homogeneous group of women where the greatest improvement opportunity exists as evidenced by the variation in rates of NTSV cesarean births indicating clinical practice patterns may affect this rate (ACOG, 2014).

- SMP comments on validity:
  - SMP reviewers did not raise substantial concerns with the methods for validity testing, but were concerned with small sample sizes and the results of the testing.
  - There was a concern by one reviewer that the funnel plot was used improperly, noting that they can be used as a tool to identify a small percentage of deviating institutions but are not meant to be used to judge whether different groups of institutions perform differently. For this reason, the reviewer stated that quality indicators can only be validly presented in funnel plots if there is no association between the values of the quality indicator and hospital characteristics.
  - Regarding validity results, SMP members were concerned that item-level validity was not good for Site 2, rendering the average performances reported not useful.
  - One reviwer raised a concern that no formal statistical testing was performed for the exclusions analysis, in particular, where social risk factors may be more affected by exclusions, this could improperly inflate facility performance.

- There were concerns raised by some reviewers that the measure is not risk-adjusted, but that testing results show that empirical examination of risk adjustment may be warranted. One reviewer noted, "The [measure developer (MD)] should consider the use of maternal age, BMI, pre-existing comorbid conditions, obstetrical conditions (pre-eclampsia), and prolonged labor. The MD provides evidence that university-based hospitals (who presumably have higher risk patients) have similar CD rates compared to other sites. This, however, does not provide sufficient evidence to demonstrate that some of the variation in CD rates would differ in individual hospitals with very degrees of patient risk."
- One reviewer noted, "In the absence of documentation of face validity or of agreement with other measures generally accepted as related to quality the validity of the measure remains to be determined."

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer, including whether the correct data elements were assessed for each measure.
- The SMP did not pass this measure on reliability or validity with concerns about testing results. Is there further argument that the testing results are in fact sufficient to demonstrate encounter level reliability / validity?

Measure #0716e ePC-06 Unexpected Newborn Complications in Term Newborns

MEASURE HIGHLIGHTS

- New Measure
- **Description:** ePC06 is a hospital level performance score reported as the rate per 1,000 full term newborns with no preexisting conditions who had Unexpected Newborn Complications, typically calculated per year.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Data
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: Not applicable
- **Ratings for reliability**: 1 high 2 moderate 4 low and 4 insufficient → Measure does not pass with LOW/INSUFFICIENT rating
  - Reliability testing was conducted at the encounter level:
  - Two health systems participated in pilot testing, each using different EHR systems (EPIC or Cerner).
  - A total of 6,699 cases were collected from 18 hospitals. Cases were randomly sampled to obtain an even distribution of measure scores across hospitals in the sample.
  - Encounter-level validity testing served as reliability testing. Results and methods are noted under validity testing (see below).
- SMP comments on reliability:
  - There were concerns that the validity of the data elements for key outcomes were not presented. There was a concern that not all data elements were included in the data element agreement analyses. In addition, the results were incomplete as only the numerator was actually assessed.

    - Reviewers found the results of the reliability (data element validity testing) to be acceptable, but some raised concerns that the results were incomplete and only provided for a small portion of the measures, and the results were highly variable.
    - There was a concern that the sample sizes were very small, and that the denominators varied substantially. Because within-hospital variance was not estimated, there were a number of outliers (without risk adjustment) and at least one hospital outside the U.S.
- **Ratings for validity:** 0 high 5 moderate 3 low and 3 insufficient → Consensus not reached
- Validity testing was conducted at the encounter level:
    - During the validity testing, 61 sample cases were successfully re-abstracted from 14 hospitals. Four hospitals from the original pilot were not included in the validity sample due to multiple hospitals having the same accreditation identifier.
    - During the virtual visits, site staff shared their screen, navigated through the EHRs of the sampled patients, while The Joint Commission staff manually re-abstracted each data element. Re-abstraction findings were compared with the original electronic data submission and any disagreements were adjudicated with reasons for discrepancies noted.
    - The testing methodology included the following: (1) all clinical data elements and all editable demographic elements were scored; (2) all measure data were re-abstracted with original data having been blinded so that the re-abstraction was not biased; (3) Re-abstracted data were compared with original data on a data element by data element basis as well as by measure result. Measure agreement and data element rates were calculated. Clinical and demographic data were scored separately. The measure agreement rate was corrected for chance variation with the kappa statistic.
    - Thirty-two records across eight different hospitals in Site 1 exhibited a match rate of 93.8 percent. Twenty-nine records across six different hospitals in Site 2 exhibited 100 percent match rate in measure outcome. The overall kappa was 0.955.
    - There were some exceptions to this agreement: (1) the secondary diagnoses (other than Single Live Term Newborn) and the procedure codes were lower since they were not always collected according to instructions; (2) the gestational age, author date/time, and birth weight were low due to differing data sources; (3) the demographic variables of race and ethnicity also had lower agreement rates for Site 2, which was due to different data sources.
    - Despite the lower agreement rates for some noted data elements, accountable entity level validity of the overall measure score was not impacted.
    - To demonstrate meaningful differences in performance, the developer calculated a funnel plot for hospital rates of the measure.
    - Of the 18 hospitals contributing data in the pilot, two were identified as statistically significant high outliers. Rates ranged from 0.3 – 8.4 percent.
    - Missing data elements are counted as mismatch. For Site 2, there were no mismatches from missing data. For Site 1, three data elements accounted for most of the missing data: Procedure EMR display, Procedure start date time, and Diagnosis EMR display other than Single Live Term Newborn. The missing secondary diagnoses and procedure codes are due to misinterpretation of measure specifications for Site 1. The missing race and ethnicity codes for Site 2 are due to different data sources.

- A match indicates that the data submitted by the hospital matched what was reabstracted during the validation visit. In other words, a match indicates the data were not missing and were accurate. As evidenced by the results above, there is variation between Pilot Site 1 and 2 with results of 93.80 percent and 100 percent, respectively. The developer outlines the root cause for the missing data for Site 2 and mitigation plans.
- For an exclusions analysis, the developer compared the frequencies of the denominator and numerator by site before and after the exclusions. The performance scores were re-calculated and checked for any significant change after exclusions. No formal statistical test was performed for the effect of exclusion on the performance score.
- Denominator exclusions ranged from 4.8-56.7 percent, indicating variability throughout the sites. Exclusions had an appreciable impact on measure rates; without excluding these cases, the overall measure rates more than quadrupled. The largest relative exclusions were for the Children's hospital, which the developer reports is to be expected since they tend to admit a larger share of at-risk babies.
- The developer stated that risk adjustment was not needed due to the three exclusions: (1) babies with congenital malformations and genetic diseases, (2) babies with pre-existing fetal conditions such as IUGR, and (3) babies who were exposed to maternal drug use in-utero.
- To further reinforce the need to not risk adjust this measure, the developer presented conceptual evidence. The 2018 California Maternal Quality Care Collaborative (CMQCC) data from all 234 California hospitals revealed Unexpected Newborn Complication rates on average were comparable between hospitals with level II and IV NICUs when compared to lower levels of care. The developer stated that to guard against potential overcoding and undercoding, babies with a length of stay greater than 5 days will count as a moderate complication even if they do not have any complication codes. In addition, the developer stated that risk adjustment is not included for maternal conditions because this would add burden to collecting the measure. Lastly, some maternal conditions are complications of labor that affect the baby, which is what the measure is trying to assess.

- SMP comments on validity:
  - SMP members had varied opinions on the acceptability of the validity testing. For example, a concern was raised that given unexpected complications in term newborns is a rare event, the number of cases reviewed was too few. Another reviewer stated that, "Testing of ability to distinguish among sites…relates more to reliability than validity".
  - There were concerns (similar to 0471e) that the funnel plot was not applied properly.
  - With regard to the results, it was noted that the agreement between the e-extraction to chart review gold standard was excellent for Site 1 and Site 2. However, SMP members noted that not all data elements were tested.
  - There was a concern that both severe and moderate complications were included in the numerator without distinction between the two. Therefore, a site with a certain percentage of moderate complications and no severe complications might appear worse than another site with a higher percentage of severe complications but fewer moderate complications.
  - There were varied opinions raised on the lack of risk adjustment for the measure.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer, including whether the correct data elements were assessed for each measure
- The SMP did not pass this measure on reliability and did not reach consensus on validity. Is there further discussion that would provide support to demonstrate reliability / validity of the measure?

Measure #2820 Pediatric Computed Tomography (CT) Radiation Dose (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** Radiation dose is measured as the dose-length product for every diagnostic brain, skull, and abdomen and pelvis CT scan performed by a reporting facility on any child less than 18 years of age during the reporting period of 12 months. The dose associated with each scan is evaluated as "high" or "acceptable," relative to the 75th percentile benchmark for that type of scan and age of patient. Median doses are calculated at the facility level for each type of scan and age of patient stratum, and then compared with the same 75th percentile benchmark. The overall proportion of high dose exams is calculated including all CT scans.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Other
- **Level of analysis:** Facility
- **Risk stratification:** subgroup - anatomic areas and age
- **Sampling allowed**: yes; (1) for assessment of the facility's median value for each stratum and whether it is acceptable or poor (For skull and abdomen and pelvis categories, 11 exams within each age and anatomic area stratum. For the brain category, 25 exams within each age and anatomic area stratum) and (2) for assessment of the overall proportion of exams above the 75% benchmark (23 CT exams across all age and anatomic area strata)
- **Ratings for reliability:** 5 high 4 moderate 0 low and 1 insufficient → Measure passes with HIGH rating
- Reliability testing conducted at the accountable entity level:

    - Data source for updated testing: UCSF international CT Dose Registry (2016-2021) representing 23,319 pediatric CT exams per year on average.
    - Hospitals may be included in both the (1) anatomic area-age strata calculations and (2) overall facility median dose calculation as long as they meet minimum sample size requirements for at least one of the 15 anatomic area-age strata.
    - Developers use sampling with replacement of CT exams within each anatomic area and age group for each hospital with 1,000 repetitions. Within each anatomic area-age group, hospitals are split into 11 subsets based on decile distribution of sample sizes. Agreement and Cohen's Kappa (between the simulated classifications and the "true class") are calculated.
    - Agreement consistently exceeds 90 percent, and Cohen's Kappa consistently exceeds 0.81 for a sample size in the range of 8 to 11 within an anatomic area-age stratum.
    - Developers assessed reliability for scoring within anatomic area-age strata but not for the overall performance based on median radiation dose overall.

- Reliability testing was conducted at the patient/encounter level:

    - In a prior submission, the developer conducted data element validation of DLP values using both manual and electronic abstraction with Kappa values greater than 95 percent.

- SMP comments on reliability:

    - Generally, reviewers found the results to support reliability.
    - One reviewer commented that the testing methods used for measure score reliability (evaluating classification and use of Cohen's Kappa) were better suited for validity testing than reliability testing.
    - Another reviewer commented that the bootstrap approach used could over-estimate reliability and offered recommendations about how to overcome this potential limitation.
    - One reviewer noted potential typos/errors in the equations with implications for the calculated sample sizes. The developer has clarified this typo in their response.

- **Ratings for validity:** 1 high 7 moderate 1 low and 1 insufficient → Measure passes with MODERATE rating
- Validity testing was conducted at the accountable entity level:

    - Developer does not assess the relationship between the measure score and other quality measures. Rather, it presents information on the relationship between radiation dose levels and organizational factors/care processes associated with high quality care. They also led an RCT examining the impact of education feedback on radiation doses used in CT imaging.
    - Studies cited by the developer demonstrate a relationship between specific organizational structures and processes of care that serve as facilitators/barriers to dose optimization using metrics similar, but not identical, to the measure under consideration.

- Validity testing conducted at the patient/encounter level:

    - Anatomic area. The developer offers data validating assignment of CT anatomic category in adults, which they indicate is a more complicated assignment than in children. They compare an algorithm that assigns categories using CPT and ICD-10-CM codes against review of the complete medical record.

        1. Based on 978 CT exams, sensitivity was 0.86 and specificity was 0.96

    - Radiation dose (dose length product – DLP). The developer notes that DLP is a standardized data element and well-validated, relying on published work and testing within the UCSF International CT Dose Registry.

        2. DLP was reported within the plausible range for 99.6% of exams.

- In the prior submission, the developer conducted data element validation of DLP values using both manual and electronic abstraction with Kappa values greater than 95 percent.

- Risk adjustment/stratification: The developer stratifies the measure by anatomic area and age. The developer provides rationales for not adjusting by the following factors: (1) clinical indicatin/protocol, (2) patient size, and (3) social risk.
- SMP comments on validity:

  - Overall, reviewers found the testing methods acceptable.
  - One reviewer noted that patient/encounter validation reported sensitivity and specificity but did not report Kappa statistic (chance-adjusted agreement) and considered this an important limitation.
  - Reviewers did not note any concerns related to exclusions or missing data.
  - Reviewers accepted developer's justification for not risk-adjusting the measure.
  - One reviewer had a concerns with how the measure is scored and classifications are interpreted: "The problem with using the median (or 50% above) is that a high proportion of scans (e.g., 20%) could be near lethal doses but the median could be below the 75th percentile. In other words, sites with median >75 percentile almost certainly have room for improvement but very bad sites can be missed."

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- Additional clarification on method for scoring this measure and how it identifies outliers

Measure #3687e ePC-07 Severe Obstetric Complications (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Hospital-level measure scores are calculated as a risk-adjusted proportion of the number of delivery hospitalizations for women who experience a severe obstetric complication, as defined by the numerator, by the total number of delivery hospitalizations in the denominator during the measurement period. The hospital-level measure score will be reported as a rate per 10,000 delivery hospitalizations.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Data
- **Level of analysis:** Facility
- **Statistical risk model:** adjusted for maternal age and 27 preexisting conditions and pregnancy characteristics
- **Risk stratification:** two subgroups - including and excluding cases where blood transfusion was the only severe obstetric complication
- **Sampling allowed**: No sampling.
- **Ratings for reliability:** 4 high 5 moderate 1 low and 0 insufficient → Measure passes with MODERATE rating
- Reliability testing was conducted at the accountable entity level:
  - Risk stratification for SES: Economic/housing instability was included in the risk model. Race/ethnicity was examined as a stratification variable rather than risk variables. It was

determined that illumination of outcome disparities by race/ethnicity, rather than adjustment of outcomes by race/ethnicity, would best inform stakeholders and patients and be most impactful in incentivizing improvements in quality of maternal care.
    - Measure scores were calculated for eight pilot sites used for risk model development, and for the 25 individual hospitals within those eight pilot sites.
    - For reliability testing, testing was conducted at several volume thresholds, including the following: no required minimum number of delivery encounters for the year, at least 25 delivery encounters for the year, and 200 delivery encounters for the year.
    - Data from Table 2a.11.01, indicate a median reliability score of 0.991 (range: 0.982 – 0.997) for any severe obstetric complication and 0.955 (range: 0.916 – 0.983) for severe obstetric complications excluding blood transfusion-only cases.
    - For hospitals with at least 25 delivery encounters (Table 2a.11.02), the median reliability score was 0.959 (0.802-0.996) for any severe obstetric complication outcome and 0.684 (0.273-0.961) for severe obstetric complications excluding blood transfusion-only cases. The signal-to-noise reliability was higher when included hospitals had at least 200 delivery encounters for the year. Particularly for the second outcome (severe complications excluding blood transfusion-only cases: the median reliability score was 0.978 (0.867-0.996) for any severe obstetric complication outcome and 0.804 (0.377-0.961) for severe obstetric complications excluding blood transfusion-only cases.
- SMP comments on reliability:
    - The reliability testing was seen as largely acceptable.
    - It was noted that the SNR reliability results indicate very high reliability. However, these results appear to change when blood transfusion cases are excluded.
- **Ratings for validity:** 2 high 6 moderate 0 low and 2 insufficient → Measure passes MODERATE rating
- Validity testing conducted at the Encounter-Level, Accountable-Entity Level, and Face Validity:
    - Validity testing was completed for 15 individual hospitals at six pilot sites. The developer reviewed 3-4 charts for each hospital in the system and 30-36 charts at each of the individual hospitals across three different EHR vendors (Epic, Meditech, Cerner).
    - Method 1 (Re-abstraction/Clinical Adjudication) – Encounter-Level Validity & Measure Score Validity

        - A statistically representative sample of the electronically submitted inpatient encounters was selected for re-abstraction. To determine validity, re-abstraction findings were compared with the original electronic data submission and any disagreements were adjudicated with reasons for discrepancies noted.
        - All clinical data elements and all editable demographic elements were scored.
        - All measure data were re-abstracted with original data having been blinded so that the re-abstraction is not biased.
        - Re-abstracted data are compared with original data for each data element.
        - Results: The PPV rate was 100 percent at Pilot Sites 1, 2, 3, 6, and 7, and 70 percent at Pilot Site 9, with an overall PPV of 94.74 percent. Overall, the data element agreement rate for all sites was excellent at a score of 90.4 percent, site range 70-97 percent, kappa range 0.703-0.963.

- Overall performance measure outcome rates were calculated on all cases submitted by each pilot site. Next, performance measure outcome rates were calculated on the adjudicated data for the sampled cases. The performance measure outcome rates were compared, and agreement rates were corrected for chance variation with the kappa statistic.
- Results: Overall, the study revealed ePC-07 to have an excellent measure outcome agreement rate of 91.2 percent with a kappa score of 0.881 indicating almost perfect agreement.

- Method 2 (Face Validity) To assess face validity, a Qualtrics survey was produced and distributed to the members of the Technical Expert Panel (TEP) for their completion. Members rated the following statements:
    - The severe obstetric morbidity and mortality captured by the Severe Obstetric Complications eCQM is an important health outcome to measure because it is an area with room for improvement.
    - The Severe Obstetric Complications eCQM will produce reliable and valid hospital measurement of severe obstetric morbidity and mortality rates across hospitals.
    - The Severe Obstetric Complications eCQM is feasible to implement because required data are routinely collected as part of clinical care and are extractable from EHRs.
    - Hospitals can use the Severe Obstetric Complications eCQM performance results for performance improvement.
    - The risk standardized rate of severe obstetric morbidity and mortality events obtained from the Severe Obstetric Complications eCQM as specified is a critical component (that is, necessary but not all-inclusive) of defining and comparing quality of obstetric care between hospitals.
    - Results: Fifteen members of the TEP completed face validity surveys. Eighty percent of TEP members strongly agree, while 20 percent moderately agree that this is an important health outcome to measure because there is room for improvement. Eighty-seven percent strongly or moderately agree the eCQM will produce reliable and valid rates, while the remaining 13 percent of respondents somewhat agree. Similarly, 87 percent strongly or moderately agree that hospitals can use the results for performance improvement, while the remaining 13 percent of respondents somewhat agree.
    - According to the developer, variation in pilot site severe obstetric complication rates indicate a clinically meaningful quality gap in the delivery of maternal care to patients experiencing a delivery hospitalization, as some sites show results indicating higher rates of risk-standardized rates of severe obstetric complications while other sites show results indicating substantially lower risk-standardized rates of severe obstetric complications.
    - The risk model was developed by Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (CORE).
    - A risk model was developed and tested with data from eight pilot sites; a total of 60,184 delivery hospitalizations were randomly divided in a 70/30 split for a

development data set (N=42,129) and a validation data set (N=18,055) using the age, and pre-existing conditions (identified with ICD-10 codes).
    - The developer estimated the hospital-specific risk standardized obstetric complications rate (RSOCR) using a hierarchical logistic regression model (hierarchical model). This accounts for within-hospital correlation of the observed outcome among patients and accommodates the assumption that underlying differences in the quality of care across hospitals lead to systematic differences in patient outcomes.
    - Decisions to include housing/economic instability as a risk factor and race/ethnicity as a stratification factor were made a priori and were not tested or influenced by analytic results.
    - The goal in selecting risk factors for adjustment was to develop parsimonious models that included clinically relevant variables strongly associated with a severe obstetric complication outcome. This used a two-stage approach, first identifying the comorbidity or clinical status risk factors that were most important in predicting the outcome, then considering the potential addition of social risk factors. Social risk factors considered were also dependent on the availability of information in the EHR.
    - Results: The calculated C-statistic for the risk model for any severe obstetric complications was 0.74 using the development data set and 0.75 using the validation data set; the calculated C-statistic for the severe obstetric complications excluding blood transfusion-only cases measure was 0.77 using the development data set and 0.73 using the validation data set.
    - According to the developer, both models show a reasonable range between the lowest decile and highest decile of predicted ability, given the low prevalence of the outcome, demonstrating the risk-adjustment model adequately controls for differences in patient characteristics.
- SMP comments on validity:
    - The validity testing approaches were seen largely as acceptable. However, a concern was raised that the face validity testing lacked testing of the exclusion for COVID and the 34 risk adjustment variables.
    - The SMP reviewers mostly thought that the validity results were acceptable. However, it it was noted that the data element validity testing was incomplete because not all elements were tested.
    - The risk adjustment methodology was seen to be appropriate but there were some questions about how stratification by social factors (i.e. race and housing insecurity) may play out.

ITEMS TO BE DISCUSSED

- Correlation of the hospital level rates of transfusions and the non-transfusion components of the measure
- How is the social risk factor variable in the risk adjustment model (economic/housing instability) collected or recorded?
- Clarification on testing results on non-transfusion cases in the risk adjustment model.

Subgroup 2

Measure #3689 First Year Standardized Waitlist Ratio (FYSWR)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The FYSWR measure tracks the number of incident patients in a practitioner (inclusive of physicians and advanced practice providers) group who are under the age of 75 and were listed on the kidney or kidney-pancreas transplant waitlist or received a living donor transplant within the first year of initiating dialysis. For each practitioner group, the First Year Standardized Waitlist Ratio (FYSWR) is calculated to compare the observed number of waitlist events in a practitioner group to its expected number of waitlist events. The FYSWR uses the expected waitlist events calculated from a Cox model, adjusted for age and patient comorbidities at incidence of dialysis. For this measure, patients are assigned to the practitioner group based on the National Provider Identifier (NPI)/Unique Physician Identifier Number (UPIN) information entered on the CMS Medical Evidence 2728 form.
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with 18 variables
- **Sampling allowed**: N/A
- **Ratings for reliability:** 0 high 10 moderate 0 low and 0 insufficient → Measure passes with MODERATE rating
- Reliability testing was conducted at the accountable entity level:
    - The developer calculated an IUR value of 0.64 for the measure, which indicates that 64 percent of the variation in the measure can be attributed to the between-facility differences (signal) and 36 percent to the within-facility variation (noise). The developer notes a moderate degree of reliability.
    - Dialysis practitioner group practices with less than 11 eligible patients and less than two expected events were excluded from this calculation.
- SMP comments on reliability:
    - SMP members requested clarity on patient attribution. Also, they noted the developer should be clear that this is a three-year measure.
    - One SMP member raised modest concern with the measure's ability to identify variation in performance with over 94 percent of facilities classified as "average" / "as expected".
- **Ratings for validity:** 1 high 5 moderate 4 low and 0 insufficient → Consensus not reached
- Empirical validity testing was conducted at the accountable entity level:
    - The developer tested the validity of the measure by evaluating the association between the dialysis practitioner group level measure performance, and subsequent mortality and overall transplant rates among all patients attributed to the practitioner groups.
    - The developers examined the Spearman correlation between the practitioner group measure value and each of the outcomes, respectively.
    - The dialysis practitioner group level second year average mortality rates are 15.3, 15.7, and 15.9 deaths per 100 patient-years for T1, T2, and T3, respectively (trend test p=0.0607). The Spearman correlation coefficient is -0.02 (p=0.3151).

- The dialysis practitioner group level second year average transplant rates are 4.7, 3.2, and 1.8 transplants per 100 patient-years for T1, T2, and T3, respectively (trend test p=<0.01). The Spearman correlation coefficient is 0.32 (p<0.01).
- The developer noted that higher FYSWR performance correlated with higher second year transplant rate, with clear separation of transplant rates across practitioner group tertiles of performance. The direction of the relationship with mortality was as expected, with numerically lower mortality with higher performance on the FYSWR measure, though it did not achieve statistical significance.

- SMP comments on validity
  - One SMP reviewer noted that tertiles are limited in their ability to demonstrate stability vs movement among levels.
  - One member noted the developers report a high missing practitioner rate (inability to attribute) of 6.2 percent. It is unclear why this problem does not exist for #3694 and #3695, when these are all measures of waitlisting among patients on dialysis under the care of dialysis practice groups. The same data source (IDR) is used for all three measures. However, it appears with analysis furnished by the measure submitter, these cases "have similar waitlisting experience to the average." This mitigates the concern of the large amount of missingness.
  - Reviewers noted several areas of concern about the risk adjustment model for consideration by the Standing Committee regarding appropriate selection of conditions and social risk factors.
  - One SMP member noted that they were unclear if risk factors in the model were present at the onset of measurement period (e.g., data elements from CMS #2728). The reviewer noted that this is important so as to limit the risk factors to those that were present at the start of care.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- Are there any concerns about the reliability or validity testing methodology, or the results?

