August 26, 2020

National Quality Forum
1099 14th Street NW
Suite 500
Washington DC 20005

RE: NQF All-Cause Admissions and Readmissions Project, Spring 2020 Cycle

Kidney Care Partners (KCP) appreciates the opportunity to comment on the measures under consideration for endorsement in the National Quality Forum's (NQF) All-Cause Admissions and Readmissions Project, Spring 2020 Cycle. KCP is a coalition of more than 30 organizations, comprised of patient advocates, dialysis professionals, care providers, researchers, and manufacturers, dedicated to working together to improve quality of care for individuals with Chronic Kidney Disease (CKD) and End-Stage Renal Disease (ESRD). We commend NQF for undertaking this important work and offer comment on all four ESRD-related measures considered in this cycle—the currently endorsed *Standardized Hospitalization Ratio for Dialysis Facilities* (SHR, NQF 1463) and *Standardized Readmission Ratio for Dialysis Facilities* (SRR, NQF 2496), as well as the two new emergency department (ED) measures, the *Standardized ED Encounter Ratio for Dialysis Facilities* (SEDR, NQF 3565) and *Standardized Ratio for ED Encounters Occurring within 30 Days of Hospital Discharge for Dialysis Facilities (*ED30, NQF 3566).

STANDARDIZED HOSPITALIZATION AND READMISSION RATIO MEASURES

Overarching Concerns

KCP is steadfast in its belief that hospitalization and readmissions are important outcomes to measure, but our longstanding concerns about a number of reliability, validity, specification, and harmonization issues remain unaddressed for both the SHR and SRR. Below we detail several overarching concerns and make several recommendations applicable to both metrics; concerns specific to the individual measures are then addressed.

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

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

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

- **Specifications and Harmonization Issues.** Measure specifications indicate the minimum data requirement for the SHR is 5 patient-years at risk, which differs from the SRR, which uses 10 patient-years at risk. Likewise, the groupings used in the risk models for the patient age and duration of ESRD variables differ between the two measures--the SHR considers age as a continuous variable while the SRR uses three distinct age groupings, and there are four SHR groupings for ESRD duration while time on dialysis is appears to be a continuous variable in the SRR model. No justification or empirical analyses are offered to justify these differences.

NQF 1463: Standardized Hospitalization Ratio for Dialysis Facilities (SHR)
KCP has identified a number of concerns and makes recommendations specific to the SHR, as follows:

- **Reliability.** We note a reliability statistic of 0.70 is often considered as "good" reliability, though we recognize the characterization also depends on the analytic method.[1] We thus have concerns about the overall IUR for the SHR of 0.53-0.59 for 2015-2018—a sizeable decline from the 2010-2013 IUR of 0.70-0.72. This finding indicates that nearly one-half of any facility's score could be attributable to random noise and not signal. KCP believes CMS should implement the measure adjusted to yield a reliable result (reliability statistic of 0.70 or greater), consistent with how the NQF bases its evaluation of measures and more generous than the literature.[2]

  Moreover, CMS did not provide testing data stratified by facility size for the measure

[1] Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California:RAND Corporation. TR-653-NCQA, 2009.
[2] Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge, p. 13; DeVellis, RF. (2012). *Scale development: Theory and applications*. Los Angeles: Sage. pp. 109–110; Adams, JL. (2009). The reliability of provider profiling. RAND Health.

iteration currently under review by NQF because it "is not required." Yet we note that prior SHR testing results indicated very poor reliability for small facilities (then defined as facilities with fewer than 50 patients for the SHR), with IURs of 0.46-0.54 for 2010-2013 data. Only large facilities (>88 patients) had a reasonable IUR of 0.81-0.82 over the same time period. Given this history and the notable decline in the overall IUR since the measure was last reviewed by NQF, we believe it's disingenuous, at best, not to provide reliability based on facility size merely because NQF "does not require" it. KCP believes penalizing facilities for performance due to random chance is not appropriate and that it is imperative that CMS provide the most recent reliability results stratified by facility size. Absent that information, we submit that the demonstrably unreliable SHR, as currently specified, is particularly unreliable and unsuitable for use in small facilities. KCP believes the measure must specifically require a minimum sample as identified through the developer's empirical testing to prevent small facilities from having scores that are highly subject to random variability.