Measure #3694 Percentage of Prevalent Patients Waitlisted in Active Status (aPPPW)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This measure tracks the percentage of patients in each dialysis practitioner group practice who were on the kidney or kidney-pancreas transplant waitlist in active status. Results are averaged across patients prevalent on the last day of each month during the reporting year. The proposed measure is a directly standardized percentage, which is adjusted for covariates (e.g. age and risk factors).
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with 23 covariates that are grouped in seven categories
- **Sampling allowed**: N/A
- **Ratings for reliability**: 5 high 3 moderate 0 low and 2 insufficient → Measure passes with HIGH rating

- Reliability testing was conducted at the accountable entity level using the inter-unit reliability (IUR) with a bootstrap (n=100) approach.
    - The developer calculated a IUR value of 0.93 for the measure, which indicates that 93 percent of the variation in the measure can be attributed to the between-facility differences and 7 percent to the within-facility variation.
    - Dialysis practitioner group practices with less than 11 eligible patients were excluded from this calculation.
- SMP comments on reliability:
    - SMP members raised concerns on the use of patient-months as the unit of counting and analysis for both numerator and denominator. The ability to count one patient up to twelve times in the measure for an entity for a given year creates questions about the independence of the observations that go into calculating the performance rate. SMP members noted that the reliability statistics may be overestimated if the observations for a given patient are highly correlated with each other.
    - Some SMP members sought clarity on whether this measure was a true outcome measure rather than a process measure.
    - SMP members sought clarification on the distinction between being waitlisted and being waitlisted with active status.
    - One SMP member raised modest concern with the measure's ability to identify variation in performance with over 92.4 percent of facilities classified as "average" / "as expected."
- **Ratings for validity:** 2 high 4 moderate 3 low and 1 insufficient → Consensus not reached
    - The developer conducted empirical validity testing at the accountable unit level
    - The developer tested the validity of the measure by evaluating the association between the dialysis practitioner group level measure performance, and mortality and overall transplant rates among all patients attributed to the practitioner groups.
    - The developers examined the Spearman correlation between the practitioner group measure value and each of the outcomes respectively.
    - The dialysis practitioner group level average mortality rates are 17.8, 18.3, and 19.2 deaths per 100 patient-years for T1, T2, and T3, respectively (trend test p=0.002). The Spearman correlation coefficient is -0.083 ($p<0.0001$).
    - The dialysis practitioner group level average transplant rates are 5.0, 4.2 and 3.1 transplants per 100 patient-years for T1, T2, and T3, respectively (trend test p=0.002). The Spearman correlation coefficient is 0.279 ($p<0.0001$).
    - The developer noted that higher aPPPW performance correlated with higher transplant rate, with clear separation of transplant rates across practitioner tertiles of performance. The direction of the relationship with mortality was as expected, and statistically significant, with numerically lower mortality with higher performance on the measure, although the magnitude of the association was smaller for transplant rate.
- SMP comments on validity:
    - SMP reviewers had several concerns with the methodological approach to risk adjustment.
    - Reviewers sought clarity on whether the comorbidities are limited to claims prior to the measurement period. This is important so as to limit the risk factors to those that were present at the start of care.

- SMP members noted an inconsistency between the risk model equation and the description, which includes two-way interaction terms.
- One SMP member noted that it is not appropriate to adjust for missingness by including a flag in the model because it incentivizes lower submission rates.
- Again, SMP members noted concerns regarding non-independence of patient-months.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- Are there any concerns about the reliability or validity testing methodology, or the results?
- Are there concerns regarding the use of patient-months?

Measure #3695 Percentage of Prevalent Patients Waitlisted (PPPW)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This measure tracks the percentage of patients in each dialysis practitioner group practice who were on the kidney or kidney-pancreas transplant waitlist. Results are averaged across patients prevalent on the last day of each month during the reporting year. The proposed measure is a directly standardized percentage, which is adjusted for covariates (e.g. age and risk factors).
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with 23 covariates that are grouped in seven categories
- **Sampling allowed**: N/A
- **Ratings for reliability**: 4 high 4 moderate 0 low and 2 insufficient → Measure passes with HIGH/MODERATE rating
- Reliability testing was conducted at the accountable entity level using the inter-unit reliability (IUR) with a bootstrap (n=100) approach.
    - The developer calculated a IUR value of 0.9409 for the measure, which indicates that over 94 percent of the variation in the measure can be attributed to the between-facility differences and 6 percent to the within-facility variation.
    - Dialysis practitioner group practices with less than 11 eligible patients were excluded from this calculation.
- SMP comments on reliability:
    - SMP members raised concerns on the use of patient-months as the unit of counting and analysis for both numerator and denominator. The ability to count one patient up to twelve times in the measure for an entity for a given year creates questions about the independence of the observations that go into calculating the performance rate. SMP members noted that the reliability statistics may be overestimated if the observations for a given patient are highly correlated with each other.
    - SMP members sought clarification on the distinction between being waitlisted and being waitlisted with active status.
    - Some SMP members sought clarity on whether this measure was a true outcome measure rather than a process measure.

    - One SMP member noted that the denominator definition is unclear. Specifically, the member noted the measure is assigned to a dialysis practitioner group practice according to each patient's treatment history during a given month during the reporting year; however, the method as to the selection of that "given month" is unstated.
- **Ratings for validity:** 2 high 4 moderate 3 low and 1 insufficient → Consensus not reached
    - The developer conducted empirical validity testing at the accountable unit level.
    - The developer tested the validity of the measure by evaluating the association between the dialysis practitioner group level measure performance, and mortality and overall transplant rates among all patients attributed to the practitioner groups.
    - The developers examined the Spearman correlation between the practitioner group measure value and each of the outcomes respectively.
    - The dialysis practitioner group level average mortality was 17.9, 18.2, 19.2 deaths per 100 patient-years for each of the threer tertiles (T1 to T3) based on their performance on the PPPW (T1 to T3, from highest to lowest waitlisting), respectively (trend test p=0.0017). The Spearman correlation coefficient was: -0.087 (p<0.0001).
    - The dialysis practitioner group level average transplant rate is 5.3, 3.9, 3.1 transplants per 100 patient-years for T1, T2, and T3 groups, respectively (trend test p<0.0001). The Spearman correlation coefficient is 0.266 (p<0.0001).
    - The developer noted that higher PPPW performance correlated with higher transplant rate, and the relationship with mortality was also as expected by the developer, and statistically significant, with numerically lower mortality with higher performance on the PPPW measure although the magnitude of the association was smaller than for transplant rate.
- SMP comments on validity:
    - One member noted concerns about the measure's ability to identify outliers.
    - Many SMP reviewers noted concerns with the risk adjustment strategy that should be considered by the Standing Committee.

ITEMS TO BE DISCUSSED

    - Additional clarifying information from the developer
    - Are there any concerns about the reliability or validity testing methodology, or the results?
    - Are there concerns regarding the use of patient-months?

Measure #3679 Home Dialysis Rate

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Percent of all dialysis patient-months in the measurement year in which the patient was dialyzing via a home dialysis modality (peritoneal dialysis and/or home hemodialysis).
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Electronic Health Data, Electronic Health Records
- **Level of analysis:** Facility
- **Risk stratification:** Risk stratification by age, gender, race, ethnicity, and dual-eligibility
- **Sampling allowed**: N/A
- **Ratings for reliability:** 6 high 0 moderate 1 low and 3 insufficient → Consensus not reached
- Reliability testing was conducted at the accountable entity level:

    - Testing conducted with two Large Dialysis Organizations that could provide data as submitted to the primary data source used for this measure (CMS EQRS/CROWNWeb) for CY 2020. These data represented 296 Hospital Referral Regions, 5,699 facilities, 417,807 patients, and 4.5 million patient-months.
    - Facility-level signal-to-noise reliability testing was conducted using the beta-binomial test, following the approach in Adams (2009). The mean reliability (n=5,694) was 0.9989. More than 90 percent of facilities had reliability greater than or equal to 0.99. The smallest facilities (<10 patient-months), 10th percentile had reliability of 0.92.
    - The mean reliability of scores aggregated to Hospital Referral Region (HRR) level were 0.9943 (minimum 0.9435).
- SMP comments on reliability
    - Regarding the specifications, two reviewers expressed concerns with lack of independence among patient-months for the same patient. Another reviewer questioned the construction of this measure as an outcome rather than a composite with the Home Dialysis Retention measure to avoid unintended consequences of a stand-alone home dialysis rate measure.
    - Some SMP members sought clarity on whether this measure was a true outcome measure, rather than a process measure.
    - Two reviewers had questions about the accountable entity for attribution. One noted that the measure uses a mix of HRRs and separate facilities, making it difficult to explain and interpret the results. One asked what units (e.g., HRRs?) are being compared and who is responsible for improvement.
    - Regarding reliability testing methods, two reviewers expressed concerns about use of the beta binomial method, with one noting a general limitation of the model that when p hat is 0 or 1, reliability will be 1. The other noted that it should not be used due to within-person correlation associated with patient-months as the unit of analysis and lack of independence of observations.
    - Reviewers also expressed concerns about alignment between the level of testing (individual facility) and the level of measure reporting (HRR).
    - The reviewers raising concerns about the methods also raised questions about the very high reliability results. The methods may not have been appropriate (non-independence of observations, high percent of facilities with 0 percent rate resulting in zero error variance in the formula, small sample sizes where sampling error is not accounted for in the beta binomial method), which may not adequately demonstrate reliability despite high results.
- **Ratings for validity:** 2 high 2 moderate 3 low and 3 insufficient → Consensus not reached
- Validity testing was conducted at the accountable entity level:
    - The developer aggregated measures scores to obtain percent home dialysis at the HRR level and compared this to the CMS "Percent Home Dialysis Utilization by HRR" (from 2018, most recent year available), using the Pearson Correlation Coefficient. The Pearson Correlation Coefficient result was 0.706 ($p<0.0001$).
    - The developer also performed systematic assessment of face validity of the measure score by convening an expert panel of nine members (five providers, two facilities, three manufacturers) and asked:

1. How likely is it that the measure score(s) provides a fair and accurate reflection of the quality of care provided in this area? (highly unlikely; unlikely; neither likely nor unlikely; likely; highly likely). A total of 88.9 percent (8 of 9) rated highly likely or likely.

- What is the likelihood that the measure score(s) can be used to effectively distinguish real differences in performance between providers in this area? (highly unlikely; unlikely; neither likely nor unlikely; likely; highly likely). A total of 88.9 percent (8 of 9) rated highly likely or likely. One panel member rated as unlikely; developer provided reason for rating and Standing Committee response.
- Missing data was evaluated for one of the two Large Dialysis Organizations (LDOs), representing more than 2 million denominator patient months. Missing data were rare overall and most common for discharge status (0.004% denominator months), nursing home LTCF residence status (4.2% denominator months), and insurance status (0.8% denominator months).
- When all exclusions were applied, less than 10 percent of patient months were removed from the denominator with an estimated effect of a 1.5 percentage point change in the measure score (13% without exclusions, 14.5% with).
- Regarding risk adjustment/stratification, the developer opted not to risk-adjust the measure and determined that stratification was more appropriate. A conceptual model is provided based on published literature and internal analysis. Poisson regression modes were used to estimate adjusted outcomes. Age, race, and dual eligibility were statistically significant, but there were small changes in the overall measure scores.
- Based both on the small impact on measure performance and the developer's perspective that risk adjustment could obscure important disparities, the determination was made not to risk-adjust and to stratify the measure instead.

- SMP comments on validity
  - Several SMP reviewers questioned the appropriateness of testing the measure against another measure that is so similar.
  - Several SMP reviewers were concerned that testing was not done at the required level analysis, that the empirical testing was done at the HRR level instead of the facility level.
  - The face validity testing also caused concerns with one reviewer noting the absence of patient or caregivers from the face validity assessments and about conflict of interest by having members of the organization developing the measure also voting on face validity. Generally, reviewers found the face validity testing acceptable except otherwise.
  - One reviewer suggested additional exclusions should be considered. One reviewer commented that patients enrolled in hospice, residing in a nursing home or other LTCF should not be excluded.
  - Some reviewers questioned the justification for the decision not to risk-adjust. Two reviewers commented that there should be more description of how risk stratification would be implemented in practice.
  - Three reviewers expressed concern about the high rates of identification of high/low outliers.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer

- Are the methods appropriate for testing reliability for this measure?
- Is the measure calculation by individual facility or HRR as structured appropriate?
- Is the high rate of identification of outliers in the reported measure scores a cause for concern?
- Are the face validity results sufficient to meet the validity requirements?

Measure #3697 Home Dialysis Retention

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Percent of all new home dialysis patients in the measurement year for whom >=3 consecutive months of home dialysis was achieved. New patients are defined as those who started a home dialysis modality during the measurement year.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Electronic Health Data, Electronic Health Records, Registry Data, ESRD Quality Reporting System (EQRS)/legacy CROWNWeb Clinical Data Repository
- **Level of analysis:** Facility
- **Risk stratification:** Risk stratification by age, gender, race, ethnicity, and dual-eligibility
- **Sampling allowed**: N/A
- **Ratings for reliability:** 0 high 2 moderate 6 low and 2 insufficient → Measure does not pass with LOW rating
- Reliability testing was conducted at the accountable entity level:
  - Testing conducted with two Large Dialysis Organizations that could provide data as submitted to the primary data source used for this measure (CMS EQRS/CROWNWeb) for CY 2020. These data represented 292 Hospital Referral Regions, 2,581 facilities, and 24,858 patients.
  - Facility-level signal-to-noise reliability testing was conducted using the beta-binomial test, following the approach in Adams (2009). The mean reliability (n=2,581) was 0.5241. Fifty percent of facilities had a reliability score of 1. By sample size, mean reliability varies from 0.41 to 0.66. Mean reliability of scores aggregated to HRR level: 0.3787.
- SMP comments on reliability:
  - Some SMP members sought clarity on whether this measure was a true outcome measure rather than a process measure.
  - Regarding reliability testing methods, some reviewers expressed concerns about use of the beta binomial method and with the small sample sizes.
  - Reviewers also expressed concerns about alignment between the level of testing (individual facility) and the level of measure reporting (HRR).
  - One reviewer expressed concern about the lack of any volume threshold given the potential for very small denominators among many facilities.
  - Reviewers generally found that the resulting reliability scores were below an acceptable threshold.
- **Ratings for validity:** 0 high 5 moderate 3 low and 2 insufficient → Consensus not reached
- Validity testing – Systematic Assessment of Face Validity of Measure Score:
  - Expert panel of nine members (five providers, two facilities, three manufacturers) were given the specifications, measure scores, and performance distributions and asked:

1. How likely is it that the measure score(s) provides a fair and accurate reflection of the quality of care provided in this area? (highly unlikely; unlikely; neither likely nor unlikely; likely; highly likely). A total of 77.77 percent (7 of 9) rated highly likely or likely.
2. What is the likelihood that the measure score(s) can be used to effectively distinguish real differences in performance between providers in this area? (highly unlikely; unlikely; neither likely nor unlikely; likely; highly likely). A total of 77.77 percent (7 of 9) rated highly likely or likely.

- Two panel members rated as "neither likely nor unlikely."
- The developer notes that the paired measure set was rated as highly likely or likely by eight of nine panel members.

- Missing data were evaluated for one of the two Large Dialysis Organizations (LDOs), representing more than 10,000 denominator patients. Missing data were rare.
- When all exclusions were applied, approximately 5 percent of patients were removed from the denominator with an estimated effect of a 2.8 percentage point change in the measure score (83.2% without exclusions, 86.0% with).
- The developer did not conduct independent risk adjustment for this measure. They stated: "The Home Dialysis Retention Measure denominator is built from our Home Dialysis Rate Measure numerator. As such, we did not perform a separate risk adjustment analysis for the Retention Measure."
- SMP comments on validity:
  - The face validity testing also caused concerns with one reviewer noting the absence of patient or caregivers from the face validity assessments and about conflict of interest by having members of the organization developing the measure also voting on face validity. Reveiwers sought clarification on the two expert panel members who expressed disagreement, given the small number of panel members.
  - One reviewer suggested additional exclusions should be considered. One reviewer commented that patients enrolled in hospice, residing in a nursing home or other LTCF should not be excluded.
  - Some reviewers questioned the justification for the decision not to risk-adjust, noting that risk adjustment modeling was not conducted for this measure; instead, the developer relied on risk adjustment data for #3679. Two reviewers commented that there should be more description of how risk stratification would be implemented in practice.
  - Four reviewers expressed concern about the high rates of identification of high/low outliers.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- Are the methods appropriate for testing reliability for this measure?
- Is the measure calculation by individual facility or HRR as structured appropriate?
- Is the high rate of identification of outliers in the reported measure scores a cause for concern?
- Are the face validity results sufficient to meet the validity requirements?

- Is the developer's rationale for not assessing risk adjustment independently (of the paired measure NQF #3679) for this measure sufficient or should independent risk adjustment analyses be conducted?

Appendix A: Measures That Passed (Not Pulled for Discussion) (Detailed)

Subgroup 1

Measure #2377 Overall Defect-Free Care for AMI

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** The proportion of acute MI patients >= 18 years of age that receive "perfect care" based upon their eligibility for each performance measures
- **Type of measure:** Composite
- **Data source:** Other
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: N/A
- **Ratings for reliability:** 4 high 6 moderate 0 low and 0 insufficient → Measure passes with MODERATE rating
- Reliability testing conducted at the accountable entity level:
    - Testing conducted using a national registry for CY 2019 with 695 hospitals and 130,279 patients represented.
    - Split sample testing (cohort split into two random samples) with calculation of Pearson correlation coefficient and Cronbach coefficient.
    - Mean scores: sample 1= 0.5711 (SD=0.22); sample 2=0.5729 (SD=0.22)
    - Pearson correlation coefficient: 0.87685
    - Cronbach Coefficient: 0.93438
- The developer's validity testing serves as a demonstration of data element reliability.
- SMP comments on reliability:
    - Reviewers generally found the testing methods to be acceptable.
- **Ratings for validity:** 2 high 7 moderate 0 low and 1 insufficient → Measure passes with MODERATE rating
- Validity testing: empirical validity testing of the composite measure score:
    - Compared hospital performance (n=526) on the composite measure of "defect-free care" (2019 data) and 30-day risk-standardized mortality rates for AMI (2013-2014 most recent data) and examined the distribution and correlation (Pearson correlation coefficient) of the two measures
    - Hypothesis: Defect-free care processes for AMI may be associated with lower mortality rates.
    - The developers found a similar distribution of hospitals by volume across both measures.
    - Pearson correlation coefficient: -0.09596 (p=0.0279)

    - Developer comment on low correlation: "The low correlation may be explained by the fact that there are a number of other unmeasured factors that could contribute to 30-day mortality rates beyond whether defect free care was delivered in-hospital (e.g., unsuccessful procedure, lack of follow-up, poor medication adherence or access to care). Further, the 30-day time period started upon admission to the hospital thus the rates also accounted for in-hospital mortality."
- Validity testing: systematic face validity of the measure scores (initial testing):
    - Seventeen-member oversight committee and 12-member steering committee, but no description of the consensus processes was used or any criteria. The developer states: "The face/content validity of this measure has been achieved by virtue of the noted expertise of those individuals who developed this measure."
    - Results: "Face validity was achieved through reaching consensus that the measure had strong clinical evidence and was reliable."
- Validity testing: patient/encounter level validity:
    - National Cardiovascular Data Registry Data Quality Program has validation checks of data completeness (missing data), consistency (logically related fields have values consistent with other fields), and accuracy (agreement between registry data and chart reviews).
    - Data accuracy results
    1. Categorical data assessed using prevalence-adjusted and bias-adjusted kappa (93,748 data points): 0.939
    2. Continuous data assessed using Pearson Correlation Coefficient (23,206 data points): 0.888

- Abstractor inter-rater reliability
    3. PABAK (8,139 data points): 0.971
    4. Pearson (1,781 data points): 0.990
- SMP comments on validity testing:
    - At the patient/encounter level, reviewers generally agreed that data element validity was high.
    - At the accountable entity level, reviewers noted the overall weak association and noted this may be due to the measure being a composite of process measures or to the different time frames used.
    - Two reviewers noted that patient-level analysis (instead of facility level) evaluating whether defect-free care is associated with lower mortality would be more appropriate, with one reviewer suggesting using a multi-level regression model.
    - One reviewer noted that 42 percent of AMIs are not eligible for inclusion in the measure, expressing the concern that if there are systematic differences in performance for eligible and ineligible patients, the measure may not be a good indicator of quality.
- **Ratings for composite construction:** 1 high 8 moderate 0 low and 0 insufficient → Measure passes with MODERATE rating

- Assessed correlation between each of the 15 hospital-level component measures with the composite using the Pearson correlation coefficient
- Pearson correlation coefficients ranged from -0.06 to 0.7279 across the 15 components. Two components were identified as not having statistically significant associations. Developer comment: "While only some of the components in the composite may not have had a moderate to strong correlation to the overall composite, all are based on Class IA or B recommendations and represent optimal clinical care for patients admitted for STEMI or NSTEMI treatment."
- Components that were incorporated previously (prior NQF submissions) and identified as "topped out" were removed from the composite.

- SMP comments on composite construction:
  - Reviewers generally agreed that the composite construction was appropriate, noting that most had moderate to strong correlation. Some reviewers noted the significant variation in correlation of individual components with the overall score.

Subgroup 2

Measure #3659 Standardized Fistula Rate for Incident Patients

MEASURE HIGHLIGHTS

- New Measure
- **Description:** Adjusted percentage of adult incident hemodialysis patient-months using an autogenous arteriovenous fistula (AVF) as the sole means of vascular access. The Standardized Fistula Rate (SFR) for Incident Patients is based on the prior SFR (NQF #2977) that included both incident and prevalent patients. This measure was initially endorsed in 2016, but as part of measure maintenance review by the NQF Standing Committee in 2020, concerns were raised about the strength of evidence supporting the prior measure. Namely, recent updates to the KDOQI guidelines downgraded the evidence supporting fistula as the preferred access type and instead focus on catheter avoidance and developing an individualized ESKD Lifeplan. However, the guidelines do suggest that under favorable circumstances an AV fistula is preferred to an AV graft in incident patients due to fewer long-term vascular access events. Given that over 80% of incident dialysis patients begin treatment with a tunneled catheter, and that 12 months after dialysis initiation AV fistula rates exceed 60%, the incident SFR was developed to focus on the subset of dialysis patients that the evidence suggests may benefit the most during a time of intense vascular access creation. Specifically, blood stream infection rates are the lowest in incident patients with AV fistula compared to long-term catheters. Therefore the goal of this new measure is to evaluate facility performance in increasing fistula use in the incident population in order to reduce the heightened risks patients face due to bacteremia and infection related hospitalizations.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 16 risk factors
- **Sampling allowed**: N/A
- **Ratings for reliability:** 3 high 4 moderate 1 low and 2 insufficient → Measure passes with MODERATE rating

- Reliability testing was conducted at the accountable entity level using the inter-unit reliability (IUR) with a bootstrap approach, and Profile IUR (PIUR).
    - The developer calculated a IUR value of 0.705, which indicates that 70.5 percent of the variation in the Incident SFR can be attributed to between-facility differences in performance (signal) and 29.5 percent to the within-facility variation (noise).
    - The developer also calculated a PIUR of 0.970. They noted that this value is higher compared to the IUR, indicating the existence of outlier facilities.
    - This calculation included facilities with at least 11 patients during the two-year period. However, the developer states in 2a.05 "Patients at those facilities with <11 attributed patients are still included in our modeling and are not excluded." The model specifications and reliability testing alignment should be clarified.
    - SMP members raised concerns on the use of patient-months as the unit of counting and analysis for both numerator and denominator. The ability to count one patient up to twelve times in the measure for an entity for a given year creates questions about the independence of the observations that go into calculating the performance rate. SMP members noted that the reliability statistics may be overestimated if the observations for a given patient are highly correlated with each other.
    - One SMP member raised modest concern with the measure's ability to identify variation in performance with over 92 percent of facilities classified as "average" / "as expected".
- **Ratings for validity:** 1 high 7 moderate 2 low and 0 insufficient → Measure passes with MODERATE rating
- The developer conducted empirical validity testing at the accountable unit level.
    - The developer created performance categories with the higher quintiles representing better care.
    - The developer assessed validity by using a Poisson regression model to examine the association between facility level quintiles of performance scores and:

1. 2018-2019 Standardized Mortality Ratio (SMR, NQF #0369)
    - The developer found that the relative risk of mortality increased as the performance measure quintile decreased from the reference group (Q5) with the highest risk in quintile 1. For quintile 4, RR=1.02 (95% CI: 1.00, 1.04; p<0.001), quintile 3, RR=1.06 (95% CI: 1.04, 1.08; p<0.001), quintile 2, RR=1.08 (95% CI: 1.06, 1.10; p<0.001), and quintile 1, RR=1.13 (95% CI: 1.11, 1.15; p<0.001).
2. 2018-2019 Standardized Hospitalization Ratio (SHR, NQF #1463)
    - The developer found that the relative risk of hospitalization increased as the performance measure quintile decreased from the reference group (Q5) with the highest risk in quintile 1. For quintile 4, RR=1.06 (95% CI: 1.05, 1.06; p<0.001), quintile 3, RR=1.07 (95% CI: 1.06, 1.07; p<0.001), quintile 2, RR=1.11 (95% CI: 1.10, 1.12; p<0.001), and quintile 1, RR=1.15 (95% CI: 1.14, 1.15; p<0.001)
3. 2018 First Year Standardized Mortality Ratio (SMR).
    - The developer found that the relative risk of mortality increased as the performance measure quintile decreased from the reference group (Q5)

with the highest risk in quintile 1. For quintile 4, RR=1.08 (95% CI: 1.03, 1.14; p=0.002), quintile 3, RR=1.11 (95% CI: 1.05, 1.16; p<0.001), quintile 2, RR=1.17 (95% CI: 1.12, 1.23; p<0.001), and quintile 1, RR=1.53 (95% CI: 1.46, 1.60; p<0.001).

4. 2018-2019 All-Cause Hospitalization
    - The developer found that the hospitalization rate decreased as the performance measure quintile increased. Hospitalization rates for quintiles 1 to 5 were 1.06, 0.99, 0.95, 0.93, and 0.87 patient-years respectively (trend test p<0.001).
5. 2018-2019 Vascular Access Related Infection Hospitalization
    - The developer found that the hospitalization rate decreased as the performance measure quintile increased. Hospitalization rates for quintiles 1 to 5 were 0.22, 0.18, 0.17, 0.16, and 0.15, respectively (trend test p<0.001).

- The developer notes that the results of the Poisson regression and trend test suggest that lower fistula use is associated with higher risk of mortality and hospitalization (measured by the respective standardized mortality, standardized hospitalization, and first year standardized mortality ratios), as well as all-cause and vascular access infection related hospitalization (measured by the hospitalization rates), as compared to facilities with higher standardized fistula rates.
- SMP comments on validity:
    - One SMP member questioned the risk model noting that the C-statistics and calibration were based on development data only, with no external validation

Measure #3696 Standardized Modality Switch Ratio for Incident Dialysis Patients (SMoSR)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The standardized modality switch ratio (SMoSR) is defined to be the ratio of the number of observed modality switches (from in-center to home dialysis—peritoneal or home hemodialysis) that occur for adult incident ESRD dialysis patients treated at a particular facility, to the number of modality switches (from in-center to home dialysis—peritoneal or home hemodialysis) that would be expected given the characteristics of the dialysis facility's patients and the national norm for dialysis facilities. The measure includes only the first durable switch that is defined as lasting 30 continuous days or longer. The SMoSR estimates the relative switch rate (from in-center to home dialysis) for a facility, as compared to the national switch rate. Qualitatively, the degree to which the facility's SMoSR varies from 1.00 is the degree to which it exceeds (> 1.00) or is below (< 1.00) the national modality switch rates for patients with the same characteristics as those in the facility. Ratios greater than 1.00 indicate better than expected performance while ratios <1.00 indicate worse than expected performance. When used for public reporting, the measure calculation will be restricted to facilities with at least one expected modality switch in the reporting year. This restriction is required to ensure patients cannot be identified due to small cell size.
- **Type of measure:** Outcome
- **Data source:** Claims, Registry Data
- **Level of analysis:** Facility
- **Risk-adjusted:** Statistical risk model with 18 risk factors
- **Sampling allowed**: N/A
- **Ratings for reliability:** 0 high 6 moderate 2 low and 0 insufficient → Measure passes with MODERATE rating

- Reliability testing was conducted at the accountable entity level using the inter-unit reliability (IUR) with a bootstrap approach.
    - This approach utilizes a resampling procedure to estimate the within facility variation that cannot be directly estimated by ANOVA. The developer also calculated a profile inter-unit reliability (PIUR). This approach assesses the measure's ability to consistently flag extreme providers.
    - The developer calculated a IUR value of 0.605 for the measure, which indicates that over 60 percent of the variation in the measure can be attributed to the between-facility differences and less than 40 percent to the within-facility variation. The PIUR is 0.606.
    - The developer notes that this IUR value is moderate and indicates that the measure can reliably detect differences in performance scores across facilities; the PIUR demonstrates a similar ability to flag outliers.
- **Ratings for validity:** 1 high 7 moderate 2 low and 0 insufficient → Measure passes with MODERATE rating
- The developer conducted empirical validity testing at the accountable unit level.
- The developer used multiple approaches to test validity of the measure, including the following:
    1. Spearman's rho Correlations with Quality Outcome Performance Measures, specifically, Standardized Mortality Ratio (SMR), First-Year Standardized Mortality Ratio (FYSMR), Standardized Hospitalization Ratio (SHR), Standardized Waitlist Ratio-Incident Dialysis Patients (SWR), ICH-CAHPS "Providing information to patients", and the percentage of home dialysis patients at the facility
        - This measure is associated with the Standardized Waitlist Ratio-Incident Dialysis Patients (SWR) (Spearman's rho=0.12, p<0.0001), in the developer's expected direction. The developer notes that facilities that do well facilitating education on transplant that results in patient waitlisting within the first year, are also performing well providing effective education on home dialysis that results in switches from in-center to home dialysis within the first year. As the developer hypothesized, all other associations between this measure and SMR, FYSMR, and SHR were very weak
    2. Gamma Tests for Concordance Analysis with Performance Classification
        - The developer found that this measure and Standardized Waitlist Ratio-Incident Dialysis Patients (SWR) had a positive Gamma coefficient 0.29 and was statistically significant (p<0.0001) indicating that facilities that perform significantly better helping patients switch to home dialysis also do significantly better in helping patients in the referral and waitlisting process for transplant.
    3. Association with patient reported outcomes: ICH-CAHPS "Providing information to patients"
        - The developer found that the facilities with a better performance on this measure have a higher ICH-CAHPS score for providing information to patients (Pearson's r = 0.191)
    4. The developer found moderate correlation between the percentage of home dialysis patients and performance on this measure (Pearson's r = 0.398)
    5. Two-Part Semi-continuous Model
        - The developer presents the logistic regression part of the model which asserts that each unit increase in this measure is associated with a 30

percent decrease in odds of observing a facility with zero home-dialysis patients (p-value <0.001).
- The linear regression part of the model which indicates that for facilities with non-zero number of home dialysis patients, the proportion of home dialysis patients is positively associated with the SMoSR (beta coefficient=2.9, p<0.0001)
- The developer states that facilities providing more effective modality switch education have higher SMoSRs.

Appendix B: Additional Information Submitted by Developers for Consideration

Subgroup 1

Measure Number: 1460

Measure Title: Bloodstream Infection in Hemodialysis Outpatients

Measure Developer/Steward: Centers for Disease Control and Prevention

Reliability

- **Issue 1: [Reliability: Specifications. Concern on reporting Denominator data used to calculate patient-months and access type reporting error]**
    - Developer Response 1: As discussed by Reviewers 1 and 3, there could be undercounting the total patients at a facility based on the question "number of patients on first 2 working days of the month." Undercounting is also possible if a different time frame is selected and can only be optimally minimized if a daily patient count is submitted by dialysis facilities for each reporting month. The protocol for counting the number of patients during the first two working days of the month was chosen to obtain an approximate patient census for a relatively stable population in lieu of the workload required to collect daily patient counts and is typically used for measurements in dialysis facilities. In a 2012 study conducted among Tennessee dialysis facilities a strong correlation was observed between monthly total denominator and NHSN denominator proxy (Pearson Correlation Coefficient [PCC]: 0.988, p <0.0001), and between the NHSN denominator proxy and the other proxy methods (PCC: 0.988–0.991). Nguyen D et al. Correlation Between Methods to Calculate Denominators for Dialysis Event Surveillance Using Electronic Health Record Data, Tennessee, 2012.
    - Reviewer 3 raises the likelihood of manual error while reporting the access type of highest risk category. To guard against this concern, the NHSN Dialysis team conducts annual Dialysis Event surveillance protocol training and provides post-training knowledge checks to promote accurate and standardized reporting. Supporting materials are available online and questions can be directly submitted to the NHSN helpdesk. In addition, end-stage renal disease, healthcare quality programs funded by the Center for Medicare & Medicaid Services required dialysis facilities to complete NHSN training and to obtain proof of course completion.

- **Issue 2: Reliability Testing at Patient/Encounter Level. Validity testing results were provided to demonstrate reliability. Reviewers identified concerns about lack of reporting Kappa**

agreement statistic when access types were validated and high level of underreporting of BSI as a concern to validity of the BSI SIR metric. Also, reviewers expressed concern on lack of Signal to Noise analysis or equivalent metric that could assess the impact of under-reporting on the BSI metric.

- Developer Response 2: The validation study regarding access type aimed to evaluate the accuracy of access type reported to NHSN by the participating dialysis facilities, where review of medical records at the dialysis clinics was considered the reference standard for obtaining these data. The methodology used required an assessment of overall agreement as described in the testing form, given that one approach is a considered a reference standard. A Kappa statistic was not calculated and would not be appropriate because the performances of two raters were not being contrasted.
- The CMS validation studies that indicated 14-19% underreporting in 2018-2019 and 34% in 2020 used the targeted sampling method, which was purposefully focused on facilities that were potentially under-reporting. For that reason, the under-reporting in those facilities does not reflect the overall estimate of under-reporting of all dialysis facilities and should not be used to assess the impact of overall BSI SIR metric.
- Validation studies designed and conducted by state health departments to estimate the extent of under-reporting/over-reporting of BSI are not based on a sampling design that accounts for variation due to factors such as facility size, Large Dialysis Organization (LDO) type, and standalone status; hence, the results are not generalizable to all dialysis facilities and should not be used to generate an overall estimate.
- Conducting any signal to noise analyses or equivalent metrics on how these under-reporting estimates would impact the overall BSI SIR estimation would not be appropriate.

Validity

- **Issue 1:** [Results from Validity assessments were used to demonstrate reliability; hence the above stated concerns were repeated under the Validity section and Developer responses are same as state above. In addition, reviewers expressed concerns on Overall Threats to Validity due to lack of inclusion of social risk factors and patient's underlying medical conditions, which might impact BSI measures. Reviewers also identified that lack of knowledge on patient and social risk factors couple of high underreporting does not make this a meaningful metric.]
    - Developer Response 1: NHSN does not recommend adjusting for social risk and racial factors in prediction models used to calculate the SIR metric because it is not appropriate to account for these factors in assessing facility level performance because facilities should be measured on their performance without being given allowance for patients who might be of varied race and ethnicity as all patients deserve the same level of care. In addition, if NHSN were to include such patient-level factors in risk adjustment models (or as part of an approach to risk stratification which is more likely to happen), the burden of collecting such data for all co-morbidities and risk factors manually is excessive because it would have to be collected for all patients in the facility in order to be used in the models. As we explore options for electronic data capture and measurement methods, it may become feasible.

- As mentioned in the reliability section, estimation of underreporting is best available to NHSN from the validation studies conducted by state health departments and CMS QIP program. Validation studies are designed and conducted by state health departments and the CMS QIP program to estimate the extent of under-reporting/over-reporting of BSI and purposefully focuses on facilities that were potentially under-reporting. For that reason, the under-reporting in those facilities does not reflect the overall estimate of under-reporting of all dialysis facilities and should not be used to assess the impact on overall BSI SIR metric. Conducting any signal to noise analyses or equivalent metrics on how these under-reporting estimates would impact the overall BSI SIR estimation would not be logical.

Measure Number: 2820

Measure Title: Pediatric Computed Tomography (CT) Radiation Dose

Measure Developer/Steward: University of California, San Francisco

Reliability

- **Issue 1:** On page 4, under "Overall Rating of Reliability," Reviewer 4 correctly pointed out typos: *I was able to reproduce the same size calculation on page 48 for brain exams (n=25) but not for skull/abdomen/pelvis (n=9). I presume there's a typo in the last line of the equations on page 48 and that 0.3 should be 0.5. After making that change, I get n=25 for brain and n=15 for skull/abdomen/pelvis. After making a slight change to the formula on page 49, I was able to reproduce the sample size calculation for detecting a 2-fold increase in the probability of receiving a dosage above the benchmark with 80% power. I think the calculated sample size (n=23) may be based on a one-sided alpha=0.05 not alpha=0.025. And I think either choice of alpha is acceptable.*
    - Developer Response 1: We wish to correct 3 typos:

    - In the third to last equation of section 2b.05 *should* say **0.5**/(k/sqrt(NM)), *not* **0.3**/(k/sqrt(NM)). The full, corrected equation is:
    - ≈ 1-Pr(Z < $z_{1-0.05}$-0.5/(k/sqrt(N_M)))
    - Related to the bullet-point above, in the first paragraph of section 2b.06, the minimum sample size required to detect a hospital with median dose in skull and abdomen and pelvis more than 0.5 standard deviations above the benchmark should be 15, not 9. This also impacts the sample size reported in Specifications, sp.25: for skull and abdomen and pelvis, the minimum sample size should be 15 (not 11), to ensure both reliability and validity. Table sp-25 should have "15" instead of "9" in the skull and abdomen columns.
    - The minimal sample size required to detect high proportion of exams above the benchmark (23) was estimated using a one-sided hypothesis, not a two-sided hypothesis. That is, the subscript for the "z" in the second to last equation in 2b.05 *should* be 0.05, *not* 0.025. The full, corrected equation is:
    - 0.8 = Pr[Z < $z_{\mathbf{0.05}}$-H(0.25,0.50) * sqrt(N_{M2})]

Validity

- **Issue 1:** On page 7, under "Validity Results," Reviewer 10 noted that we did not look at chance-adjusted (Kappa) agreement, but rather sensitivity and specificity of CT exam classification relative to a gold standard manual chart review.
    - Developer Response 1: It is our understanding that chance-adjusted agreement is required for reliability, but criterion validity (comparison against a gold standard) is preferred for validity testing.

- **Issue 2:** On page 10, under "Overall Rating of Validity," Reviewer 2 expressed concern that our binary assessment of poor or acceptable performance based on the median or proportion of exams exceeding the 75th percentile could well miss "very bad sites" with a high proportion of exams (up to 49%) with excessive radiation doses, so long as their median per strata, or the majority of their exams, fall below the 75th percentile per strata. The measure also does not differentiate between high radiation doses (meaning above the threshold) and "near lethal doses."
    - Developer Response 2: This is a valid and important concern, and the reviewer is correct that not all radiation doses above the 75th percentile are equally bad. The scenario that the reviewer is describing, however, is not typical. It is true that some sites may have a small number of imaging protocols that drive excessively high radiation dose in specific exam types. However, the measure is intended to assess routine practice and not to identify extreme outliers, as there are already safety mechanisms in place that focus on extreme variation or sentinel events. Each state in the U.S. has a radiation control agency responsible for regulating radiation-producing equipment, and these bodies set protocols for reporting extreme radiation doses. Facilities should already be monitoring and reporting "near lethal doses" under these regulations. The purpose of this measure is to evaluate routine facility-level performance and to encourage corrective action for sites where the median radiation doses are higher than the population 75th percentile.

    - The other issue raised by this reviewer is that sites with a high proportion of exams with excessive radiation doses can still "pass" the measure. We acknowledge there is a critical difference between 0% and 49% of exams dosed above the 75th percentile and that our binary assessment does not account for this quality problem. However, our binary categorization (poor and acceptable) is highly reliable and will drive quality over time. We explored more granular levels of optimality (e.g., very poor, poor, acceptable, ideal) but these classifications were unreliable due to the sample size of our data. The only sites that did have sufficient sample size to achieve reliability at these different levels were pediatric hospitals; thus, it may be worthwhile to implement more categories for this subset of entities in the future.
- **Issue 3:** On page 11, under "Overall Rating of Validity," Reviewer 4 highlights an important concept that quality measures assessing radiation dose must also account for image quality, to avoid the unintended consequence of deteriorating image quality as a result of radiation dose reduction. They ask the questions: *Are there other existing measures that do this? If not, is there any risk of unintended consequences from measuring one but not the other?*
    - Developer Response 3: This is something that we will discuss at length in our full measure submission. Currently, no existing measure in the NQF inventory of

endorsed measures or in any known quality program addresses image quality because (1) there is a lack of consensus on what constitutes image quality in the radiology community, and (2) measuring image quality would require assessment of CT images, which has been technologically infeasible in the absence of electronic tools. We have developed a related, adult CT quality measure (submitted for NQF endorsement in the Fall 2021 cycle) that incorporates both radiation dose and image quality into an electronic clinical quality measure (eCQM). If the adult measure is approved and adopted we intend to develop a similar measure specifically for the pediatric population, and when the current measure #2820 is due for the next round of maintenance review, we aim to submit it as an eCQM that assesses both radiation dose and image quality.

Measure Number: 0471e

Measure Title: ePC-02 Cesarean Birth

Measure Developer/Steward: The Joint Commission

Reliability

- **Issue 1:** Reliability Testing Methods (Question 2a10). Reviewers' comments were mixed; some note data element validity testing satisfies NQF requirements for reliability testing (Reviewer 6), while others state the data reported as data element validity testing is reliability testing. (Reviewer 11). Some note that no accountable entity reliability testing was performed (Reviewers 3, 7, 10).
    - Developer Response 1: NQF submission instructions state if accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required. We elected to follow this guidance and referred reviewers to the validity section. We used NQF guidance for eCQM testing for data element validity and are providing a correlation of the eCQM 0471e Cesarean Birth with the chart- abstracted 0471 Cesarean Birth measure for further concurrent validity. We did not conduct accountable entity level reliability; however, per NQF Guidance for Evaluating Validity, it is not required. We followed NQF approach for eCQM testing, which is to assess agreement between the EHR and chart-abstracted data for validity. We have provided the additional chart (including kappa scores) below for clarity.

- Table 1. Data Element Agreement Rates

Data Element Name	Site #1 Match Rate	Site #1 N	Site #1 Mismatch	Site #1 Due to missing	Site #2 Match Rate	Site #2 N	Site #2 Mismatch	Site #2 Due to missing	Overall Total Match Rate	Overall Total N	Overall Total Mismatch	Overall Total Due to missing	kappa
DOB	100.00%	89	*	*	100.00 %	34	*	*	100.00%	123	*	*	1
ONC Administrative Sex Code	100.00%	89	*	*	100.00 %	34	*	*	100.00%	123	*	*	1
Race	98.90%	89	1	0	100.00 %	34	*	*	99.20%	123	1	0	0.99
Ethnicity	98.90%	89	1	0	100.00 %	34	*	*	99.20%	123	1	0	0.99
Payer	100.00%	89			100.00 %	34	*	*	100.00%	123	*	*	1
Encounter, Performed : Encounter Inpatient	23.60%	89	68	0	100.00 %	34	*	*	44.70%	123	68	0	0.45
Admission Date Time (Relevant Period Start Time)	100.00%	89	*	*	100.00 %	34	*	*	100.00%	123	*	*	1
Discharge Date Time (Relevant Period End Time)	100.00%	89	*	*	100.00 %	34	*	*	100.00%	123	*	*	1
Abnormal Presentation Diagnosis Code	93.30%	89	6	0	100.00 %	34	*	*	95.10%	123	6	0	0.95
Delivery of Singleton Diagnosis Code (ICD10)	100.00%	89	*	*	100.00 %	34	*	*	100.00%	123	*	*	1
Delivery of Singleton Diagnosis (SNOMED)	100.00%	89	*	*		0	*	*	100.00%	89	*	*	1
Placenta Previa Diagnosis Code	100.00%	89	*	*	100.00 %	34	*	*	100.00%	123	*	*	1
Cesarean Section Procedure Code	100.00%	39	*	*	100.00 %	12	*	*	100.00%	51	*	*	1
Cesarean Section Procedure Date	100.00%	39	*	*	100.00 %	12	*	*	100.00%	51	*	*	1
Delivery Procedure Code	100.00%	89	*	*	100.00 %	34	*	*	100.00%	123	*	*	1
Delivery Procedure Date	100.00%	89	*	*	97.10%	34	1	0	99.20%	123	1	0	0.99
Assessment, Performed: Date and time of obstetric delivery, Author Date Time	100.00%	89	*	*	41.20%	34	20	15	83.70%	123	20	15	0.84
Assessment, Performed: Estimated Gestational Age at Delivery, Author Date Time	100.00%	89	*	*	8.80%	34	31	6	74.80%	123	31	6	0.75
Assessment, Performed: Estimated Gestational Age at Delivery, result	100.00%	89	*	*	79.40%	34	7	6	94.30%	123	7	6	0.94
Assessment, Performed: Births.preterm - Author Date Time	100.00%	89	*	*	0.00%	9	9	9	90.80%	98	9	9	0.91
Assessment, Performed: Births.preterm – Result	100.00%	89	*	*	0.00%	12	12	12	88.10%	101	12	12	0.88
Assessment, Performed: Births.term - Author Date Time	100.00%	89	*	*	0.00%	15	15	15	85.60%	104	15	15	0.86
Assessment, Performed: Births.term – Result	96.60%	89	3	0	0.00%	17	17	17	81.10%	106	20	17	0.81
Assessment, Performed: Parity - Author Date Time	100.00%	89	*	*	44.10%	34	19	3	84.60%	123	19	3	0.85
Assess Perf Parity - Result	98.90%	89	1	0	73.50%	34	9	3	91.90%	123	10	3	0.92
Assessment, Performed: pregnancies (gravida) - Author Date Time	100.00%	89	*	*	50.00%	34	17	3	86.20%	123	17	3	0.86

Data Element Name	Site #1 Match Rate	Site #1 N	Site #1 Mismatch	Site #1 Due to missing	Site #2 Match Rate	Site #2 N	Site #2 Mismatch	Site #2 Due to missing	Overall Total Match Rate	Overall Total N	Overall Total Mismatch	Overall Total Due to missing	kappa
Assessment, Performed: pregnancies (gravida) - Result	100.00%	89	*	*	91.20%	34	3	3	97.60%	123	3	3	0.98
TOTALS	96.50%	2303	80	0	78.90%	757	160	92	92.20%	3060	240	92	*

*Cell intentionally left blank

- **Issue 2:** Overall Rating of Reliability: Reviewers noted that data element validity results indicate that variations in measure score may relate to variations in data element validity, due to poor accuracy at the stand-alone hospital (pilot site 2). Reviewer 4 states, "This illustrates the potential for errors in EHR- based data capture and suggests that accuracy results may be heavily site-specific. If issues can be tested and resolved for each site that participates, then it's reasonable to assume the elements will be captured accurately."
    - Developer Response 2: In late 2018, it was decided that in 2020 we would implement ePC02 as an optional eCQM for our accreditation program and pursue NQF endorsement in the future using production data. For this current NQF submission, we evaluated production data received in 2021 representing 2020 discharges.
    - A total of 6 sites consisting of 15 hospitals submitted production data for calendar year 2020. In August of 2021 (in the middle of the pandemic), TJC reached out to all 15 hospitals to recruit sites willing to participate in validity testing on the data submitted. Under normal circumstances, it is incredibly difficult to field test eCQMs. Hospitals generally do not have the substantial resources required to implement eCQMs when they are not part of a regulatory quality program. Considering the additional burden placed on hospitals by the COVID pandemic, TJC was extremely grateful to have 2 sites consisting of 7 hospitals volunteer to participate in pilot testing.
    - Site 2 used a standalone OB documentation system that did not interface completely with the electronic health record (Meditech). The OB documentation was present in Meditech in non-discrete fields in a .pdf format. Most mismatches were in the delivery date/time, estimated gestational age, gravida, para, preterm or term birth fields. The site has since implemented changes where the data is now stored in discrete fields and therefore the data is able to be captured by the eCQM; however, the site was unable to submit updated data in time for NQF submission. Lessons learned during pilot testing will help us prepare implementation support materials to address challenges like those observed at Site 2.
    - We developed the eCQM version of PC02 to reduce administrative burden for sites able to report it. We accept either eCQM or chart-abstracted data (or both) for The Joint Commission accreditation program. Thirteen Joint Commission accredited hospitals submitted PC-02 data for both the eCQM and chart-abstracted measures in calendar year 2020. The ePC-02 rates for the 13 hospitals who submitted both eCQM and chart-abstracted measure results to The Joint Commission for 2020 discharges were correlated. All of these correlations are in the expected direction. A correlation of 0.1 - 0.3 was considered weak, 0.3 - 0.5 was considered moderate, and over 0.5 was considered strong. ePC-02 and the chart-based NQF endorsed PC-02 measure correlate at 0.88, which is strong and is statistically significant ($p<0.01$).
    - Table 2. Correlation Results

Measure	Correlation
Cesarean Section (chart-abstracted)	0.88

Validity

- **Issue 1:** Methods for establishing validity (Question 2b.02). A reviewer notes that while we correlated results with PC-06, Newborn complications, we did not perform direct face or empirical validity testing.
    - Developer Response 1: We have provided correlation to another quality measure, chart-abstracted ePC-02 Cesarean Birth (NQF #0471). Please see developer response to Issue 2 "overall rating of reliability" for details.

- **Issue 2:** Assess the results for establishing validity (Questions 2b.03-04). Most reviewers commented while overall kappa scores indicate excellent agreement, sensitivity of the numerator was a function of testing site. 6 hospitals from Site 1 performed well, while 1 hospital (site 2) was unable to collect data elements necessary to calculate the measure numerator.
    - Developer Response 2: Please see developer response to Issue 2 "overall rating of reliability" for details regarding resolving issues with data collection for Site 2.