Finally, to assess more directly the value of SHR in identifying facilities with extreme outcomes, CMS and UM-KECC crafted an additional metric of reliability termed the Profile-IUR (PIUR).[3] Per CMS, "The PIUR indicates the presence of outliers or heavier tails among the providers, which is not captured in the IUR itself. . . . [When] there are outlier providers, even measures with a low IUR can have a relatively high PIUR and can be very useful for identifying extreme providers."[4] The PIUR for the SHR was 0.75-0.85, which CMS interprets as demonstrating that "the SHR is effective at detecting outlier facilities and statistically meaningful differences in performance scores across dialysis facilities."[5] Yet we note that NQF's Scientific Methods Panel (SMP), none of whom were familiar with the PIUR, disagrees that it is an appropriate measure of reliability for *any* measure used in the ESRD Quality Incentive Program (QIP), which are used to distinguish performance between providers falling in the *middle* of the curve to determine penalties. The SMP concluded that the IUR is and remains the appropriate measure of reliability for this purpose. KCP concurs with this position.

- **Validity.** In previous comments to CMS, KCP noted that many of the prevalent comorbidities in the final model had p-values significantly greater than 0.05. CMS responded that the large number of clinical factors in the model generates multicollinearity among covariates, likely resulting in some unexpected results in direction of coefficient sign and levels of statistical significance. However, KCP remains concerned that this strategy results in a model that will not be generalizable. In the current model, for example, asthma is associated with a higher risk of hospitalization than critical illness myopathy, and 'complete AV block' is protective while 'mood disorders' are harmful. We posit these inexplicable findings are a function of collinearity and coding idiosyncrasy. KCP supports prevalent comorbidity adjustment, but we are concerned that the proposed collection of adjusters will be less robust with each year that passes from initial model development.

  KCP also notes that validity testing yielded a c-statistic for the SHR of 0.621. We are concerned the model will not adequately discriminate performance—particularly that

[3] He K, Dahlerus C, Xia L, Li Y, Kalbfleisch JD. The profile inter-unit reliability. *Biometrics*. 2019 Oct 23. doi: 10.1111/biom.13167. [Epub ahead of print.]
[4] Kalbfleisch JD, He K, Xia L, Li Y. Does the inter-unit reliability (IUR) measure reliability? *Health Services and Outcomes Research Methodology*. 2018;18(3):215-225. Doi: 10.1007/s10742-018-0185-4.
[5] CMS. SMR measures submission materials to NQF.

smaller units might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model's goodness of fit and validity to represent meaningful differences among facilities and encourage continuous improvement of the model.

NQF 2496: Standardized Readmission Ratio for Dialysis Facilities (SRR)

KCP has significant concerns with the SRR that are consistent with the Standing Committee's recommendation against continued endorsement. Most concerning in our view is the metric's extremely poor reliability.

- **Reliability.** The overall IUR for this iteration of the SRR was found to be 0.35, such that nearly *two-thirds* of a facility's score on the measure can be attributed to noise and not signal. This value represents a sizeable decline from the 2009 IUR of 0.55. Though we recognize the characterization also depends on the analytic method, we again note a reliability statistic less than 0.50 is considered "unacceptable."[6] KCP believes CMS should implement the measure adjusted to yield a reliable result (reliability statistic of 0.70 or greater), consistent with how the NQF bases its evaluation of measures and more generous than the literature.[7]

  Here, too, CMS did not provide testing data stratified by facility size because it "is not required" by NQF. Prior SRR testing indicated notably poorer reliability for small facilities (defined in 2009 as facilities with fewer than 70 patients for the SRR), with an IUR of 0.46 compared to the overall IUR of 0.55. Given this history and the notable decline in the overall IUR since the measure was last reviewed by NQF, we believe it is imperative that CMS provide the most recent SRR reliability results stratified by facility size. Absent that information, we submit that the demonstrably unreliable SRR, as currently specified, is particularly unreliable and unsuitable for use in small facilities. As with the SHR, KCP believes the measure must specifically require a minimum sample as identified through the developer's empirical testing to prevent small facilities from having scores that are highly subject to random variability.