- **Issue 3:** Assessment of threats to validity (Questions 2b.15-18). One reviewer stated, "they said because the number of sites was small, no formal statistical test was performed for the effect of exclusion on the performance scores. This is concerning. It is not clear where there are large numbers of exclusions whether patients with social risk factors were affected more by exclusions and whether this would inappropriately inflate facility performance."
    - Developer Response 3: The Joint Commission's eCQM Cesarean Birth measure ePC-02 has been used in accreditation programs since 2020. A 2010 study by Huesch and Doctor examined the relationship between African American race and cesarean delivery. They found that the prevalence of the malpresentation risk factor for elective primary cesareans was less for African Americans (21.5%) compared to Other Race/Ethnicity (32.9%) (P<.001).1 Because there have been conflicting study results on race and placenta previa, a study by Kim et al. (2011) was done to examine the prevalence of placenta previa among five major racial and ethnic groups: African American, Asian, Caucasian, Hispanic and Native American. Results showed Asian women, followed by African American women, have a significantly increased risk of pregnancy complicated by placenta previa, compared with Caucasian women: Asian0.64%, Native American 0.60%, African American 0.44%, Caucasian 0.36%, Hispanic 0.34% and unknown 0.31% (P<0.001).2 While this shows an increased risk, the measure targets the nulliparous population which has a lower overall rate of placenta previa. According to a 2015 study by Ahmed et al., placenta previa complicates approximately 0.3–0.5% of pregnancies with no prior cesarean delivery.3 Prior to ePC-02's use, a pilot study was completed and included statistical testing for the measure rates across patients of different races and insurance payers (see study results below). Based on the information from literature and the results of the pilot study, it is reasonable to determine that the exclusions for malpresentation and placenta previa would not inappropriately inflate facility performance. There are still statistically significant differences in cesarean birth rates among race when exclusions have been applied.
    - The results of the pilot testing are as follows (see table 3):

- The tests rejected the hypotheses that the measure rate is the same for patients of different races, and insurance payer. This indicates that the measure rate is statistically different across these groups. But the test failed to reject this hypothesis for patients of different ethnicities, likely because so few patients were classified as Hispanic in the electronic extract. We identified disparities in performance in terms of patients' race, and insurance status.
- We see differences in performance by race. Compared with White patients, Black patients had higher rates at seven out of ten hospitals. Across hospitals, the performance rate of Black patients was 29.7 percent, while White patients had a performance rate of 25.1 percent. We also see differences in measure performance by insurance payer type. Medicaid patients had lower rates compared with patients not covered by Medicare or Medicaid (that is, patients with private insurance and uninsured patients) at eight out of ten hospitals. The performance rate of Medicaid patients was 23.8 percent across hospitals, while patients who reported having "other" types of insurance (i.e., patients with private insurance and uninsured patients) had a rate of 26.8 percent. Although the number of patients covered by Medicare was small, this population had higher rates (34.5 percent) than both the Medicaid and "other insurance" populations. These differences between patient groups in the overall sample were statistically significant at the .05 level.
- Chi-square tests performed on the aggregated data (the combined data for all hospitals in the sample) rejected the hypotheses that the measure rate is the same for patients of different races, and insurance payer. These differences between patient groups in the overall sample were statistically significant at the .05 level. The results demonstrated that statistically significant differences can be detected among hospitals and demographic characteristics (race and insurance payer). The gaps in performance between hospitals and demographic groups indicate that there is room for improvement in performance rates.

1. Huesch, M., & Doctor, J. N. (2015). Factors associated with increased cesarean risk among African American women: evidence from California, 2010. *American journal of public health*, *105*(5), 956–962. https://doi.org/10.2105/AJPH.2014.302381
2. Kim, L. H., Caughey, A. B., Laguardia, J. C., & Escobar, G. J. (2012). Racial and ethnic differences in the prevalence of placenta previa. *Journal of perinatology : official journal of the California Perinatal Association*, *32*(4), 260–264. https://doi.org/10.1038/jp.2011.86
3. Ahmed, S. R., Aitallah, A., Abdelghafar, H. M., & Alsammani, M. A. (2015). Major Placenta Previa: Rate, Maternal and Neonatal Outcomes Experience at a Tertiary Maternity Hospital, Sohag, Egypt: A Prospective Study. *Journal of clinical and diagnostic research: JCDR*, *9*(11), QC17–QC19. https://doi.org/10.7860/JCDR/2014/14930.6831

Table 3. Performance rate by disparity group (hospital level)

Category	Measure Result (%) Hospital 1	Measure Result (%) Hospital 2	Measure Result (%) Hospital 3	Measure Result (%) Hospital 4	Measure Result (%) Hospital 5	Measure Result (%) Hospital 6	Measure Result (%) Hospital 7	Measure Result (%) Hospital 8	Measure Result (%) Hospital 9	Measure Result (%) Hospital 10	Measure Result (%) Across Hospitals (Pooled Data)
Race	%	%	%	%	%	%	%	%	%	%	Race
White	27.4	31	21.9	24.5	35.8	27.8	24	20	27.6	21.5	25.10%
Black	23.3	32	32.4	36.7	35.6	34.2	29	24.1	27.6	27.1	29.70%
Other	20.3	28.9	21.1	27.6	31.6	25.2	25.9	19.7	29.4	18.3	24.50%
(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer	(Primary) Payer
Medicare	0	42.9	0.0*	0	NaN*	66.7*	33.3	75.0*	0	25	34.50%
Medicaid	18.6	27.8	23.6	26.1	34.1	24.7	25.3	19.6	23.7	22.4	23.80%
Others*	28.8	32.5	22.3	27	35	30.5	26	24.1	28.9	22.2	26.80%
Overall performance	**22.1**	**30**	**22.7**	**26.5**	**34.3**	**27.6**	**25.6**	**20.8**	**27.7**	**22.1**	*

Source: Data from January 1, 2014-December 31, 2015

* Indicates a sample size of fewer than 10 patients

Notes: Table does not include patients with missing or unknown characteristics data.

NaN: Not calculable because the denominator in the equation is equal to zero.

Medicare includes Original Medicare, Medicare Managed Care, and a combination of Medicare and another private payer (e.g., Medicare and private insurance).

Medicaid includes traditional Medicaid, Medicaid Managed care, or a combination of Medicaid and another payer (e.g., Medicaid and private insurance).

Other payers includes private insurance (e.g., health maintenance organization, preferred provider organization), the uninsured, and other payers (e.g., worker's compensation).

- **Issue 4:** Risk Adjustment (Questions 2b19-32). Reviewer responses were mixed; some supported the decision not to risk adjust the outcome (Reviewer 2, 3, 5, 9) while others suggested further exploration of risk factors was needed or deferred to the standing committee for clinical review (Reviewer 1, 6, 7, 8).
    - Developer Response 4: The chart-abstracted PC-02 Cesarean Birth Measure was released in 2010 and the eCQM ePC-02 in 2020. The issue of risk adjustment has been discussed with Technical Advisory Panels over the years. The Joint Commission's Perinatal Care Technical Advisory Panel recommends using the simple cesarean birth rates without further risk adjustment. In 2016, the decision to remove all risk-adjustment from this measure was made based on analysis of data on this measure received by The Joint Commission which indicates that age is only a weak predictor of outcome, and that age standardization could potentially distort the age-standardized measure rates for hospitals with small sample sizes. Additionally, the California Maternal Quality Care Collaborative (CMQCC) has done extensive research (provided in submission) which supports the rationale for no additional risk adjustment.
    - The Cesarean Birth measure is designed to measure the rates of cesarean births among a subset of the general obstetric population of women while also keeping the burden of data collection to a minimum. The measure focuses on mothers having their first birth who are at the highest risk of primary cesarean birth when

compared to mothers who have experienced a previous vaginal birth. By setting aside twins, breech presentations, and premature births, this measure focuses on a more homogeneous group of women where the greatest improvement opportunity exists. Other typical cesarean birth measures report rates of either primary cesarean (a mix of first and subsequent births and therefore dependent on the proportion of first births in the facilities' population) or repeat cesarean (identifies a different set of issues focused on emergency support capabilities). Because the measure focuses on nulliparous women with a term, singleton baby in a vertex position, the only exclusions to the denominator population are diagnosis codes in the value sets Abnormal Presentation. (2.16.840.1.113762.1.4.1045.105) and Placenta Previa (2.16.840.1.113762.1.4.1110.37). The eCQM measure logic only allows nulliparous women with a term singleton delivery into the measure population. Extensive testing by The Joint Commission made it clear that there is no need to exclude for all known indications for performing cesareans, since these types of medical conditions are less common and would not significantly increase a hospital's cesarean rates. Maternal age, race, and weight are known cesarean risk factors for individuals but commonly cancel each other when analyzing for hospital PC-02 rates. Thus, including a comprehensive set of maternal medical exclusions would add data collection burdens without commensurate benefit.

    - The measure is designed to be an accurate way for leaders to identify whether a hospital's rate of cesarean births for women included in this select population is consistent with the rate of cesareans within this same population at another hospital. Hospitals whose Cesarean Birth measure rates are higher than rates at other hospitals are encouraged to explore and evaluate differences in the medical and nursing management of women in labor.

- **Issue 5:** Missing Data (Questions 2b08-10). Reviewers noted that the high rate of missingness at site 2 raised questions about why the site had higher rates of missing data, whether the challenges at this site could be generalized to other sites, and how challenges might be addressed
    - Developer Response 5: We developed the eCQM version of PC02 to reduce administrative burden for sites able to report it but accept both eCQM and chart-abstracted data in The Joint Commission accreditation program. The eCQM has been correlated with the chart-based version of PC-02 as shown above in Developer Response 2 under Issue 2 Overall Rating of Reliability. Hospitals are encouraged to adopt eCQM measures when their hospitals can support the technology. Hospitals are in varying stages of EHR capability; however, we have learned from site 2 that although the system had some interoperability issues, the hospital implemented changes and will be able to submit valid data in the future. Please see Developer Response 2 under Issue 2 Overall Rating of Reliability for more details.

- **Issue 6:** Overall Rating of Validity. Lacks kappa agreement for exclusions. No testing of exclusions.
    - Developer Response 6: Please refer to Developer Response 2 for Issue 2 Overall Rating of Reliability and address above. Table 2b.17.01 in the Intent to Submit submission provides the exclusion data for the value sets Abnormal Presentation and Placenta Previa. We have provided an additional breakdown of the exclusion information in Table

Table 4 Denominator Exclusion Data

Hospital Number	Delivery Encounters	Denominator: Denominator Exclusions (Total Placenta Previa & Abnormal Presentation) (N)	Denominator: Abnormal Presentation (N)	Denominator: Placenta Previa (N)	Denominator: Denominator Exclusions Placenta Previa & Abnormal Presentation) (%)	Denominator: Denominator less Exclusions (Placenta Previa & Abnormal Presentation)	Rate With Placenta Previa & Abnormal Presentation Cases Excluded	Rate with Placenta Previa & Abnormal Presentation Cases Included
1.1	35	1	1	0	2.9%	34	32.4%	34.3%
1.2	13	0	0	0	0.0%	13	30.8%	30.8%
1.3	86	7	7	0	8.1%	79	22.8%	29.1%
1.4	12	0	0	0	0.0%	12	16.7%	16.7%
1.5	38	6	6	0	15.8%	32	21.9%	34.2%
1.6	32	5	5	0	15.6%	27	18.5%	31.3%
2	11	0	0	0	0.0%	11	0.0%	0.0%
3.1	71	0	0	0	0.0%	71	71.8%	71.8%
3.2	38	0	0	0	0.0%	38	55.3%	55.3%
3.3	9	0	0	0	0.0%	9	55.6%	55.6%
4	2	0	0	0	0.0%	2	0.0%	0.0%
5	122	10	10	0	8.2%	112	25.0%	31.1%
6.1	399	26	21	5	6.5%	373	20.6%	25.8%
6.2	41	4	3	1	9.8%	37	18.9%	26.8%
6.3	88	5	5	0	5.7%	83	25.3%	29.5%
Total	997	64	58	6	6.4%	933	27.5%	32.2%

Measure Number: 0716e

Measure Title: ePC-06 Unexpected Newborn Complications in Term Newborns

Measure Developer/Steward: The Joint Commission

Reliability

- **Issue 1:** Reliability Specifications: Reviewer 9: I have a question about the exclusion for babies exposed to maternal drug use in-utero - Does this include illicit substances as well as prescription teratogenic medications as well?
    - **Developer Response 1:** The Maternal Drug Use value set includes the ICD-10CM and SNOMEDCT codes used to exclude newborns affected by maternal drug use. The value set can be found on the VSAC with the OID 2.16.840.1.113762.1.4.1029.127. Both illicit substances and prescription teratogenic medications are included in the value set. Here is a list of the code descriptions for the included codes:
    - Fetal exposure to teratogenic substance (disorder)
    - Fetal or neonatal effect of hallucinogenic agent transmitted via placenta and/or breast milk (disorder)
    - Fetal or neonatal effect of noxious substance transmitted via placenta (disorder)
    - Fetal valproate syndrome (disorder)

- Fetal alcohol syndrome (disorder)
- Fetal or neonatal effect of maternal use of alcohol (disorder)
- Fetal or neonatal effect of placental or breast transfer of narcotics (disorder)
- Fetal or neonatal effect of placental or breast transfer of hallucinogen (disorder)
- Fetal or neonatal effect of placental or breast transfer of immune sera (disorder)
- Fetal or neonatal effect of placental or breast transfer of anticonvulsant (disorder)
- Fetal or neonatal effect of placental or breast transfer of anticoagulant (disorder)
- Fetal or neonatal effect of placental or breast transfer of chemotherapeutic agent (disorder)
- Fetal or neonatal effect of placental or breast transfer of uterine depressant (disorder)
- Fetal or neonatal effect of placental or breast transfer of hypoglycemic agent (disorder)
- Fetal or neonatal effect of placental or breast transfer of endocrine agent (disorder)
- Fetal or neonatal effect of noxious influences transmitted via placenta or breast milk (disorder)
- Fetal or neonatal effect of maternal use of tobacco (disorder)
- Fetal or neonatal effect of maternal use of nutritional chemical substance (disorder)
- Fetal or neonatal effect of maternal exposure to environmental chemical substances (disorder)
- Neonatal withdrawal symptoms from maternal use of drugs of addiction (finding)
- Withdrawal symptoms from therapeutic use of drugs in newborn (finding)
- Neonatal effect of noxious substance transmitted via breast milk (disorder)
- Fetal minoxidil syndrome (disorder)
- Fetal primidone syndrome (disorder)
- Fetal or neonatal effect of placental or breast transfer of anti-infective (disorder)
- Fetal or neonatal effect of antibiotic transmitted via placenta and/or breast milk (disorder)
- Fetal or neonatal effect of immune serum transmitted via placenta and/or breast milk (disorder)
- Fetal or neonatal effect of maternal use of antihypertensive drug (disorder)
- Fetal disorder caused by chemicals (disorder)
- Fetal warfarin syndrome (disorder)
- Fetal or neonatal effect of diethylstilbestrol transmitted via placenta and/or breast milk (disorder)
- Fetal or neonatal effect of maternal transmission of substance (disorder)
- Fetus affected by placental transfer of anticonvulsant (disorder)
- Fetal Alcohol Spectrum Disorder (disorder)
- Fetal or neonatal effect of maternal alcohol addiction (disorder)
- Drug withdrawal syndrome in neonate of dependent mother (disorder)
- Fetal or neonatal effect of anti-infective agent transmitted via placenta and/or breast milk (disorder)
- Neonatal effect of alcohol transmitted via breast milk (disorder)
- Fetal or neonatal effect of anti-infective agent transmitted via placenta (disorder)
- Neonatal effect of anti-infective agent transmitted via breast milk (disorder)
- Fetal or neonatal effect of hallucinogenic agent transmitted via placenta (disorder)

  - Fetal or neonatal effect of hallucinogenic agent transmitted via breast milk (disorder)
  - Fetal or neonatal effect of diethylstilbestrol transmitted via placenta (disorder)
  - Fetal or neonatal effect of diethylstilbestrol transmitted via breast milk (disorder)
  - Fetal or neonatal effect of narcotic transmitted via placenta (disorder)
  - Fetal or neonatal effect of narcotic transmitted via breast milk (disorder)
  - Fetal or neonatal effect of medicinal agent transmitted via placenta and/or breast milk (disorder)
  - Withdrawal symptom (finding)
  - Fetal or neonatal effect of narcotic transmitted via placenta and/or breast milk (disorder)
  - Fetal or neonatal effect of toxic substance transmitted via placenta and/or breast milk (disorder)
  - Newborn affected by maternal antineoplastic chemotherapy
  - Newborn affected by maternal cytotoxic drugs
  - Newborn affected by maternal use of anticonvulsants
  - Newborn affected by maternal use of opiates
  - Newborn affected by maternal use of antidepressants
  - Newborn affected by maternal use of amphetamines
  - Newborn affected by maternal use of sedative-hypnotics
  - Newborn affected by other maternal medication
  - Newborn affected by maternal use of unspecified medication
  - Newborn affected by maternal use of anxiolytics
  - Newborn affected by maternal use of alcohol
  - Newborn affected by maternal use of unspecified drugs of addiction
  - Newborn affected by maternal use of cocaine
  - Newborn affected by maternal use of hallucinogens
  - Newborn affected by maternal use of other drugs of addiction
  - Newborn affected by maternal use of cannabis
  - Newborn affected by other maternal noxious substances
  - Neonatal withdrawal symptoms from maternal use of drugs of addiction
  - Withdrawal symptoms from therapeutic use of drugs in newborn

- **Issue 2:** Reliability Specifications: Reviewer 11 stated that the Length of stay calculations were not exactly clear; LOS > maternal vs. LOS >5 days for uncomplicated birth
  - Developer Response 2: The severe complications value set Neonatal Severe Septicemia has a length of stay modifier of greater than 4 days. Cases with a code in this value set and a length of stay greater than 4 days regardless of delivery type will be in the numerator as a severe complication. There are 2 sets of moderate complication value sets. The first set are Moderate Birth Trauma, Moderate Respiratory Complications, and Moderate Respiratory Complications Procedures. These value sets do not use a length of stay modifier. If a code in the newborn record is in one of these value sets, the case would be in the numerator. The second set of moderate complication value sets includes length of stay modifiers. These value sets are Moderate Birth Trauma with LOS, Moderate Respiratory Complications with LOS, Moderate Neurological Complications with LOS Procedures, Moderate Respiratory Complications with LOS Procedures, and Moderate Infection with LOS. The length of stay (LOS) calculation is used to determine if

the case meets the LOS modifier requirements. A length of stay greater than 4 days for a cesarean delivery and a length of stay greater than 2 days for a vaginal delivery would qualify the logic to look in these value sets. If a code is present in any of the moderate complication with LOS value sets and the case meets the LOS calculation requirements the case would be in the numerator. Per the original chart-based developer CMQCC, the length of stay calculation of greater than 5 days is used as "safety net" to capture cases that are under-coded, however cases which have a LOS greater than 5 days and codes in value sets Neonatal Jaundice, Phototherapy, or Social Indications would be excluded from the numerator if they have no other severe or moderate complication codes**.**

- **Issue 3:** Reliability Testing: There may be a lack of consensus among reviewers about reliability versus validity testing. Use of data element level testing was described as both reliability and validity by reviewers. Several reviewers stated that Entity level testing was not presented. Reviewer 10 stated there was no exclusion data element testing presented.
- Developer Response 3: As accepted by the NQF guidance, we chose to submit data element validity testing, which would make section 2a.10-2a.12 not applicable. Reviewer 10 felt the results were incomplete, although recognized the process we were following. Please see Issue 4 under validity for additional entity level testing and Issue 2 under Validity for additional data element testing information.

- **Issue 4: Reviewer 8 and 11 noted highly variable results.**
  - Developer Response 4: Funnel plot methodology showed that while there was variability in the rates only 2 pilot sites had statistically meaningful differences in rates. The purpose of showing the funnel plot was to demonstrate that there is variability in the measure rates and that this variability could potentially be used to determine meaningful organizational differences in measure rates after adjusting for inter-hospital variability. As noted in the reference given for construction of the funnel plot in our submission, the confidence limits were adjusted for inter-hospital variability, although this variability was not explicitly reported separately.

Validity

- **Issue1:** Reviewer 10 stated that data element validity is acceptable, however not all data elements: were tested.
  - Developer Response 2: An additional level of data is provided here from our pilot study. The data element agreement rate analysis by value set indicates that agreement at the data element level is generally excellent. The only data element, used for measure outcome calculation, under 0.60 was Neonatal Severe Respiratory Procedures where the original data failed to identify this in two records. As noted in 2b.04 of our submission, for Site 2 the Severe Shock and Resuscitation procedures codes were saved as pdf format. Site 2 agreed that an EHR future improvement would be able to make this procedure code available in a structured format.

Table 1 Pilot Site Agreement Rates for Data Elements by Value Set

Value Set Name	Used for	Total	Match	kappa	PPV	NPV	Incidence
Ethnicity	Demographics	61	49	0.63	1.00	0.93	61
ONC Administrative Sex	Demographics	61	61	1.00	1.00	1.00	61
Payer	Demographics	61	61	1.00	1.00	1.00	61
Race	Demographics	61	61	1.00	1.00	1.00	61
Birth Weight	Denominator	61	61	1.00	1.00	1.00	61
Encounter Inpatient	Denominator	61	61	1.00	1.00	1.00	61
Single Live Birth	Denominator	61	61	1.00	1.00	1.00	61
Single Live born Newborn	Denominator	61	61	1.00	1.00	1.00	61
Single Live Born Newborn Born in Hospital	Denominator	61	61	1.00	1.00	1.00	61
Congenital Malformations	Denominator Exclusion	61	60	0.79	1.00	0.67	7
Discharge To Acute Care Facility	Numerator	61	61	*	*	1.00	0
Discharged to Health Care Facility for Hospice Care	Numerator	61	61	*	*	1.00	0
Fetal Conditions	Denominator Exclusion	61	61	1.00	1.00	1.00	1
Maternal Drug Use	Denominator Exclusion	61	61	1.00	1.00	1.00	2
Other Health Care Facility	Numerator	61	61	*	*	1.00	0
Moderate Birth Trauma	Numerator	61	61	*	*	1.00	0
Moderate Birth Trauma with LOS	Numerator	61	61	1.00	1.00	1.00	4
Moderate Infection with LOS	Numerator	61	61	1.00	1.00	1.00	2
Moderate Neurological Complications with LOS Procedures	Numerator	61	61	*	*	1.00	0
Moderate Respiratory Complications	Numerator	61	60	0.95	1.00	0.93	15
Moderate Respiratory Complications Procedures	Numerator	61	61	1.00	1.00	1.00	2
Moderate Respiratory Complications with LOS	Numerator	61	61	1.00	1.00	1.00	15
Moderate Respiratory complications with LOS Procedures	Numerator	61	61	1.00	1.00	1.00	8

Value Set Name	Used for	Total	Match	kappa	PPV	NPV	Incidence
Neonatal Jaundice	Numerator	61	61	1.00	1.00	1.00	8
Neonatal Severe Infection	Numerator	61	61	*	*	1.00	0
Neonatal Severe Neurological Complications	Numerator	61	61	*	*	1.00	0
Neonatal Severe Neurological Procedures	Numerator	61	61	*	*	1.00	0
Neonatal Severe Respiratory Complications	Numerator	61	61	*	*	1.00	0
Neonatal Severe Respiratory Procedures	Numerator	61	59	0.49	1.00	0.33	3
Neonatal Severe Septicemia	Numerator	61	61	1.00	1.00	1.00	2
Patient Expired	Numerator	61	61	*	*	1.00	0
Phototherapy	Numerator	61	61	1.00	1.00	1.00	3
Severe Birth Trauma	Numerator	61	61	1.00	1.00	1.00	6
Severe Hypoxia/Asphyxia	Numerator	61	61	1.00	1.00	1.00	1
Severe Shock and Resuscitation	Numerator	61	61	*	*	1.00	0
Severe Shock and Resuscitation Procedures	Numerator	61	61	1.00	1.00	1.00	1
Single Liveborn Newborn Cesarean	Numerator	61	61	1.00	1.00	1.00	19
Single Liveborn Newborn Vaginal	Numerator	61	61	1.00	1.00	1.00	9
Social Indications	Numerator	61	61	1.00	1.00	1.00	1

- *Cell intentionally left blank

- **Issue 3:** Assess the results(s) for establishing validity: Reviewer 1 comments were related to the 0471e Cesarean Birth Measure not 0716e Unexpected Complications in Term Newborns measure. Reviewer 8 comments were concerns about Funnel Plot results and no Brier score. Reviewer 10 stated data element validity is acceptable, however not all data elements were tested**.**
    - Developer Response 3: We have concerns that our 0716e measure was evaluated with the 0471e measure information or that transcription error occurred. We request clarification of the comment. In response to the concerns for using Funnel Plot methodology instead of the Brier score, ePC-06 is not risk-adjusted and the Brier score is most meaningful when there is a risk model available. The purpose of showing the funnel plot was to demonstrate that there is variability in the measure rates and that this variability could potentially be used to determine meaningful organizational differences in measure rates after adjusting for inter-hospital variability. As noted in the reference given for construction of the funnel plot in our submission, the confidence limits were adjusted for inter-hospital variability, although this variability was not explicitly reported separately. It should not be

interpreted that the funnel plot actually did identify differences since the reporting period is limited and there were a limited number of hospitals in the pilot. Once data are collected on a greater number of hospitals and for longer reporting periods, then the funnel plot methodology could be applied using the methods given by the reference the reviewer mentioned, including checking whether the measure was associated with institutional characteristics and excepting the steps in the reference having to do with risk adjustment since this measure was not risk-adjusted. Additional results for data element validity are provided in Issue 2 under Validity.

- **Issue 4:** VALIDITY: ASSESSMENT OF THREATS TO VALIDITY: Reviewer 11 states that both severe and moderate complications are included in the numerator without any distinction between the 2.
    - Developer Response 4: We have provided a table with the breakdown of severe and moderate complication rates for the pilot sites. The predictive accuracy for the two strata measures was excellent, with all the PPV and NPV statistics greater than 0.80 which is in the excellent to perfect predictive accuracy range. The severe rate should be the focus of the measure and current literature suggests hospital rates of severe unexpected complications in term newborns ranged from 0.6 to 89.9 per 1000 births (median, 15.3 per 1000 births [interquartile range, 9.6-22.0 per 1000 births]). 1 This measure is also an important balancing measure and results should be evaluated with other Perinatal Care measures such as PC-02 Cesarean Birth.

1. Clapp, M. A., James, K. E., Bates, S. V., & Kaimal, A. J. (2020). Patient and Hospital Factors Associated With Unexpected Newborn Complications Among Term Neonates in US Hospitals. JAMA network open, 3(2), e1919498. https://doi.org/10.1001/jamanetworkopen.2019.19498

Table 2 Pilot Site Measure Rates

Hospital	Overall Numerator	Denominator	Overall rate per 1,000	Severe rate per 1,000	Moderate rate per 1,000
1	5	184	27.2	21.7	5.4
2	7	111	63.1	18.0	45.0
3	6	331	18.1	3.0	15.1
4	50	596	83.9	63.8	20.1
5	37	724	51.1	16.6	34.5
6	11	182	60.4	27.5	33.0
7	5	74	67.6	40.5	27.0
8	3	179	16.8	0.0	16.8

Hospital	Overall Numerator	Denominator	Overall rate per 1,000	Severe rate per 1,000	Moderate rate per 1,000
9	4	388	10.3	5.2	5.2
10	2	26	76.9	38.5	38.5
11	6	592	10.1	3.4	6.8
12	8	455	17.6	8.8	8.8
13	2	609	3.3	1.6	1.6
14	1	129	7.8	0.0	7.8
15	2	112	17.9	8.9	8.9
16	6	308	19.5	13.0	6.5
17	6	493	12.2	8.1	4.1
18	2	584	3.4	0.0	3.4

Table 2.1 Measure-level agreement statistics

*	*	**Overall Rate**	**Overall Rate**	**Overall Rate**	**Overall Rate**	**Severe Rate**	**Severe Rate**	**Moderate Rate**	**Moderate Rate**
System	N	Agreement Rate	Kappa	NPV	PPV	NPV	PPV	NPV	PPV
Site 1	32	93.8%	0.91	1.00	0.95	0.96	1.00	1.00	0.86
Site 2	29	100.0%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Total	61	96.7%	0.96	1.00	0.97	0.97	1.00	1.00	0.92

*Cell intentionally left blank

- **Issue 5:** Risk adjustment: Reviewer 1 references "the exclusions result in a homogenous population that represents the group where most improvement opportunity exists." This is a statement from 0471e Cesarean Birth not 0716 Unexpected Complications in Term Newborns. Reviewer 6 stated they were not convinced that including maternal conditions would add burden and that some maternal risk factors are associated with newborn outcomes. Reviewer 7 expressed concern about other abuses such as alcohol be an exclusion and may want to consider other social risk factors. Reviewer 11 stated since the length of stay is included in the numerator certain social risk factors might determine length of stay even in the presence of excellent patient care.

- Developer Response 5: We have concerns that our 0716e measure was evaluated with the 0471e measure information or that transcription error occurred. Adding maternal conditions codes would increase the burden of abstraction because the data elements are taken from the newborn's record for the 0716e measure and maternal codes are not provided in the newborn record, only the maternal record. One of the conditions referenced was Placental conditions which are listed in the Fetal Conditions value set found on the Value Set Authority Center with OID **2.16.840.1.113762.1.4.1029.130.** Having a placental condition code on this table would exclude these cases from the denominator. Other abuse, such as alcohol, would be excluded from the denominator if the case had a code in the Maternal Drug Use value set. Please see issue 1 under reliability for additional information on the Maternal Drug Use value set. Other social risk factors are used to exclude cases whose length of stay is >5 days and no severe or moderate complications codes are found in the record. Currently, payer, race, ethnicity, and sex are collected by measure and could be used for internal hospital use or to be considered for future enhancements to the measure. To guard against social risk factors increasing a newborn's length of stay with out the presence of a severe or moderate complication, the measure logic uses a length of stay >5 days and NO codes in the value sets Neonatal Jaundice (OID **2.16.840.1.113762.1.4.1029.124**), Phototherapy (OID **2.16.840.1.113762.1.4.1029.121)**, and Social Indications (OID **2.16.840.1.113762.1.4.1029.136**). If a code is present on one or all of these value sets and the length of stay is >5 days, AND there is no other severe or moderate complication code, the case would be excluded from the numerator.

- **Issue 6:** Missing Data: Reviewer 1 expressed concern for missing data from one of the pilot sites, reviewer 8 referred to Table 2b.09.01, which is a table from 0471e Cesarean Birth not 0716e Unexpected Complications in Term Newborns. Reviewer 9 and 11 also noted some missing data issues.
    - Developer Response 6: We have concerns that our 0716e measure was evaluated with the 0471e measure information or that transcription error occurred. The 0716e measure has excellent agreement rates and chance-corrected reliability. Site 1 exhibited a match rate of 93.8% and Site 2 exhibited 100% match rate in measure outcome. The overall kappa is 0.955 which indicates excellent agreement. Missing data elements were due to the pilot site not consistently collecting the data according to instructions and different data sources. To resolve missing data issues, the latest version of ePC-06 has been updated with Procedure Start Date only data element without time to resolve the Datetime issue and Assessment Performed Author Datetime was replaced with Relevant Datetime to align with new QDM guidance, where it is used to capture assessment time instead of documentation time. Site 2 agreed that future EHR improvements would resolve other missing data issues.

- **Issue 7: Overall rating of validity:** Reviewer 4 stated "Developers argue that risk-adjustment is not required because the population is relatively homogeneous after the application of exclusion criteria which were selected intentionally in order to obviate the need for adjustment. Subject matter experts are in the best position to assess the potential for residual confounding due to any possible risk factors that were not excluded or adjusted. For example, is it safe to

assume that maternal age isn't important?" Reviewer 10 stated data element validity testing is incomplete and there is a lack of risk adjustment. Reviewer 11 Stated that given the equivalence given in the numerator to severe and moderate complications and absence of validation with any other measure of quality, it is not clear that this measure has been adequately validated.

- Developer Response 7: The Joint Commission works with a Technical Expert Panel and one of our clinical expert consultants is Dr. Elliott Main from the California Maternal Quality Care Collaborative (CMQCC). As a subject matter expert, Dr. Main has provided feedback on risk adjustment for the Unexpected Complications in Term Newborns measure. As this eCQM is developed from the chart-based PC-06 measure (NQF endorsed #0716) which was developed by CMQCC, the risk adjustment rationale is used for the chart-based version of the measure is also applicable here. Risk adjustment is not included for maternal conditions because this would add burden to collecting the measure. Maternal condition codes are located in the maternal record, and this measure uses data from the newborn record as it is the population in the measure. Also, some maternal conditions are complications of labor that affect the baby which is what the measure is trying to assess. 1 CMQCC provided additional rationale in their 2020 submission for why the chart-based PC-06 (NQF #0716) measure is not risk adjusted. This measure is not risk-adjusted but rather risk-stratified, using a series of exclusions to identify a standard low-risk population. When constructing the measure, the exclusion criteria were chosen to ensure that the target population would be healthy, term babies with no pre-existing complications, thus reducing bias due to case mix complications. Babies more at risk for experiencing adverse outcomes (premature babies, low birth weight infants, babies with congenital malformations, exposure to maternal substance use and other pre-existing conditions) were excluded from the target population. The measure is not risk-adjusted for patient factors that could possibly obscure disparities such as sex or insurance status of the newborns. The measure does not adjust for gestational age (within the term, 37-43 weeks of gestation, population) recognizing that some morbidities are more prevalent at different gestational ages because timing of labor induction is part of obstetric practice. In short, we did not want to mask morbidities resulting from early elective delivery practices (under 39 weeks of gestational age) or non-interventional practices in some hospitals (who do not induce women who are over 41 weeks pregnant, thus increasing the risk of stillbirth and morbidity in post term infants). Variables related to quality of care are purposely not included in risk models for performance measures used to assess quality. Risk adjustment should not mask or adjust for the very factors that are driving the differences in neonatal health outcomes at hospitals. The measure does not adjust for a hospital's neonatal intensive care unit level, birth volume, ownership status, teaching status or number of maternal-fetal care specialists. The list of exclusions account for most conditions that have been linked to social risk factors such as preterm birth and poor fetal growth (small-for-dates infants) so we did not further assess social risk. As the eCQM ePC-06 (NQF #0716e) is correlated with the chart-abstracted NQF endorsed measure, the rationale for risk adjustment applies.

Additional information has been provided in Developer Response 2 under Validity on data element validity. Measure rates for severe and moderate complications have been provided in Developer Response 4 under Validity.

- For validation with another measure of quality, the ePC-06 rates for the pilot hospitals were correlated with rates on chart-based PC-06 (0716 NQF) measure reported to The Joint Commission for 2019 discharges. A correlation of 0.1 - 0.3 was considered weak, 0.3 - 0.5 was considered moderate, and over 0.5 was considered strong. The ePC-06 measure has strong correlation with the chart-based PC-06 NQF endorsed measure.
- Table 3 Correlations

Measure	Correlation
Unexpected Complications in Term Newborns - Overall Rate	0.91
Unexpected Complications in Term Newborns - Severe Rate	0.87
Unexpected Complications in Term Newborns - Moderate Rate	0.66

1. California Maternal Quality Care Collaborative (2020, March 26) Overview & Frequently Asked Questions Unexpected Newborn Complications (UNC) Measure PC-06 and NQF #716. Retrieved from https://www.cmqcc.org/focus-areas/quality-metrics/unexpected-complications-term-newborns

Subgroup 2

Measure Number: 3679

Measure Title: Home Dialysis Rate

Measure Developer/Steward: Kidney Care Quality Alliance

Reliability

- **Issue 1: Use of patient-month construct.**
  From Reviewer 1: “In this measure as well as others in this set, I'm concerned about the use of patient-month as the unit of counting for numerators and denominators. It would seem as if these cannot be independent of each other when a given patient contributes up to 12 observations per year.”

  From Reviewer 3: “The key concern I have with this measure is that it doesn't account for lack of independence among patient-months from the same patient.”

  - Developer Response 1: We thank the SMP Reviewers for your consideration and comments. The intent behind our use of the patient-month construct is to account for patients’ potentially variable time contributions to both the numerator and denominator. We recognize from an empirical standpoint that observations from the same person will never by truly independent of each other. However, given the considerable influx and efflux of patients contributing to a provider’s home dialysis rate throughout a given measurement year, we believe use of the patient-month construct is necessary to provide an accurate annual assessment of that rate. This is of particular importance with the recent implementation of the End-Stage Renal Disease (ESRD) Treatment Choices (ETC) Model under which we expect the program’s strong focus on increasing home dialysis utilization will result in both a rapid uptake of these modalities and a marked increase in treatment failure rates,

with the resulting potential for wide variations in patient-level performance from month to month. We in fact posit that a "patient construct" would provide an inaccurate and effectively meaningless picture of providers' home dialysis rates. For instance, use of a "patient construct" here would erroneously equate a provider that achieves a home dialysis rate of 20 percent for even a single month of the measurement year with a provider that maintains a 20 percent rate across the full twelve months. This would create a scenario in which a provider could start a large number of patients on a home modality towards the end of the measurement year simply to perform well on the measure. Conversely, a "patient-month construct" incentivizes earlier starts (i.e., the more months a patient is on a home modality in a given year, the higher the score) and tracks a provider's performance across time, minimizing such opportunities for measure "gaming."

- We also note that a substantial proportion of NQF-endorsed measure do, in fact, utilize patient-months and question why this construct would be prohibitive in this instance. Nevertheless, as our intent is to develop meaningful, reliable, and valid measures, we value and welcome the SMP's input and recommendations on how they believe this measure might be modified to better address its concerns in this regard.

- **Issue 2: Use of parent companies within Hospital Referral Regions as accountable entity.**
From Reviewer 5: "I am having trouble understanding the accountable entity of level of attribution. The developers explain that for dialysis facilities that are not subsidiaries of a parent organization, the individual facility is the accountable entity. But for facilities that are owned by a parent company, all dialysis facilities owned by the same parent within an HRR are aggregated (because they may refer all home dialysis patients to a separate, wholly or partially owned entity within that HRR). This is a complicated arrangement that will be very difficult to explain. Later in the submission packet, the developers report reliability at the HRR and facility levels, and validity at the HRR level. But the actual accountable entities appear to be a hybrid of facilities and HRRs; in other words, they are groups of facilities under the same ownership within the same HRR. Since the actual accountable entities are NEITHER the 5,694 separate facilities nor the 296 HRRs, it is very hard to interpret the reliability and validity data presented later."

  From Reviewer 11: "Based on the measure description, is this a comparison among HRRs? Who (person) will be responsible for improving this value? What is the consequence of being the worst (or best) performing HRR on this measure?"

  - Developer Response 2: We thank the SMP Reviewers for your consideration and comments. As noted in the submission documents, we have developed this measure for potential use in the recently launched ESRD Treatment Choices (ETC) program. We have thus modeled our level of analysis to reflect the reality of how this issue is being addressed within the ETC model; specifically, CMS will aggregate the home dialysis rate across dialysis facilities under the same legal entity (parent organization) within the same Hospital Referral Region (HRR). We believe this approach is fair and respects the existing business structure many organizations have developed around home dialysis. That is, **t**o account for home dialysis-only facilities within an HRR, particularly if many facilities within a given

organization/parent company send its home dialysis patients to such a provider, the measure aggregates facilities owned by the same company within a given HRR. As such, the accountable entity for this measure is the parent organization and the level of analysis is the aggregate of that organization's facilities within a given HRR. The parent organization would be responsible for improving the values obtained from the measure. We maintain that it would be impractical for us to develop a measure for a CMS Model that creates a different aggregation than that established through federal rulemaking and is the current law. We encourage the reviewers to acknowledge this reality and support measure development for this model, as well as other potential models in the future.

It should also be noted that KCQA is not a measure implementer and will not implement this measure or impose penalties or rewards for performance on the measure. However, that the current penalty structure of the ETC Model is as follows:

"Modality Performance Score (MPS) Calculation and Benchmarking: The MPS, a number between 0 and 6 points based on performance on home dialysis and transplant rates, will guide the providers' Performance Payment Adjustment (PPA). Specifically, CMS will take the better of a provider's achievement and improvement performance to calculate the MPS for a given measurement period. Providers will be scored from 0 to 2 points in 0.5-point increments based on the established Achievement Benchmark for the corresponding Benchmark Year, where the 30th, 50th, 75th, and 90th percentile values from the Comparison HRR group distribution would define ranges for each score;[a] and the Improvement Benchmark established by the provider's own historic performance, where no improvement, >0-5%, >5-10%, and >10% improvement define ranges for each score."

Achievement Benchmark	Improvement Benchmark	MPS Points
90th + percentile	**Not a scoring option**	**2**
75th + percentile	**>10%**	**1.5**
50th + percentile	**>5%**	**1**
30th + percentile	**>0%**	**0.5**
<30th + percentile	**No improvement**	**0**

- **Issue 3: Reliability testing at HRR level.**
  From Reviewer 11: "Measure claims to compare performance at HRR level—not the facility level. Either measure is not described properly or no reliability test is provided at HRR level."
    - Developer Response 3: We thank the SMP Reviewers for your consideration and comments. As noted in the submission documents, we performed reliability testing at both the facility and parent organization HRR levels. At the HRR level, reliability estimates were as follows:

HRRs	Alpha	Beta	Min	10th Pctl	Median	90th Pctl	Max	Mean
296	**4.415**	**22.62**	**0.9435**	**0.9857**	**0.997**	**0.9995**	**1**	**0.9943**

- **Issue 4: Paired measure set vs composite.**
  From Reviewer 7: "From the description in sp.03, it seems that the developers' recommendation to pair the proposed measure with the Home Dialysis Retention measure is essential to avoid unintended consequences of a stand-alone home dialysis rate measure. Can the developers provide a reasonable explanation on why these two measures are not submitted as a composite measure?

    - Developer Response 4: We thank the SMP Reviewers for your consideration and comments. We conferred with NQF Staff on whether the measures should be approached as a paired set or as a composite and were advised that the former would be more appropriate. Specifically, because the two measures result in distinct, individual scores rather than a single, rolled-up score, they are more consistent with NQF's definition of paired measures than composite. Given NQF staff suggested the paired set as the preferred option, we are confused by the SMR Reviewers suggesting that this advice was inappropriate to follow. However, we again note that we value and welcome the SMP's input and recommendations as to the best approach in this regard.

  We would also like to express our disappointment that the SMP has removed KCQA's Home Dialysis Retention Measure from further consideration without the opportunity for discussion. As we note in the submission documents, we believe the Retention Measure's reliability estimates were low because of uniformly high performance across providers during the testing period—2020, prior to implementation of the ETC Program. However, as we explain, the Retention Measure is unusual in that it is a *proactive* metric that seeks to identify and pre-empt the predicted *future* performance gap that will occur with the rapidly changing home dialysis landscape expected with deployment of the ETC Model. We emphasize that patients have consistently and vociferously raised concerns that CMS efforts to incentivize home dialysis through the ETC model and other programs will undoubtedly lead to cases wherein home dialysis modalities are prescribed for clinically and otherwise inappropriate patients. Currently, there is little to no problem in this regard—this is a *new* concern stemming from application of a *new* program. The KCQA Retention Measure seeks to elevate the patient-voice in the ETC Program by creating a counterbalancing deterrent to starting potentially inappropriate patients on home dialysis, as might be incentivized by the financial bonus associated with this step.

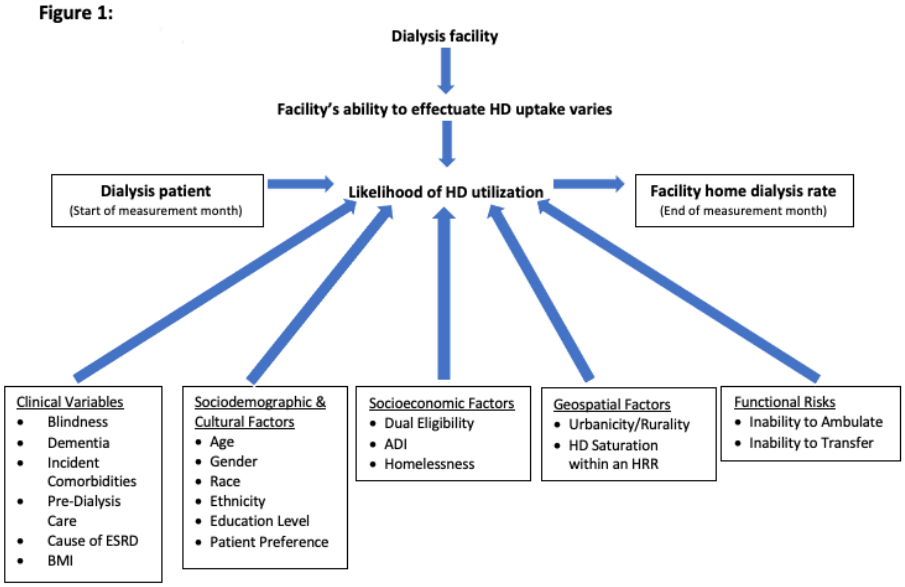
  Without a Retention Measure, there is no such counterbalance. Patients have prioritized the need for this balance. We would hope that the SMR Reviewers would prioritize patient-centered concerns and recognize that some measures meant to avoid future problem will not show gaps in current data, but are necessary to endorse to address anticipate gaps because of changes in federal law.

- **Issue 5: Risk stratification vs adjustment.**

- From Reviewer 7: "SP.22: There is no description of exactly how and at what stage the measure is stratified. Also, from the specifications it is not clear why the developers have selected a stratification approach and not a statistical risk adjustment model."
    - Developer Response 5: We thank the SMP Reviewers for your consideration and comments. We again note that KCQA is not a measure implementer and thus, while we are making the recommendation that results be stratified as indicated, we will not ultimately determine if and how the measure is stratified when deployed.

Additionally, as noted in the submission documents, our approach was based, precisely, on NQF's August 2021 report on social and functional status-related risk model guidance.[b] Specifically, as directed in that report, we first developed a conceptual model illustrating the pathway between the social and functional status-related risk factors, patient clinical factors, healthcare processes, and the measured healthcare outcome (home dialysis rate/retention). The NQF report notes that all demographic, clinical risk factors, social and functional risks, and patient preferences related to the outcome of interest, regardless of whether they can be operationalized in available data, should be considered for inclusion in the conceptual model. In particular, NQF specifically recommends that the following variables be evaluated when assessing the need for risk adjustment or stratification: age; gender; race/ethnicity; urbanicity/rurality; Medicare and Medicaid dual eligibility; indices of social vulnerability such as the Area Deprivation Index (ADI); and markers of functional risk such as frailty. For the KCQA Home Dialysis Rate Measure, based on our literature reviews and expert opinion from our Home Dialysis Workgroup and Steering Committee, we identified numerous such risk factors believed to impact home dialysis rates:

Figure 1:



[b] National Quality Forum. *Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance.*

This home dialysis conceptual model then guided our selection of candidate risk factors. We identified patient sociodemographic, socioeconomic, and geospatial factors and clinical variables, including comorbidities and measures of frailty and disability. These reflect the characteristics of the patients at the start of each measurement month and are independent of the quality of care provided. Potential clinical variables included not only incident clinical comorbidities, but also measures of pre-dialysis care, cause of ESRD, BMI, and frailty/functional status. We also considered social risk factors that may influence patients' access to home dialysis (e.g., geospatial considerations) and other barriers (e.g., homelessness) outside the control of a given dialysis facility. Variables in all of these domains have been found or are hypothesized to be associated with home dialysis utilization.[c,d,e] However, the domains differ in the extent to which we expect an individual dialysis facility or group of facilities to be able to mitigate the barriers to home dialysis conferred by such variables. These differences inform their potential use as risk adjusters, since adjusting for factors that can be more easily mitigated by higher quality care is more likely to mask low-quality care.

As noted in NQF's report, however, some of these variables were ultimately eliminated during the testing phase when we were able to better identify issues with data availability, statistical issues (e.g., confounding), and model performance. For instance, we found that several necessary data elements were not consistently available across dialysis providers and/or payers, such as pre-dialysis care, incident comorbidities, homelessness, education level, and proxy markers of functional decline (i.e., inability to transfer/ambulate). Operationalizing the ADI data element was unfeasible without considerable additional burden to our testing sites, and the Workgroup and Steering Committee agreed that patient preference would be difficult to accurately and reliably capture—and might introduce considerable risk of "gaming" the measure. Ultimately, the risk variables our committees agreed are not easily mitigable (and are thus appropriate variables for risk adjustment and/or stratification) and are operationalizable include age, gender, race, ethnicity, and dual eligibility status.

Finally, and again in accordance with NQF's recent guidance, we used Poisson regression models[f] and reliability measures to estimate adjusted outcomes to assess the effect of various social risk factors on the measures. As indicated in our submission documents, models for age, race, and dual eligibility were statistically significant, but changes in

---

[c] United States Renal Data System. ***2020 USRDS Annual Data Report:*** Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[d] Mehrotra R et al. Racial and ethnic disparities in use of and outcomes with home dialysis in the United States. J Am Soc Nephrol. 2016;27:2123–2134.

[e] Weiner D and Meyer K. Home dialysis in the United States: To increase utilization, address disparities. (Editorial.) Kidney Medicine. 2020;2(2):95-97.

[f] Due to overdispersion in the data, a quasi-Poisson regression model was fit to each risk factor—quasi-Poisson models explicitly model an overdispersion parameter.

overall measure scores were slight with application of the models, indicating that risk-adjustment has little impact on measure performance. Taken in conjunction with the concern that adjustment for such sociodemographic variables could obscure important, well-documented, and persistent disparities in home dialysis use in the US,[g,h] potentially setting lower standards of quality for more disadvantaged patient populations, KCQA agreed that risk-adjustment of this measure is both unnecessary and inappropriate.

Yet while risk-adjustment has little impact on overall measure performance, stratification by risk category highlights appreciable variations in performance across various sociodemographic and socioeconomic variables:

Category	Min	Q1	Median	Mean	Q3	Max	Facilities Included
Age 0 to < 18	0.0%	0.0%	0.0%	39.5%	100.0%	100.0%	132
Age 18 to < 25	0.0%	0.0%	0.0%	23.2%	37.5%	100.0%	2316
Age 25 to < 35	0.0%	0.0%	0.0%	18.6%	27.9%	100.0%	4954
Age 35 to < 45	0.0%	0.0%	0.0%	17.4%	26.7%	100.0%	5477
Age 45 to < 55	0.0%	0.0%	0.0%	15.9%	23.5%	100.0%	5641
Age 55 to < 65	0.0%	0.0%	0.0%	14.5%	20.3%	100.0%	5670
Age 65 to < 75	0.0%	0.0%	0.0%	13.6%	17.4%	100.0%	5665
Age 75 to < 85	0.0%	0.0%	0.0%	12.1%	13.0%	100.0%	5636
Age 85+	0.0%	0.0%	0.0%	8.5%	0.0%	100.0%	5041
Male	0.0%	0.0%	0.0%	14.5%	20.2%	100.0%	5690
Female	0.0%	0.0%	0.0%	14.4%	20.1%	100.0%	5685
White	0.0%	0.0%	0.0%	15.4%	22.4%	100.0%	5671
Black	0.0%	0.0%	0.0%	12.7%	13.5%	100.0%	5349
Other Race	0.0%	0.0%	0.0%	17.3%	22.6%	100.0%	4422
Dual eligible	0.0%	0.0%	0.0%	11.8%	11.5%	100.0%	5570
Overall	0.0%	0.0%	0.1%	**14.5%**	19.9%	100.0%	5699

Stratified analysis demonstrates that White patients (15.4%) are considerably more likely to utilize home dialysis modalities than Black patients (12.7%). There is also an incremental and steady decline in home dialysis with increasing age, with nearly 40% of patients < 18 years on home modalities, 17% among those aged 35-45, and < 12% among the 75+ age group. And less than <12% of dual-eligible patients use a home modality. While risk-adjustment might obscure these important inequities, potentially setting lower standards of quality for more sociodemographically vulnerable populations, we believe providers can and should use these stratified performance results to facilitate quality improvement efforts and focus resources on disparities reduction strategies. As such, we recommend that performance scores for the Home Dialysis Rate Measure be stratified by these sociodemographic variables.

- **Issue 6: Ethnicity data.**
  - From Reviewer 7: "The decision to use ethnicity for measure stratification is pending as explained, and should be clarified before the proposed stratification approach can be vetted."

[g] United States Renal Data System. ***2020 USRDS Annual Data Report:*** Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[h] Thorsness R, Wang V, Patzer R, et al. Association of social risk factors with home dialysis and kidney transplant rates in dialysis facilities. *JAMA.* 2021;326(22):2323-2325.

- Developer Response 6: We thank the SMP Reviewers for your consideration and comments. We are reviewing our ethnicity data and will provide these results to the SMP ASAP.

- **Issue 7: Beta-binomial reliability testing.**
  - From Reviewer 3: "The developers used the beta-binomial approach for reliability testing. Although this method has been used widely in the past and supported by NQF, there is an issue with this approach. The results of this testing actually highlights this problem. When P hat is 0 or 1, reliability will be 1 according to the formula."
  - Developer Response 7: We thank the SMP Reviewers for your consideration and comments. We recognize the limitations of the beta-binomial method. However, we also note the accuracy of the Reviewer's comment that this approach is supported by NQF and is widely used by developers submitting measures for endorsement consideration. We add that NQF has specifically advised developers to use the beta-binomial methodology, leading to our application of the approach here. If, however, NQF has new guidance on specific situations requiring the use of other methods and what other methods they now prefer be applied in those situations, we welcome and appreciate such feedback.
  - In the meantime, we posit it would be inconsistent for the SMR Reviewers to reject an accepted NQF method on a measure-by-measure basis, and inappropriate for a subcommittee to reject individual measures that have applied an NQF-endorsed method as recommended.
- **Issue 8: Reliability testing results.**
  - From Reviewer 3: "The results provided by the developers indicate close to perfect reliability for more than 90% facilities. This is not reassuring, rather it is unsettling. The developers show that for 19 facilities with less than 10 patient-months, the median reliability is 1. The developers conclude that "signal-to-noise reliability is extremely high for almost all facilities, regardless of size." This is a bug, rather than a feature. I think this is partly due to more than 40% facilities have 0% rate (page 27, that is, p hat is 0, resulting in 0 error variance based on the formula.)" Related, from Reviewer 5: "The numbers appear high on their face, but they are incorrect. The beta binomial assumptions do not apply here, given up to 12 non-independent observations on each patient within each facility. Further, the stratified results are inconsistent with the overall results. Reliability of 1 for small facilities with <10 patient-months are impossible; the Adams method does not account for sampling error with such small sample sizes."
  - Reviewer 5: "Reliability testing is invalid, for two reasons: 1. It was performed at the facility level, when the accountable entity is actually some aggregate of one or more facilities under the same ownership within the same HRR. 2. More importantly, the unit of this analysis was the patient-month. Patient-months are not independent of each other. Given that patients tend to stay on center dialysis, once started, or to stay on home dialysis, once started, there is a high level of within-facility correlation simply due to within-person correlation (autocorrelation). This analytic error grossly exaggerates reliability at the facility level (or any aggregate of facilities). The beta binomial method

from Adams cannot be used for this situation because of its assumptions (see RAND report)."

- Developer Response 8: We thank the SMP Reviewers for your consideration and comments. As previously noted, the intent behind our use of the patient-month construct is to account for patients' potentially variable time contributions to both the numerator and denominator. Given the considerable influx and efflux of patients contributing to a facility's home dialysis rate throughout a given measurement year, we believe use of the patient-month construct is necessary to provide an accurate assessment of that rate across that year; this is of particular importance with the recent implementation of the ESRD Treatment Choices (ETC) Model in which we expect the program's strong focus on increasing home dialysis will result in both a rapid uptake of these modalities and increased treatment failure rates, with the resulting potential for wide variations in performance from month-to-month. Again, however, as we do recognize that consecutive patient-months in the same patient are not entirely independent of each other, as well as the limitations of the beta-binomial method. We welcome SMP and NQF guidance on alternative approaches.

As noted above, however, we are deeply troubled that NQF supports and allows for the use of the methodology submitted with this measure, and that the SMR Reviewers seem to be overturning the broader NQF policy in this regard. It is important for NQF's credibility and leadership in the measure development community that it be consistent in its approach and that methods indicated as acceptable not be used against measure developers during the review process.

- **Issue 9: Reliability testing results.**
  From Reviewer 8: "I would appreciate a clarification about how the stratification approach was taken into account when assessing STN reliability. It looks as it was not, and that the only stratification tested was by number of patients-months. Given the recommendation to stratify the measure using 5 variables with over 200 strata (9 age categories, 2 sex categories, 3 race categories, and binary variables for dual eligibility and ethnicity), wouldn't the reliability testing be more appropriate if conducted per strata? As mentioned above, I may be misinterpreting the way the stratification approach is intended to be applied, thus the request for additional information."

  - Developer Response 9: Secondary to privacy concerns from the dialysis organizations that participated in our testing, we were not provided with patient-level data and so we could not perform certain analyses. We were limited to numerators and denominators within race group, age group, sex group, and the binary variables for dual eligibility and ethnicity. We performed risk-adjust reliability analyses to the extent that the data allowed:

Variable	N	Alpha	Beta	Min	10th Pctl	Median	90th Pctl	Max	Mean
Age	**5694**	**0.1968**	**0.8042**	**0.3912**	**0.9962**	**0.9998**	**1**	**1**	**0.9977**
Gender	**5694**	**0.1474**	**0.5724**	**0.4352**	**0.9970**	**0.9999**	**1**	**1**	**0.9981**
Race	**5694**	**0.1582**	**0.7272**	**0.3881**	**0.9962**	**0.9998**	**1**	**1**	**0.9977**

Variable	N	Alpha	Beta	Min	10th Pctl	Median	90th Pctl	Max	Mean
Dual-Eligibility	**5694**	**0.1725**	**0.8209**	**0.3694**	**0.9959**	**0.9998**	**1**	**1**	**0.9975**

Validity

- **Issue 1: Face Validity Panel.**
  From Reviewer 8: "In regard to the panel used to assess face validity, I'd suggest it's inappropriate to draw on members of the organization that is developing the measure. The conflict of interest is concerning here."

    - Developer Response 1: We thank the SMP Reviewers for your consideration and comments. Per NQF guidance, face validity of the measure was assessed through a systematic and transparent process by identified experts. We note that there is no rule within NQF or any other measure development organization that prohibits experts being from the community that will be subject to the measure. Some, but not all, of the expert panel are with organizations that are members of the KCQA, but that is no different than other specialty societies that develop measures. In addition, experts in the field of ESRD and dialysis care and uninvolved in the KCQA measure development process were identified from Kidney Care Partner (KCP) member organizations and were invited to participate in a formal face validity assessment of the KCQA Home Dialysis Measures. KCQA is a group of voluntary individuals led by a Steering Committee and two Co-Chairs. None of these individuals were members of the expert panel. The full membership of KCQA includes patients and patient organizations, dialysis facilities, manufacturers, and healthcare professionals from more than 30 different organizations in the kidney care community. There is no dues requirement and participation is voluntary. The membership votes on the final measure specifications for the purpose of moving a measure forward, but is not involved in the face validity evaluation.

    The reviewer inappropriately implies unethical behavior, including conflicts of interest. However, there is no indication that any member of the expert panel was unethical in his/her engagement or evaluation. There should be evidence of such behavior or other conflict before an accusation of abusive behavior is made. The KCQA follows the practice of peer-reviewed medical publications in providing the information about the relationships of the experts to members of the organization. It is unclear if the reviewer is suggesting that NQF should adopt a new standard that prohibits any expert who is also a member of an organization associated with a measure developer from participating. Such a standard would seem unreasonable as it would mean, for instance, that any physician who is a member of the AMA or one of its specialty societies could not participate in the development of physician measures. Similarly, CMS would need to revamp its technical expert panel process and remove all experts who had an interest in the outcome of the measure because they or their organization would be potentially subject to it. That would also fall within the conflict of interest contemplated by this comment. It seems that such a perfect standard would result in a lack of experts being available to support the development of meaningful measures.

- **Issue 2: Empirical validity testing.**

From Reviewer 7: "Although the empirical validity testing yielded acceptable results, I wonder why the correlation between basically two very similar measures was not higher than 0.7. Was this tested? Are the differences between HRR-level home dialysis rates from both sources mainly related to the different years (2018/2020)? Face validity results emphasize the recommendation that the Home Dialysis Rate measure be paired with the Home Dialysis Retention Measure."

- Developer Response 2: We thank the SMP Reviewers for your consideration and comments. While the answer to the question you raise is not wholly clear to use, we do speculate that the difference is related to the variation in our measurement years, as well as to the fact that CMS indicates that the scores presented in the Public Use File are calculated from data submitted for Dialysis Facility Reports and thus are not identical to the KCQA measure or the home dialysis specifications currently used in the ETC Model. They caution that the values reported in the PUF should be regarded only as "guides." We note that the intent of performing this correlative analysis was merely to demonstrate that the data used in our analyses are valid and consistent with the most recently available (2018) similar, but larger, data set collected by CMS. We also note that despite the difference in dates and organizations contributing to the two data sets, the resulting correlation is regarded as "high" (0.5-1.0), indicating a strong and positive correlation, as would be expected.

Issue 3: Exclusions.

From Reviewer 5: "The exclusions described are essential. However, additional exclusions should be considered to reflect the challenges of home dialysis in patients who have unstable housing, do not have secure utilities, do not have an ability to receive home deliveries of equipment, live alone with poor social support, have terminal illnesses such as advanced cardiopulmonary disease or metastatic cancer, or have severe cognitive impairment."

- Developer Response 3: We thank the SMP Reviewers for your consideration and comments. At this time, CMS does not collect social determinants of health (SDoH) data in a systematic way that would allow for such data to be considered reliable or valid for the creation of such exclusions. In addition, KCQA has consistently been careful not to allow exclusions to encompass the core population that the measure is intended to help identify as potential gaps. In the case of both the Rate and the Referral Measures, the SDoH factors the Reviewer identifies do create challenges for patients selecting home dialysis; however, these are the very patients that the kidney care community seeks to encourage having greater access to home dialysis. We encourage the SMP to avoid the perfect being the enemy of the good and, as the KCQA expert Workgroup determined, support a broader measure to incentivize the ability of individuals with these SDoH to access home dialysis.

- **Issue 4: Additional exclusions.**

  From Reviewer 8: "Some of exclusions can occur during the measurement period which may be reflective of poor quality care. Specifically: -Patients enrolled in hospice at any time in the measurement month. Patients residing in a nursing home or other LTCF at any time in the

measurement month. Thus, by omitting such circumstances, we may be decreasing our ability to measure quality."

- Developer Response 4: We thank the SMP Reviewers for your consideration and comments. The KCQA Workgroup and Steering Committee did debate whether to include nursing home and similar long-term care patients in the home dialysis measures. In the end, as the intent of the measures is to incentivize a shift in care from the in-center setting to home, the conclusion was dialysis in such long-term care settings is not consistent with the primary focus of the metrics. Additionally, it was noted that as all nursing home patients receiving dialysis with the nursing home are considered to be on "home" dialysis, the rate measure score would always be 100 percent and would thus be meaningless in this setting. Finally, the Workgroup and Steering Committee have also noted that these exclusions align with the ETC Model structure, which would facilitate measure deployment within the program.

- **Issue 5: Selection of risk stratification over adjustment.**
  From Reviewer 5: "The developers present a strong conceptual model and rationale for risk-adjustment, but then they fall back on stratification for unclear reasons. Clearly, other measures in the same package of dialysis measures use risk-adjustment, so these measures are outliers in relying on stratification. Stratification on four separate features (age, gender, race, ethnicity, dual eligibility) simultaneously seems infeasible for an accountability measure."

  - Developer Response 5: We thank the SMP Reviewers for your consideration and comments. Specific to this measure, we note that adjustment for such sociodemographic variables could obscure important, well-documented, and persistent disparities in home dialysis use in the US,i,j potentially setting lower standards of quality for more disadvantaged patient populations; we maintain that risk stratification is the better approach when such disparities are known to exist. As described above, in accordance with NQF's recent risk adjustment guidance,k we identified potential risk variables using NQF's recommended approach to conceptual modeling, then used Poisson regression modelsl and reliability measures to estimate adjusted outcomes to assess the effect of various social risk factors on the measures. As indicated in our submission documents, models for age, race, and dual eligibility were statistically significant, but changes in overall measure scores were slight with application of the models, indicating that risk-adjustment has little impact on measure performance. KCQA thus agreed that risk-adjustment of this measure is both unnecessary and inappropriate. Notably, this decision is consistent

[i] United States Renal Data System. ***2020 USRDS Annual Data Report:*** Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[j] Thorsness R, Wang V, Patzer R, et al. Association of social risk factors with home dialysis and kidney transplant rates in dialysis facilities. *JAMA.* 2021;326(22):2323-2325.

[k] National Quality Forum. ***Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance.***

[l] Due to overdispersion in the data, a quasi-Poisson regression model was fit to each risk factor—quasi-Poisson models explicitly model an overdispersion parameter.

with our longstanding position against risk adjustment under certain circumstances—including for numerous measures addressing care in dialysis facilities.

- Yet, as noted previously, while risk-adjustment has little impact on overall measure performance, we have demonstrated that stratification by risk category highlights appreciable variations in performance across various sociodemographic and socioeconomic variables. Stratified analysis demonstrates that White patients (15.4%) are considerably more likely to utilize home dialysis modalities than Black patients (12.7%). There is also an incremental and steady decline in home dialysis with increasing age, with nearly 40% of patients < 18 years on home modalities, 17% among those aged 35-45, and < 12% among the 75+ age group. And less than <12% of dual-eligible patients use a home modality. While risk-adjustment might obscure these important inequities, potentially setting lower standards of quality for more sociodemographically vulnerable populations, we believe providers can and should use these stratified performance results to facilitate quality improvement efforts and focus resources on disparities reduction strategies. As such, we recommend that performance scores for the Home Dialysis Rate Measure be stratified by these sociodemographic variables.

- **Issue 6: Risk stratification application and selection.**
  From Reviewer 7: "Although risk-stratification is recommended, there is no clear description on how this should be done in practice other than stating that "we recommend that performance scores for the Home Dialysis Rate Measure be stratified by age, gender, race, ethnicity, and dual-eligibility."; and as noted above, reliability testing was conducted without taking into account risk-stratification. As a general comment regarding risk-adjustment, it seems that on one hand the developers recommend not to risk-adjust to avoid adjusting away inequalities, potentially setting lower standards of quality for more sociodemographically vulnerable populations, but then they recommend risk-stratification of those same factors. If risk-stratification is advised, it is not clear to me why the stratification approach is preferred to statistical risk-adjustment. If developers decide not to risk-adjust, why is risk-stratification recommended, as it is a way to risk-adjust? I find this to be confusing and contradictory, unless there is a major point that I am missing. This leads me to an insufficient rating of validity until this point is clarified."

    - Developer Response 6: We thank the SMP Reviewers for your consideration and comments. As noted above, KCQA is committed to addressing health inequities in the provision of dialysis care in the United States. As such, we have been on record many times indicating that the vast majority of dialysis-related measures should *not* be risk-adjusted because doing so would result in perpetuating inequities in care. This concern is appropriate to take into account when thinking about home dialysis measures. The adjusted prevalence of ESRD in Black individuals in 2019 was 78.7% higher than in the next highest group (Native Americans) and more than fourfold higher than their White counterparts;[m] in that same year, only 7.8% of Black and

[m] USRDS *2021 Annual Data Report*.

7.9% of Hispanic ESRD patients selected home dialysis as their modality, compared to 10.8% of non-Hispanic Whites.[n] 22.3% of point prevalent and 9.3% of incident dialysis patients used dual Medicare and Medicaid in 2019.[o] Such inequities are one of the things this measure seeks to address.

In terms of the specification questions, we provide the following responses: Again, KCQA is not a measure implementer and we are thus unable to dictate how—and if—the measure will ultimately be stratified. As a measure developer creating metrics intended for use in CMS's federal accountability programs, we do and will continue to work closely with CMS to help ensure that our measures are adopted and implemented as intended, but the practical application of implementation issues such as risk stratification is beyond our realm of influence or control. And as noted in a prior response, KCQA believes risk stratification is preferred to risk adjustment in scenarios in which there are known and/or empirically demonstrated disparities in care, the rationale being that adjustment obscures ("adjusts away") real differences in performance between different socioeconomic or sociodemographic groups. Risk stratification is typically the preferred approach in such instances, as this approach instead highlights the disparities by drawing explicit attention to performance variations across these groups.

- **Issue 7:**
  From Reviewer 8: "There is a lack of a systematic review, consideration and testing of potential risk factors that may be salient to the given measure. The measure submitter puts forth several variables to assess, but neglects to provide an explanation as to the rationale for selecting such variables and not other variables."
    - Developer Response 7: We thank the SMP Reviewers for your consideration and comments. As described above and as noted in the submission documents, we used NQF's August 2021 report[p] on social and functional status-related risk model guidance to develop a conceptual model illustrating the pathway between the social and functional status-related risk factors, patient clinical factors, healthcare processes, and the measured healthcare outcome. The report notes that all demographic, clinical risk factors, social and functional risks, and patient preferences related to the outcome of interest, regardless of whether they can be operationalized in available data, should be considered for inclusion in the conceptual model. In particular, NQF recommends that the following variables be evaluated when assessing the need for risk adjustment or stratification: age; gender; race/ethnicity; urbanicity/rurality; Medicare and Medicaid dual eligibility; indices of social vulnerability such as the Area Deprivation Index (ADI); and markers of

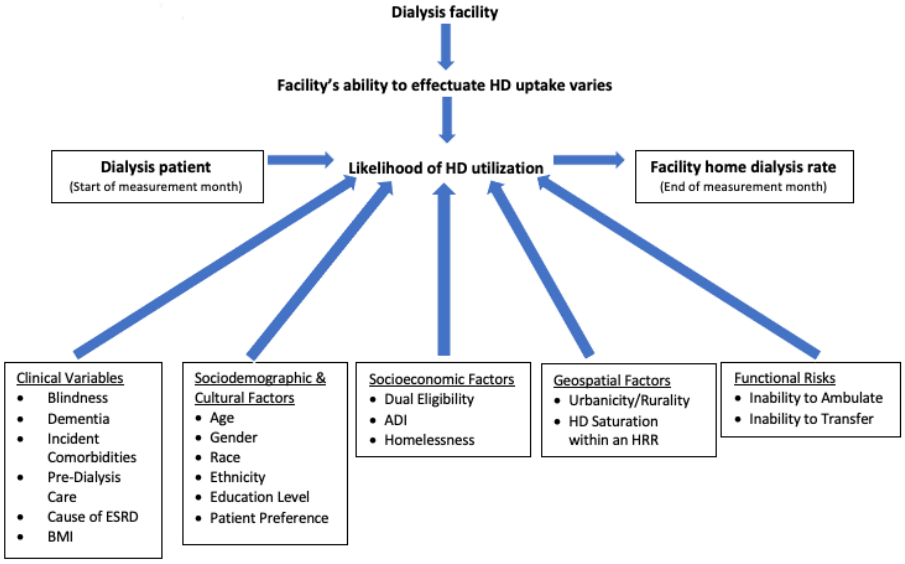
[n] IBID.

[o] IBID.

[p] National Quality Forum. ***Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance.***

functional risk such as frailty. For the KCQA Home Dialysis Rate Measure, based on our literature reviews and expert opinion from our Home Dialysis Workgroup and Steering Committee, we identified numerous risk factors believed to impact home dialysis rates:

Figure 1:



- This home dialysis conceptual model guided our selection of candidate risk factors. We identified patient sociodemographic, socioeconomic, and geospatial factors and clinical variables, including comorbidities and measures of frailty and disability. These reflect the characteristics of the patients at the start of each measurement month and are independent of the quality of care provided. Potential clinical variables included not only incident clinical comorbidities, but also measures of pre-dialysis care, cause of ESRD, BMI, and frailty/functional status. We also considered social risk factors that may influence patients' access to home dialysis (e.g., geospatial considerations) and other barriers (e.g., homelessness) outside the control of a given dialysis facility. Variables in all of these domains have been found or are hypothesized to be associated with home dialysis utilization.[q,r,s] However, the domains differ in the extent to which we expect an individual dialysis facility or group of facilities to be able to mitigate the barriers to home dialysis conferred by such variables. These differences inform their potential use as risk adjusters, since adjusting for factors that can be more easily mitigated by higher quality care is more likely to mask low-quality care.

---

[q] United States Renal Data System. ***2020 USRDS Annual Data Report:*** Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[r] Mehrotra R et al. Racial and ethnic disparities in use of and outcomes with home dialysis in the United States. J Am Soc Nephrol. 2016;27:2123–2134.

[s] Weiner D and Meyer K. Home dialysis in the United States: To increase utilization, address disparities. (Editorial.) Kidney Medicine. 2020;2(2):95-97.

As noted in NQF's report, however, some of these variables were ultimately eliminated during the testing phase when we were able to better identify issues with data availability, statistical issues (e.g., confounding), and model performance. For instance, we found that several necessary data elements were not consistently available across dialysis providers and/or payers, such as pre-dialysis care, incident comorbidities, homelessness, education level, and proxy markers of functional decline (i.e., inability to transfer/ambulate). Operationalizing the ADI data element was unfeasible without considerable additional burden to our testing sites, and the Workgroup and Steering Committee agreed that patient preference would be difficult to accurately and reliably capture—and might introduce considerable risk of "gaming" the measure. Ultimately, the risk variables our committees agreed are not easily mitigable (and are thus appropriate variables for risk adjustment) and are operationalizable include age, gender, race, ethnicity, and dual eligibility status.

- **Issue 8:**
  From Reviewer 12: "I am not certain if no risk adjustment is the right choice here. Some data are hard to get, but does that mean they don't matter? They recc stratified interpretation without RA."
    - **Developer Response 8:** We thank the SMP Reviewers for your consideration and comments. Please see our preceding response for an explanation as to why we selected risk stratification over risk adjustment, and how we selected our risk variables. Also, consistent with our response above, individuals living with ESRD and receiving dialysis are disproportionally Black and Brown, with the adjusted prevalence of ESRD in Black individuals 78.7% higher than in the next highest group (Native Americans) and more than fourfold higher than their White counterparts in 2019.[t] Many of these individuals are also low-income, as evidenced by their disproportionate dual eligibility for Medicare and Medicaid.[u] KCQA is committed to addressing health inequities in the provision of dialysis care for these patients. However, adjustment for social risks has remained controversial for fear of masking disparities or tacitly forgiving lower quality of care for socially marginalized patients.[v] We note that the Minimum Standards put forth in NQF's recent risk adjustment guidance report[w] indicate that stratification can be an appropriate alternative to risk adjustment, subject to the developer's assessment of the role of social and functional risk factors in the context of the specific intended use of the measure. We agree, and we maintain that stratification is indeed the most appropriate approach to social risk in many instances, allowing providers and other

[t] USRDS *2021 Annual Data Report*.

[u] IBID.

[v] HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE). *Report to Congress: Social Risk Factors and Performance in Medicare's Value-Based Purchasing Programs*. March 2020.

[w] National Quality Forum. ***Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance.***

healthcare stakeholders to identify and prioritize differences in care, outcomes, and experiences across different sociodemographic groups, and to develop and implement equity-focused practices to better address disparities and understand the experiences of patients from marginalized communities.[x] Such insights would be obscured if the same measures were instead adjusted for social risks. If this measure were to be risk-adjusted, it would result in the very patients we are trying to track and create incentives for adopting home dialysis out of the measure. To eliminate current inequities, it is not appropriate to "risk-adjust them out of the measure" and more appropriate to use a stratification approach.

Measure Number: 3689

Measure Title: First Year Standardized Waitlist Ratio (FYSWR)

Measure Developer/Steward: University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services

Validity

- **Issue 1 (section 18): The developers report a surprisingly high missing-practitioner rate (inability to attribute) of 6.2%. It is extremely unclear why this problem does not exist for #3694 and #3695, when these are all measures of waitlisting among patients on dialysis under the care of dialysis practice groups. The same data source (IDR) is used for all three measures.**
    - **Developer Response 1:** Although the IDR is used in all three measures to link a given provider to their practice group, the attribution of patients to individual providers is done through the CMS 2728 form for this measure (this allowed inclusion of patients with all forms of health insurance, not just Medicare primary), whereas it is done through Medicare dialysis claims for the other measures.

- **Issue 2 (section 19e): Concerns about inclusion of social risk adjustment in the models**
    - Developer Response 2: This was an area of significant concern particularly for Reviewer 5. We did not take our decision to include these factors lightly, and certainly are very aware of existing disparities in access to the transplant waitlist; our decision to propose this measure is in large part motivated by a desire to reduce such disparities. For this reason, we did not adjust for race, as it may serve to sustain known racial disparities and structural racism. However, the factors we chose (ADI, dual eligibility) do have a conceptual basis in that they are proxies for financial and social resources that can affect success following transplantation. Although the KDIGO candidate guidelines are appropriately circumspect about the influence on candidacy given limited empirical evidence, they clearly advocate psychosocial support assessment, which, among other things, includes "social history (e.g., education, occupation, financial resources, important relationships, and living circumstances)". This leads many transplant centers to incorporate such judgements into their decision-making, affecting or at least delaying waitlisting in

[x] *See* Advancing Health Equity. "Using Data to Reduce Disparities and Improve Quality." https://www.solvingdisparities.org/sites/default/files/Using%20Data%20Strategy%20Overview%20Oct.%202020.pdf (accessed June 22, 2021).

ways that may be outside dialysis practitioner control. A Technical Expert Panel consisting of a range of stakeholders, including several patients with ESRD, discussed these issues and were in consensus about the need for social risk adjustment. A dominant concern was that in the absence of such adjustment, dialysis practitioners caring for a disproportionate share of socially vulnerable patients may inappropriately be penalized by the measure, leading to unintended adverse consequences in terms of access to care for these patients.

- **Issue 3 (sections 16 and 17): Request for more detail in the approach, and concerns about strength of associations for construct validity.**
    - Developer Response 3: In response to requests for more detail on the calculations of mortality and transplant mortality rates for the construct validity analyses, we provide the following description: Mortality and transplant rates are calculated as count of total events (i.e. death or transplant) divided by sum of total patient years at risk within each practitioner group. We divided practitioner groups into 3 tertiles based on their measure (i.e. SWR/PPPW/aPPPW) performance. In each tertile, we computed means of mortality and transplant rates of the practitioner groups. We also tested the trend of these outcomes across tertiles. Spearman correlations were performed between the practitioner group measure values and their mortality and transplant rates respectively.

With respect to concerns that the associations demonstrated as part of our construct validity assessment are modest, we agree they are but, as we discussed in the relevant section, we believe this is expected given that mortality depends on many factors unrelated to waitlisting and transplantation depends on additional system level factors such as organ availability. The direction of the relationships were uniformly in the expected direction. Beyond construct validity, although we did not conduct a formal face validity assessment with an independent panel, we did engage with a Technical Expert Panel (TEP) during the development process. This TEP included stakeholders and experts (dialysis nephrologists, transplant nephrologists, transplant surgeon, researchers) and patient advocates, and reviewed a systematic compilation of relevant literature. The TEP expressed majority support for development of this waitlisting quality measure directed at dialysis practitioners.

- **Issue 4 (section 20): Concerns were expressed about the low % of flagged (i.e. worse or better than expected performance) practitioner groups and ability to distinguish clinically meaningful performance.**
    - Develop Response 4: Our intention was to reliably and validly identify outlying performance which clearly stood above or below typical performance levels. To achieve this with appropriate levels of certainty, we ultimately identified only a minority of practitioner groups with outlying levels of performance. Nevertheless we still feel this is impactful. For example, with a mean of 134 patients per practitioner group, the 135 practitioner groups identified with outlying performance (better or worse than expected) still represents care of thousands of patients. Further, despite the average FYSWR in the worse than expected group of 0.19, there was a wide range of performance within these practitioner groups, ranging from 0 to as high as 0.66.

- **Issue 5 (section 19e): A concern was raised about timing of when data elements from the CMS-2728 form were incorporated into the measure assessment.**
    - **Developer Response 5:** By design this measure assesses outcomes with a start date equal to the first initiation of dialysis for each patient in order to examine performance in the first year following start of dialysis. The CMS-2728 form is required for registration of ESRD status and is therefore completed near the start of dialysis. As such the comorbid conditions reflected in the form and used in the risk adjustment are likely to be present at dialysis start, and not reflective of care delivered by dialysis practitioners over the measurement period.
- **Issue 6 (section 19e): There were questions about the absence of validation with an external data set.**
    - Developer Response 6: We did not perform validation with an external data set given we are already using national data inclusive of essentially the universe of patients to which the measure would be directed.

Measure Number: 3694

Measure Title: Percentage of Prevalent Patients Waitlisted in Active Status (aPPPW)

Measure Developer/Steward: University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services

Validity

- **Issue 1 (section 19e): Concerns about inclusion of social risk adjustment in the models**
    - Developer Response 1: This was an area of significant concern particularly for Reviewer 5. We did not take our decision to include these factors lightly, and certainly are very aware of existing disparities in access to the transplant waitlist; our decision to propose this measure is in large part motivated by a desire to reduce such disparities. For this reason, we did not adjust for race, as it may serve to sustain known racial disparities and structural racism. However, the factors we chose (ADI, dual eligibility) do have a conceptual basis in that they are proxies for financial and social resources that can affect success following transplantation. Although the KDIGO candidate guidelines are appropriately circumspect about the influence on candidacy given limited empirical evidence, they clearly advocate psychosocial support assessment, which, among other things, includes “social history (e.g., education, occupation, financial resources, important relationships, and living circumstances)”. This leads many transplant centers to incorporate such judgements into their decision-making, affecting or at least delaying waitlisting in ways that may be outside dialysis practitioner control. A Technical Expert Panel consisting of a range of stakeholders, including several patients with ESRD, discussed these issues and were in consensus about the need for social risk adjustment. A dominant concern was that in the absence of such adjustment, dialysis practitioners caring for a disproportionate share of socially vulnerable patients may inappropriately be penalized by the measure, leading to unintended adverse consequences in terms of access to care for these patients.
- **Issue 2 (sections 16 and 17): Request for more detail in the approach, and concerns about strength of associations for construct validity.**

- Developer Response 2: In response to requests for more detail on the calculations of mortality and transplant mortality rates for the construct validity analyses, we provide the following description: Mortality and transplant rates are calculated as count of total events (i.e. death or transplant) divided by sum of total patient years at risk within each practitioner group. We divided practitioner groups into 3 tertiles based on their measure (i.e. SWR/PPPW/aPPPW) performance. In each tertile, we computed means of mortality and transplant rates of the practitioner groups. We also tested the trend of these outcomes across tertiles. Spearman correlations were performed between the practitioner group measure values and their mortality and transplant rates respectively.

With respect to concerns that the associations demonstrated as part of our construct validity assessment are modest, we agree they are but, as we discussed in the relevant section, we believe this is expected given that mortality depends on many factors unrelated to waitlisting and transplantation depends on additional system level factors such as organ availability. The direction of the relationships were uniformly in the expected direction. Beyond construct validity, although we did not conduct a formal face validity assessment with an independent panel, we did engage with a Technical Expert Panel (TEP) during the development process. This TEP included stakeholders and experts (dialysis nephrologists, transplant nephrologists, transplant surgeon, researchers) and patient advocates, and reviewed a systematic compilation of relevant literature. The TEP expressed majority support for development of this waitlisting quality measure directed at dialysis practitioners

- **Issue 3: Questions about whether transplant waitlisting is a process or outcome measure.**
  - Developer Response 3: We consider transplant waitlisting to be an outcome measure, as it represents achievement and maintenance of a beneficial health status (reflecting absence of acute or unstable health conditions) suitable for transplant candidacy, which is dependent on dialysis practitioner interventions such as optimization of dialysis prescription, maintenance of optimal dialysis access, and proper management of underlying chronic health conditions such as cardiovascular disease and diabetes mellitus. Unlike transplantation itself, which is substantially affected by system level and random factors such as organ availability and willingness to donate, achievement of waitlisting is more directly related to interventions under dialysis practitioner control.

- **Issue 4 (section 20): Concerns were expressed about the low % of flagged (i.e. worse or better than expected performance) practitioner groups and ability to distinguish clinically meaningful performance.**
  - Developer Response 4: Our intention was to reliably and validly identify outlying performance which clearly stood above or below typical performance levels. To achieve this with appropriate levels of certainty, we ultimately identified only a minority of practitioner groups with outlying levels of performance. Nevertheless we still feel this is impactful. For example, with a mean of 134 patients per practitioner group, the 172 practitioner groups identified with outlying performance (better or worse than expected) still represents care of thousands of patients. Further, despite the average aPPPW in the worse than expected group of 3.4% (versus 12.5% in the as expected group), there was a wide range of performance within these practitioner groups, ranging from 0 to as high as 8.5%.

- **Issue 5 (section 19e): There were questions about the absence of validation with an external data set.**
    - Developer Response 5: We did not perform validation with an external data set given we are already using national data inclusive of essentially the universe of patients to which the measure would be directed.

- **Issue 6 (section 19e): A clarification was requested regarding from when comorbid conditions identified from Medicare claims would be included in the risk adjustment.**
    - Developer Response 6: As noted in the NQF form section sp29, Medicare claims from the year prior to the reporting period were used for the prevalent comorbidities.

- **Issue 7 (section 19e): Concern from Reviewer #3: 'Model equation on page 29 is not consistent with model specifications in the text. For example, on page 28, both transplant center fixed characteristics and random effect are listed. On page 41, it says "two-way interactions were examined and selected for the final model based on both the magnitude and statistical significance of the estimates."'**
    - Developer Response 7: In the formula, we denoted alpha for transplant center random effects and Z for patient characteristics. To clarify, Z includes both patient characteristics and transplant center fixed characteristics. The inclusion of the sentence "two-way interactions were examined and selected for the final model based on both the magnitude and statistical significance of the estimates" was an error, as the final model doesn't include interactions.

- **Issue 8 (section 19e): There were concerns about the non-independence of patient-months.**
    - Developer Response 8: To clarify, patients are contributing up to 12 observations per year (one observation per patient-month). To account for non-independence among patient-months we implemented the empirical null method (see full citation below) for determining if statistically significant differences can be identified. This empirical null method aims to separate underlying intrinsic variation (e.g. overdispersion due to correlations among patient-months) in dialysis practitioner group outcomes from variation that might be attributed to poor or excellent care. In section 2b.05 of the NQF forms, we provided the implementation details of the empirical null method. Citations for the methods are:

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. Journal of the American Statistical Association, 99(465):96–104.
Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Institute of Mathematical Statistics Monographs). Cambridge: Cambridge University Press.
Kalbfleisch, J. and Wolfe, R. (2013). On Monitoring Outcomes of Medical Providers. Statistics in Biosciences, 5(2):286–302.

Measure Number: 3695

Measure Title: Percentage of Prevalent Patients Waitlisted (PPPW)

Measure Developer/Steward: University of Michigan Kidney and Epidemiology Cost Center/Centers for Medicare & Medicaid Services

Validity

- **Issue 1 (section 19e): Concerns about inclusion of social risk adjustment in the models**
    - Developer Response 1: This was an area of significant concern particularly for Reviewer 5. We did not take our decision to include these factors lightly, and certainly are very aware of existing disparities in access to the transplant waitlist; our decision to propose this measure is in large part motivated by a desire to reduce such disparities. For this reason, we did not adjust for race, as it may serve to sustain known racial disparities and structural racism. However, the factors we chose (ADI, dual eligibility) do have a conceptual basis in that they are proxies for financial and social resources that can affect success following transplantation. Although the KDIGO candidate guidelines are appropriately circumspect about the influence on candidacy given limited empirical evidence, they clearly advocate psychosocial support assessment, which, among other things, includes "social history (e.g., education, occupation, financial resources, important relationships, and living circumstances)". This leads many transplant centers to incorporate such judgements into their decision-making, affecting or at least delaying waitlisting in ways that may be outside dialysis practitioner control. A Technical Expert Panel consisting of a range of stakeholders, including several patients with ESRD, discussed these issues and were in consensus about the need for social risk adjustment. A dominant concern was that in the absence of such adjustment, dialysis practitioners caring for a disproportionate share of socially vulnerable patients may inappropriately be penalized by the measure, leading to unintended adverse consequences in terms of access to care for these patients.

- **Issue 2 (sections 16 and 17): Request for more detail in the approach, and concerns about strength of associations for construct validity.**
    - Developer Response 2: In response to requests for more detail on the calculations of mortality and transplant mortality rates for the construct validity analyses, we provide the following description: Mortality and transplant rates are calculated as count of total events (i.e. death or transplant) divided by sum of total patient years at risk within each practitioner group. We divided practitioner groups into 3 tertiles based on their measure (i.e. SWR/PPPW/aPPPW) performance. In each tertile, we computed means of mortality and transplant rates of the practitioner groups. We also tested the trend of these outcomes across tertiles. Spearman correlations were performed between the practitioner group measure values and their mortality and transplant rates respectively.

    With respect to concerns that the associations demonstrated as part of our construct validity assessment are modest, we agree they are but, as we discussed in the relevant section, we believe this is expected given that mortality depends on many factors unrelated to waitlisting and transplantation depends on additional system level factors such as organ availability. The direction of the relationships were uniformly in the expected direction. Beyond construct validity, although we did not conduct a formal face validity assessment with an independent panel, we did engage with a Technical Expert Panel (TEP) during the development process. This TEP included stakeholders and experts (dialysis nephrologists, transplant nephrologists, transplant surgeon, researchers) and patient advocates, and reviewed a systematic compilation of relevant

literature. The TEP expressed majority support for development of this waitlisting quality measure directed at dialysis practitioners.

- **Issue 3: Questions about whether transplant waitlisting is a process or outcome measure.**
    - Developer Response 3: We consider transplant waitlisting to be an outcome measure, as it represents achievement and maintenance of a beneficial health status (reflecting absence of acute or unstable health conditions) suitable for transplant candidacy, which is dependent on dialysis practitioner interventions such as optimization of dialysis prescription, maintenance of optimal dialysis access, and proper management of underlying chronic health conditions such as cardiovascular disease and diabetes mellitus. Unlike transplantation itself, which is substantially affected by system level and random factors such as organ availability and willingness to donate, achievement of waitlisting is more directly related to interventions under dialysis practitioner control.

- **Issue 4 (section 20): Concerns were expressed about the low % of flagged (i.e. worse or better than expected performance) practitioner groups and ability to distinguish clinically meaningful performance.**
    - Developer Response 4: Our intention was to reliably and validly identify outlying performance which clearly stood above or below typical performance levels. To achieve this with appropriate levels of certainty, we ultimately identified only a minority of practitioner groups with outlying levels of performance. Nevertheless, we still feel this is impactful. For example, with a mean of 134 patients per practitioner group, the 186 practitioner groups identified with outlying performance (better or worse than expected) still represents care of thousands of patients. Further, despite the average PPPW in the worse than expected group of 6.7% (versus 18.6% in the as expected group), there was a wide range of performance within these practitioner groups, ranging from 0 to as high as 15.3%.

- **Issue 5 (section 19e): There were questions about the absence of validation with an external data set.**
    - Developer Response 5: We did not perform validation with an external data set given we are already using national data inclusive of essentially the universe of patients to which the measure would be directed.

- **Issue 6 (section 19e): A clarification was requested regarding from when comorbid conditions identified from Medicare claims would be included in the risk adjustment.**
    - Developer Response 6: As noted in the NQF form section sp29, Medicare claims from the year prior to the reporting period were used for the prevalent comorbidities.

- **Issue 7 (section 19e): There were concerns about the non-independence of patient-months.**
    - Developer Response 7: To clarify, patients are contributing up to 12 observations per year (one observation per patient-month). To account for non-independence among patient-months we implemented the empirical null method (see full citation below) for determining if statistically significant differences can be identified. This empirical null method aims to separate underlying intrinsic variation (e.g. overdispersion due to correlations among patient-months) in dialysis practitioner group outcomes from variation that might be attributed to poor or excellent care. In

section 2b.05 of the NQF forms, we provided the implementation details of the empirical null method. Citations for the methods are:

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. Journal of the American Statistical Association, 99(465):96–104.
Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Institute of Mathematical Statistics Monographs). Cambridge: Cambridge University Press.
Kalbfleisch, J. and Wolfe, R. (2013). On Monitoring Outcomes of Medical Providers. Statistics in Biosciences, 5(2):286–302.

Measure Number: 3697

Measure Title: Home Dialysis Retention

Measure Developer/Steward: Kidney Care Quality Alliance

Reliability

- **Issue 1: Use of parent companies within Hospital Referral Regions as accountable entity.** From Reviewer 5: "I am having trouble understanding the accountable entity of level of attribution. The developers explain that for dialysis facilities that are not subsidiaries of a parent organization, the individual facility is the accountable entity. But for facilities that are owned by a parent company, all dialysis facilities owned by the same parent within an HRR are aggregated (because they may refer all home dialysis patients to a separate, wholly or partially owned entity within that HRR). This is a complicated arrangement that will be very difficult to explain. Later in the submission packet, the developers report reliability at the HRR and facility levels, and validity at the HRR level. But the actual accountable entities appear to be a hybrid of facilities and HRRs; in other words, they are groups of facilities under the same ownership within the same HRR. Since the actual accountable entities are NEITHER the 5,694 separate facilities nor the 296 HRRs, it is very hard to interpret the reliability and validity data presented later."

  From Reviewer 11: "Based on the measure description, is this a comparison among HRRs? Who (person) will be responsible for improving this value? What is the consequence of being the worst (or best) performing HRR on this measure?"

  - Developer Response 1: We thank the SMP Reviewers for your consideration and comments. As noted in the submission documents, we have developed this measure for potential use in the recently launched ESRD Treatment Choices (ETC) program. We have thus modeled our level of analysis to reflect the reality of how this issue is being addressed within the ETC model; specifically, CMS will aggregate the home dialysis rate across dialysis facilities under the same legal entity (parent organization) within the same Hospital Referral Region (HRR). We believe this approach is fair and respects the existing business structure many organizations have developed around home dialysis. That is, to account for home dialysis-only facilities within an HRR, particularly if many facilities within a given organization/parent company send its home dialysis patients to such a provider, the measure aggregates facilities owned by the same company within a given HRR. As

such, the accountable entity for this measure is the parent organization and the level of analysis is the aggregate of that organization's facilities within a given HRR. The parent organization would be responsible for improving the values obtained from the measure. We maintain that it would be impractical for us to develop a measure for a CMS Model that creates a different aggregation than that established through federal rulemaking and is the current law. We encourage the reviewers to acknowledge this reality and support measure development for this model, as well as other potential models in the future.

It should also be noted that KCQA is not a measure implementer and will not implement this measure or impose penalties or rewards for performance on the measure. However, that the current penalty structure of the ETC Model is as follows:

"Modality Performance Score (MPS) Calculation and Benchmarking: The MPS, a number between 0 and 6 points based on performance on home dialysis and transplant rates, will guide the providers' Performance Payment Adjustment (PPA). Specifically, CMS will take the better of a provider's achievement and improvement performance to calculate the MPS for a given measurement period. Providers will be scored from 0 to 2 points in 0.5-point increments based on the established Achievement Benchmark for the corresponding Benchmark Year, where the 30th, 50th, 75th, and 90th percentile values from the Comparison HRR group distribution would define ranges for each score;[y] and the Improvement Benchmark established by the provider's own historic performance, where no improvement, >0-5%, >5-10%, and >10% improvement define ranges for each score."

Achievement Benchmark	Improvement Benchmark	MPS Points
90th + percentile	**Not a scoring option**	**2**
75th + percentile	**>10%**	**1.5**
50th + percentile	**>5%**	**1**
30th + percentile	**>0%**	**0.5**
<30th + percentile	**No improvement**	**0**

- **Issue 2: Reliability testing at HRR level.**
  From Reviewer 11: "Measure claims to compare performance at HRR level—not the facility level. Either measure is not described properly or no reliability test is provided at HRR level."
    - Developer Response 2: We thank the SMP Reviewers for your consideration and comments. As noted in the submission documents, we performed reliability testing at both the facility and parent organization HRR levels. At the HRR level, reliability estimates were as follows:

HRRs	Alpha	Beta	Min	10th Pctl	Median	90th Pctl	Max	Mean
292	**154.8**	**14.16**	**0.0036**	**0.0766**	**0.317**	**0.781**	**1**	**0.3787**

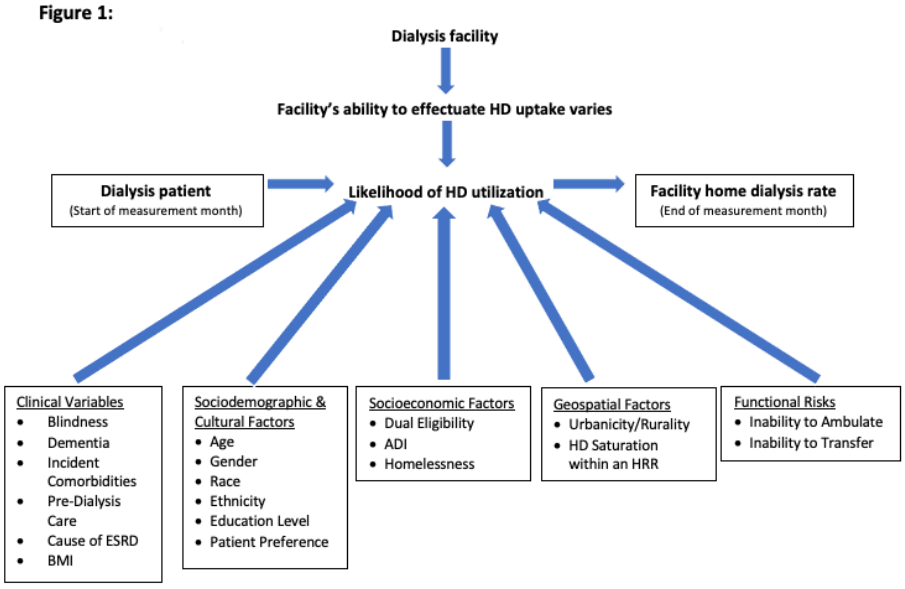
- **Issue 3: Paired measure set vs composite.**
  From Reviewer 7: "From the description in sp.03, it seems that the developers' recommendation to pair the proposed measure with the Home Dialysis Retention measure is essential to avoid unintended consequences of a stand-alone home dialysis rate measure. Can the developers provide a reasonable explanation on why these two measures are not submitted as a composite measure?
    - Developer Response 3: We thank the SMP Reviewers for your consideration and comments. We conferred with NQF Staff on whether the measures should be approached as a paired set or as a composite and were advised that the former would be more appropriate. Specifically, because the two measures result in distinct, individual scores rather than a single, rolled-up score, they are more consistent with NQF's definition of paired measures than composite. Given NQF staff suggested the paired set as the preferred option, we are confused by the SMR Reviewers suggesting that this advice was inappropriate to follow. However, we again note that we value and welcome the SMP's input and recommendations as to the best approach in this regard.
    - We would also like to express our disappointment that the SMP has removed KCQA's Home Dialysis Retention Measure from further consideration without the opportunity for discussion. As we note in the submission documents, we believe the Retention Measure's reliability estimates were low because of uniformly high performance across providers during the testing period—2020, prior to implementation of the ETC Program. However, as we explain, the Retention Measure is unusual in that it is a *proactive* metric that seeks to identify and pre-empt the predicted *future* performance gap that will occur with the rapidly changing home dialysis landscape expected with deployment of the ETC Model. We emphasize that patients have consistently and vociferously raised concerns that CMS efforts to incentivize home dialysis through the ETC model and other programs will undoubtedly lead to cases wherein home dialysis modalities are prescribed for clinically and otherwise inappropriate patients. Currently there is little to no problem in this regard—this is a *new* concern stemming from application of a *new* program. The KCQA Retention Measure seeks to elevate the patient-voice in the ETC Program by creating a counterbalancing deterrent to starting potentially inappropriate patients on home dialysis, as might be incentivized by the financial bonus associated with this step.
    - Without a Retention Measure, there is no such counterbalance. Patients have prioritized the need for this balance. We would hope that the SMR Reviewers would prioritize patient-centered concerns and recognize that some measures meant to avoid future problem will not show gaps in current data, but are necessary to endorse to address anticipate gaps because of changes in federal law.

- **Issue 4: Risk stratification vs adjustment.**
  - From Reviewer 7: "SP.22: There is no description of exactly how and at what stage the measure is stratified. Also, from the specifications it is not clear why the developers have selected a stratification approach and not a statistical risk adjustment model.

- Developer Response 4: We thank the SMP Reviewers for your consideration and comments. We again note that KCQA is not a measure implementer and thus, while we are making the recommendation that results be stratified as indicated, we will not ultimately determine if and how the measure is stratified when deployed.

Additionally, as noted in the submission documents, our approach was based, precisely, on NQF's August 2021 report on social and functional status-related risk model guidance.[z] Specifically, as directed in that report, we first developed a conceptual model illustrating the pathway between the social and functional status-related risk factors, patient clinical factors, healthcare processes, and the measured healthcare outcome (home dialysis rate/retention). The NQF report notes that all demographic, clinical risk factors, social and functional risks, and patient preferences related to the outcome of interest, regardless of whether they can be operationalized in available data, should be considered for inclusion in the conceptual model. In particular, NQF specifically recommends that the following variables be evaluated when assessing the need for risk adjustment or stratification: age; gender; race/ethnicity; urbanicity/rurality; Medicare and Medicaid dual eligibility; indices of social vulnerability such as the Area Deprivation Index (ADI); and markers of functional risk such as frailty. For the KCQA Home Dialysis Rate Measure, based on our literature reviews and expert opinion from our Home Dialysis Workgroup and Steering Committee, we identified numerous such risk factors believed to impact home dialysis rates:

Figure 1:



- This home dialysis conceptual model then guided our selection of candidate risk factors. We identified patient sociodemographic, socioeconomic, and geospatial factors and clinical variables, including comorbidities and measures of frailty and disability. These reflect the characteristics of the patients at the start of each measurement month and

[z] National Quality Forum. ***Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance.***

are independent of the quality of care provided. Potential clinical variables included not only incident clinical comorbidities, but also measures of pre-dialysis care, cause of ESRD, BMI, and frailty/functional status. We also considered social risk factors that may influence patients' access to home dialysis (e.g., geospatial considerations) and other barriers (e.g., homelessness) outside the control of a given dialysis facility. Variables in all of these domains have been found or are hypothesized to be associated with home dialysis utilization.[aa,bb,cc] However, the domains differ in the extent to which we expect an individual dialysis facility or group of facilities to be able to mitigate the barriers to home dialysis conferred by such variables. These differences inform their potential use as risk adjusters, since adjusting for factors that can be more easily mitigated by higher quality care is more likely to mask low-quality care.

As noted in NQF's report, however, some of these variables were ultimately eliminated during the testing phase when we were able to better identify issues with data availability, statistical issues (e.g., confounding), and model performance. For instance, we found that several necessary data elements were not consistently available across dialysis providers and/or payers, such as pre-dialysis care, incident comorbidities, homelessness, education level, and proxy markers of functional decline (i.e., inability to transfer/ambulate). Operationalizing the ADI data element was unfeasible without considerable additional burden to our testing sites, and the Workgroup and Steering Committee agreed that patient preference would be difficult to accurately and reliably capture—and might introduce considerable risk of "gaming" the measure. Ultimately, the risk variables our committees agreed are not easily mitigable (and are thus appropriate variables for risk adjustment and/or stratification) and are operationalizable include age, gender, race, ethnicity, and dual eligibility status.

Finally, and again in accordance with NQF's recent guidance, we used Poisson regression models[dd] and reliability measures to estimate adjusted outcomes to assess the effect of various social risk factors on the measures. As indicated in our submission documents, models for age, race, and dual eligibility were statistically significant, but changes in overall measure scores were slight with application of the models, indicating that risk-adjustment has little impact on measure performance. Taken in conjunction with the concern that adjustment for such sociodemographic variables could obscure important,

---

[aa] United States Renal Data System. ***2020 USRDS Annual Data Report***: Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[bb] Mehrotra R et al. Racial and ethnic disparities in use of and outcomes with home dialysis in the United States. J Am Soc Nephrol. 2016;27:2123–2134.

[cc] Weiner D and Meyer K. Home dialysis in the United States: To increase utilization, address disparities. (Editorial.) Kidney Medicine. 2020;2(2):95-97.

[dd] Due to overdispersion in the data, a quasi-Poisson regression model was fit to each risk factor—quasi-Poisson models explicitly model an overdispersion parameter.

well-documented, and persistent disparities in home dialysis use in the US,[ee,ff] potentially setting lower standards of quality for more disadvantaged patient populations, KCQA agreed that risk-adjustment of this measure is both unnecessary and inappropriate.

Yet while risk-adjustment has little impact on overall measure performance, stratification by risk category highlights appreciable variations in performance across various sociodemographic and socioeconomic variables:

Category	Min	Q1	Median	Mean	Q3	Max	Facilities Included
Age 0 to < 18	0.0%	0.0%	0.0%	39.5%	100.0%	100.0%	132
Age 18 to < 25	0.0%	0.0%	0.0%	23.2%	37.5%	100.0%	2316
Age 25 to < 35	0.0%	0.0%	0.0%	18.6%	27.9%	100.0%	4954
Age 35 to < 45	0.0%	0.0%	0.0%	17.4%	26.7%	100.0%	5477
Age 45 to < 55	0.0%	0.0%	0.0%	15.9%	23.5%	100.0%	5641
Age 55 to < 65	0.0%	0.0%	0.0%	14.5%	20.3%	100.0%	5670
Age 65 to < 75	0.0%	0.0%	0.0%	13.6%	17.4%	100.0%	5665
Age 75 to < 85	0.0%	0.0%	0.0%	12.1%	13.0%	100.0%	5636
Age 85+	0.0%	0.0%	0.0%	8.5%	0.0%	100.0%	5041
Male	0.0%	0.0%	0.0%	14.5%	20.2%	100.0%	5690
Female	0.0%	0.0%	0.0%	14.4%	20.1%	100.0%	5685
White	0.0%	0.0%	0.0%	15.4%	22.4%	100.0%	5671
Black	0.0%	0.0%	0.0%	12.7%	13.5%	100.0%	5349
Other Race	0.0%	0.0%	0.0%	17.3%	22.6%	100.0%	4422
Dual eligible	0.0%	0.0%	0.0%	11.8%	11.5%	100.0%	5570
Overall	0.0%	0.0%	0.1%	14.5%	19.9%	100.0%	5699

Stratified analysis demonstrates that White patients (15.4%) are considerably more likely to utilize home dialysis modalities than Black patients (12.7%). There is also an incremental and steady decline in home dialysis with increasing age, with nearly 40% of patients < 18 years on home modalities, 17% among those aged 35-45, and < 12% among the 75+ age group. And less than <12% of dual-eligible patients use a home modality. While risk-adjustment might obscure these important inequities, potentially setting lower standards of quality for more sociodemographically vulnerable populations, we believe providers can and should use these stratified performance results to facilitate quality improvement efforts and focus resources on disparities reduction strategies. As such, we recommend that performance scores for the Home Dialysis Rate Measure be stratified by these sociodemographic variables.

- **Issue 5: Ethnicity data.**
  - From Reviewer 7: "The decision to use ethnicity for measure stratification is pending as explained, and should be clarified before the proposed stratification approach can be vetted."

[ee] United States Renal Data System. 2020 USRDS Annual Data Report: Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[ff] Thorsness R, Wang V, Patzer R, et al. Association of social risk factors with home dialysis and kidney transplant rates in dialysis facilities. *JAMA.* 2021;326(22):2323-2325.

- Developer Response 5: We thank the SMP Reviewers for your consideration and comments. We are reviewing our ethnicity data and will provide these results to the SMP ASAP.

- **Issue 6: Beta-binomial reliability testing.**
  - From Reviewer 3: "The developers used the beta-binomial approach for reliability testing. Although this method has been used widely in the past and supported by NQF, there is an issue with this approach. The results of this testing actually highlights this problem. When P hat is 0 or 1, reliability will be 1 according to the formula."
  - Developer Response 6: We thank the SMP Reviewers for your consideration and comments. We recognize the limitations of the beta-binomial method. However, we also note the accuracy of the Reviewer's comment that this approach is supported by NQF and is widely used by developers submitting measures for endorsement consideration. We add that NQF has specifically advised developers to use the beta-binomial methodology, leading to our application of the approach here. If, however, NQF has new guidance on specific situations requiring the use of other methods and what other methods they now prefer be applied in those situations, we welcome and appreciate such feedback.
    In the meantime, we posit it would be inconsistent for the SMR Reviewers to reject an accepted NQF method on a measure-by-measure basis, and inappropriate for a subcommittee to reject individual measures that have applied an NQF-endorsed method as recommended.

Validity

- **Issue 1: Face Validity Panel.**
  From Reviewer 8: "In regard to the panel used to assess face validity, I'd suggest it's inappropriate to draw on members of the organization that is developing the measure. The conflict of interest is concerning here."
  - Developer Response 1: We thank the SMP Reviewers for your consideration and comments. Per NQF guidance, face validity of the measure was assessed through a systematic and transparent process by identified experts. We note that there is no rule within NQF or any other measure development organization that prohibits experts being from the community that will be subject to the measure. Some, but not all, of the expert panel are with organizations that are members of the KCQA, but that is no different than other specialty societies that develop measures. In addition, experts in the field of ESRD and dialysis care and uninvolved in the KCQA measure development process were identified from Kidney Care Partner (KCP) member organizations and were invited to participate in a formal face validity assessment of the KCQA Home Dialysis Measures. KCQA is a group of voluntary individuals led by a Steering Committee and two Co-Chairs. None of these individuals were members of the expert panel. The full membership of KCQA includes patients and patient organizations, dialysis facilities, manufacturers, and healthcare professionals from more than 30 different organizations in the kidney care community. There is no dues requirement and participation is voluntary. The membership votes on the final measure specifications for the purpose of moving a measure forward, but is not involved in the face validity evaluation.

The reviewer inappropriately implies unethical behavior, including conflicts of interest. However, there is no indication that any member of the expert panel was unethical in his/her engagement or evaluation. There should be evidence of such behavior or other conflict before an accusation of abusive behavior is made. The KCQA follows the practice of peer-reviewed medical publications in providing the information about the relationships of the experts to members of the organization. It is unclear if the reviewer is suggesting that NQF should adopt a new standard that prohibits any expert who is also a member of an organization associated with a measure developer from participating. Such a standard would seem unreasonable as it would mean, for instance, that any physician who is a member of the AMA or one of its specialty societies could not participate in the development of physician measures. Similarly, CMS would need to revamp its Technical Expert Panel process and remove all experts who had an interest in the outcome of the measure because they or their organization would be potentially subject to it. That would also fall within the conflict of interest contemplated by this comment. It seems that such a perfect standard would result in a lack of experts being available to support the development of meaningful measures.

- **Issue 2: Exclusions.**
  From Reviewer 5: "The exclusions described are essential. However, additional exclusions should be considered to reflect the challenges of home dialysis in patients who have unstable housing, do not have secure utilities, do not have an ability to receive home deliveries of equipment, live alone with poor social support, have terminal illnesses such as advanced cardiopulmonary disease or metastatic cancer, or have severe cognitive impairment."
    - Developer Response 2: We thank the SMP Reviewers for your consideration and comments. At this time, CMS does not collect social determinants of health (SDoH) data in a systematic way that would allow for such data to be considered reliable or valid for the creation of such exclusions. In addition, KCQA has consistently been careful not to allow exclusions to encompass the core population that the measure is intended to help identify as potential gaps. In the case of both the Rate and the Referral Measures, the SDoH factors the Reviewer identifies do create challenges for patients selecting home dialysis; however, these are the very patients that the kidney care community seeks to encourage having greater access to home dialysis. We encourage the SMP to avoid the perfect being the enemy of the good and, as the KCQA expert Workgroup determined, support a broader measure to incentivize the ability of individuals with these SDoH to access home dialysis.
- **Issue 3: Selection of risk stratification over adjustment.**
  From Reviewer 5: "The developers present a strong conceptual model and rationale for risk-adjustment, but then they fall back on stratification for unclear reasons. Clearly, other measures in the same package of dialysis measures use risk-adjustment, so these measures are outliers in relying on stratification. Stratification on four separate features (age, gender, race, ethnicity, dual eligibility) simultaneously seems infeasible for an accountability measure."
    - Developer Response 3: We thank the SMP Reviewers for your consideration and comments. Specific to this measure, we note that adjustment for such sociodemographic variables could obscure important, well-documented, and

persistent disparities in home dialysis use in the US,[gg,hh] potentially setting lower standards of quality for more disadvantaged patient populations; we maintain that risk stratification is the better approach when such disparities are known to exist. As described above**,** in accordance with NQF's recent risk adjustment guidance,[ii] we identified potential risk variables using NQF's recommended approach to conceptual modeling, then used Poisson regression models[jj] and reliability measures to estimate adjusted outcomes to assess the effect of various social risk factors on the measures. As indicated in our submission documents, models for age, race, and dual eligibility were statistically significant, but changes in overall measure scores were slight with application of the models, indicating that risk-adjustment has little impact on measure performance. KCQA thus agreed that risk-adjustment of this measure is both unnecessary and inappropriate. Notably, this decision is consistent with our longstanding position against risk adjustment under certain circumstances—including for numerous measures addressing care in dialysis facilities.

Yet, as noted previously, while risk-adjustment has little impact on overall measure performance, we have demonstrated that *stratification* by risk category highlights appreciable variations in performance across various sociodemographic and socioeconomic variables. Stratified analysis demonstrates that White patients (15.4%) are considerably more likely to utilize home dialysis modalities than Black patients (12.7%). There is also an incremental and steady decline in home dialysis with increasing age, with nearly 40% of patients < 18 years on home modalities, 17% among those aged 35-45, and < 12% among the 75+ age group. And less than <12% of dual-eligible patients use a home modality. While risk-adjustment might obscure these important inequities, potentially setting lower standards of quality for more sociodemographically vulnerable populations, we believe providers can and should use these stratified performance results to facilitate quality improvement efforts and focus resources on disparities reduction strategies. As such, we recommend that performance scores for the Home Dialysis Rate Measure be stratified by these sociodemographic variables.

- **Issue 4: Risk stratification application and selection.**
  From Reviewer 7: "Although risk-stratification is recommended, there is no clear description on how this should be done in practice other than stating that "we recommend that performance scores for the Home Dialysis Rate Measure be stratified by age, gender, race, ethnicity, and

[gg] United States Renal Data System. *2020 USRDS Annual Data Report:* Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[hh] Thorsness R, Wang V, Patzer R, et al. Association of social risk factors with home dialysis and kidney transplant rates in dialysis facilities. *JAMA.* 2021;326(22):2323-2325.

[ii] National Quality Forum. ***Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance***.

[jj] Due to overdispersion in the data, a quasi-Poisson regression model was fit to each risk factor—quasi-Poisson models explicitly model an overdispersion parameter.

dual-eligibility."; and as noted above, reliability testing was conducted without taking into account risk-stratification. As a general comment regarding risk-adjustment, it seems that on one hand the developers recommend not to risk-adjust to avoid adjusting away inequalities, potentially setting lower standards of quality for more sociodemographically vulnerable populations, but then they recommend risk-stratification of those same factors. If risk-stratification is advised, it is not clear to me why the stratification approach is preferred to statistical risk-adjustment. If developers decide not to risk-adjust, why is risk-stratification recommended, as it is a way to risk-adjust? I find this to be confusing and contradictory, unless there is a major point that I am missing. This leads me to an insufficient rating of validity until this point is clarified."

- Developer Response 4: We thank the SMP Reviewers for your consideration and comments. As noted above, KCQA is committed to addressing health inequities in the provision of dialysis care in the United States. As such, we have been on record many times indicating that the vast majority of dialysis-related measures should *not* be risk-adjusted because doing so would result in perpetuating inequities in care. This concern is appropriate to take into account when thinking about home dialysis measures. The adjusted prevalence of ESRD in Black individuals in 2019 was 78.7% higher than in the next highest group (Native Americans) and more than fourfold higher than their White counterparts;[kk] in that same year, only 7.8% of Black and 7.9% of Hispanic ESRD patients selected home dialysis as their modality, compared to 10.8% of non-Hispanic Whites.[ll] 22.3% of point prevalent and 9.3% of incident dialysis patients used dual Medicare and Medicaid in 2019.[mm] Such inequities are one of the things this measure seeks to address.

In terms of the specification questions, we provide the following responses: Again, KCQA is not a measure implementer and we are thus unable to dictate how—and if—the measure will ultimately be stratified. As a measure developer creating metrics intended for use in CMS's federal accountability programs, we do and will continue to work closely with CMS to help ensure that our measures are adopted and implemented as intended, but the practical application of implementation issues such as risk stratification is beyond our realm of influence or control. And as noted in a prior response, KCQA believes risk stratification is preferred to risk adjustment in scenarios in which there are known and/or empirically demonstrated disparities in care, the rationale being that adjustment obscures ("adjusts away") real differences in performance between different socioeconomic or sociodemographic groups. Risk stratification is typically the preferred approach in such instances, as this approach instead highlights the disparities by drawing explicit attention to performance variations across these groups.

[kk] USRDS *2021 Annual Data Report*.

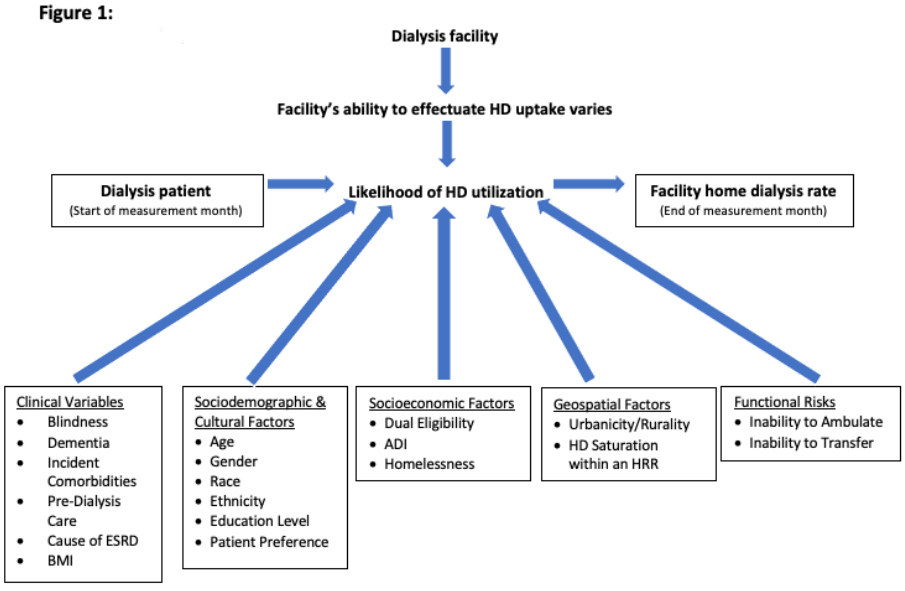
[ll] IBID.

[mm] IBID.

- **Issue 5:**
  From Reviewer 8: "There is a lack of a systematic review, consideration and testing of potential risk factors that may be salient to the given measure. The measure submitter puts forth several variables to assess, but neglects to provide an explanation as to the rationale for selecting such variables and not other variables."

    - Developer Response 7: We thank the SMP Reviewers for your consideration and comments. As described above and as noted in the submission documents, we used NQF's August 2021 report[nn] on social and functional status-related risk model guidance to develop a conceptual model illustrating the pathway between the social and functional status-related risk factors, patient clinical factors, healthcare processes, and the measured healthcare outcome. The report notes that all demographic, clinical risk factors, social and functional risks, and patient preferences related to the outcome of interest, regardless of whether they can be operationalized in available data, should be considered for inclusion in the conceptual model. In particular, NQF recommends that the following variables be evaluated when assessing the need for risk adjustment or stratification: age; gender; race/ethnicity; urbanicity/rurality; Medicare and Medicaid dual eligibility; indices of social vulnerability such as the Area Deprivation Index (ADI); and markers of functional risk such as frailty. For the KCQA Home Dialysis Rate Measure, based on our literature reviews and expert opinion from our Home Dialysis Workgroup and Steering Committee, we identified numerous risk factors believed to impact home dialysis rates:

Figure 1:



- This home dialysis conceptual model guided our selection of candidate risk factors. We identified patient sociodemographic, socioeconomic, and geospatial factors and clinical variables, including comorbidities and measures of frailty and disability. These reflect

[nn] National Quality Forum. ***Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance.***

the characteristics of the patients at the start of each measurement month and are independent of the quality of care provided. Potential clinical variables included not only incident clinical comorbidities, but also measures of pre-dialysis care, cause of ESRD, BMI, and frailty/functional status. We also considered social risk factors that may influence patients' access to home dialysis (e.g., geospatial considerations) and other barriers (e.g., homelessness) outside the control of a given dialysis facility. Variables in all of these domains have been found or are hypothesized to be associated with home dialysis utilization.[oo,pp,qq] However, the domains differ in the extent to which we expect an individual dialysis facility or group of facilities to be able to mitigate the barriers to home dialysis conferred by such variables. These differences inform their potential use as risk adjusters, since adjusting for factors that can be more easily mitigated by higher quality care is more likely to mask low-quality care.

As noted in NQF's report, however, some of these variables were ultimately eliminated during the testing phase when we were able to better identify issues with data availability, statistical issues (e.g., confounding), and model performance. For instance, we found that several necessary data elements were not consistently available across dialysis providers and/or payers, such as pre-dialysis care, incident comorbidities, homelessness, education level, and proxy markers of functional decline (i.e., inability to transfer/ambulate). Operationalizing the ADI data element was unfeasible without considerable additional burden to our testing sites, and the Workgroup and Steering Committee agreed that patient preference would be difficult to accurately and reliably capture—and might introduce considerable risk of "gaming" the measure. Ultimately, the risk variables our committees agreed are not easily mitigable (and are thus appropriate variables for risk adjustment) and are operationalizable include age, gender, race, ethnicity, and dual eligibility status.

- **Issue 6:**
    - From Reviewer 12: "I am not certain if no risk adjustment is the right choice here. Some data are hard to get, but does that mean they don't matter? They recc stratified interpretation without RA."
    - Developer Response 6: We thank the SMP Reviewers for your consideration and comments. Please see our preceding response for an explanation as to why we selected risk stratification over risk adjustment, and how we selected our risk variables. Also, consistent with our response above, individuals living with ESRD and receiving dialysis are disproportionally Black and Brown, with the adjusted prevalence of ESRD in Black individuals 78.7% higher than in the next highest group

[oo] United States Renal Data System. 2020 USRDS Annual Data Report: Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2020.

[pp] Mehrotra R et al. Racial and ethnic disparities in use of and outcomes with home dialysis in the United States. J Am Soc Nephrol. 2016;27:2123–2134.

[qq] Weiner D and Meyer K. Home dialysis in the United States: To increase utilization, address disparities. (Editorial.) Kidney Medicine. 2020;2(2):95-97.

(Native Americans) and more than fourfold higher than their White counterparts in 2019.[rr] Many of these individuals are also low-income, as evidenced by their disproportionate dual eligibility for Medicare and Medicaid.[ss] KCQA is committed to addressing health inequities in the provision of dialysis care for these patients. However**,** adjustment for social risks has remained controversial for fear of masking disparities or tacitly forgiving lower quality of care for socially marginalized patients.[tt] We note that the Minimum Standards put forth in NQF's recent risk adjustment guidance report[uu] indicate that stratification can be an appropriate alternative to risk adjustment, subject to the developer's assessment of the role of social and functional risk factors in the context of the specific intended use of the measure. We agree, and we maintain that stratification is indeed the most appropriate approach to social risk in many instances, allowing providers and other healthcare stakeholders to identify and prioritize differences in care, outcomes, and experiences across different sociodemographic groups, and to develop and implement equity-focused practices to better address disparities and understand the experiences of patients from marginalized communities.[vv] Such insights would be obscured if the same measures were instead adjusted for social risks. If this measure were to be risk-adjusted, it would result in the very patients we are trying to track and create incentives for adopting home dialysis out of the measure. To eliminate current inequities, it is not appropriate to "risk-adjust them out of the measure" and more appropriate to use a stratification approach.

[rr] USRDS *2021 Annual Data Report*.

[ss] IBID.

[tt] HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE). *Report to Congress: Social Risk Factors and Performance in Medicare's Value-Based Purchasing Programs*. March 2020.

[uu] National Quality Forum. *Developing and Testing Risk Adjustment Models for Social and Functional Status-Related Risk within Healthcare Performance Measurement: Final Technical Guidance.*

[vv] *See* Advancing Health Equity. "Using Data to Reduce Disparities and Improve Quality." https://www.solvingdisparities.org/sites/default/files/Using%20Data%20Strategy%20Overview%20Oct.%202020.pdf (accessed June 22, 2021).