  Finally, CMS again uses its additional metric of reliability, the PIUR, to demonstrate that while the SRR is not adequately discriminating performance among providers, the PIUR (0.61) indicates the measure can still "be very useful for identifying extreme [outlier] providers."[8] However, KCP concurs with the NQF Scientific Methods Panel that as distinguishing outliers is not the purpose of the program, the PIUR is not an appropriate measure of reliability for any QIP measure—which are intended to distinguish performance among providers falling in the *middle* of the curve to determine penalties. The IUR is and remains the appropriate measure of reliability for measures proposed for the QIP.

- **Validity.** The Admissions/Readmissions Standing Committee did *not* pass the SRR on *Validity*, a "must pass" criterion. While in the expected directions, correlations with other outcomes measures were demonstrably weak, with a rho of 0.39 with the SHR, 0.10 with the SMR, and 0.04 with the long-term catheter rate measure. We thus concur with the Standing

[6] Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California:RAND Corporation. TR-653-NCQA, 2009.
[7] Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge, p. 13; DeVellis, RF. (2012). *Scale development: Theory and applications*. Los Angeles: Sage. pp. 109–110; Adams, JL. (2009). The reliability of provider profiling. RAND Health.
[8] Kalbfleisch JD, He K, Xia L, Li Y. Does the inter-unit reliability (IUR) measure reliability? *Health Services and Outcomes Research Methodology.* 2018;18(3):215-225. Doi: 10.1007/s10742-018-0185-4.

Committee's conclusion that the measure is not a valid representation of the quality of care provided by dialysis facilities in this regard.

KCP also notes that validity testing yielded a c-statistic for the SHR of 0.6768. We are concerned the model will not adequately discriminate performance—particularly that smaller units might look worse than reality. We believe a minimum c-statistic of 0.8 is a more appropriate indicator of the model's goodness of fit and validity to represent meaningful differences among facilities and encourage continuous improvement of the model.

EMERGENCY DEPARTMENT MEASURES

KCP recognizes the importance of assessing emergency department (ED) utilization by individuals with ESRD. We have previously communicated this in a comment letter to the Committee prior to the Standing Committee's evaluation meeting. Specifically, we have validity concerns about the exclusion of MA patients, the all-cause construct, the lack of inclusion for urgent care center visits, risk model fit, and detecting meaningful differences in performance. For example, CMS's testing results indicate that the ED30 can only distinguish differences in performance in less than 6 percent of facilities—specifically, 2.85 percent of facilities were classified as "better than expected" and 3.05 percent as "worse than expected." Simply put, the measure is unable to assess meaningful variations in performance in the overwhelming majority (94.10 percent) of facilities. Similarly, CMS's testing indicates that the SEDR can only distinguish differences in performance in approximately 5.65 percent of facilities (0.60 percent were characterized as "better than expected" and 5.05 percent as "worse than expected"); the measure was unable to assess meaningful variations in performance in 94.35 percent of facilities. Additionally, KCP re-emphasizes its concerns about the reliability of both measures.

- **NQF 3566: Standardized Ratio for ED Encounters within 30 Days of Hospital Discharge (ED30).** KCP posits the ED30 is not reliable as specified. Reliability testing for the measure yielded an overall IUR of 0.451 across all facilities, indicating that only 45 percent of the variation in a score can be attributed to between-facility differences (signal) and 55 percent to within-facility differences (noise)—by statistical convention, a "poor" degree of measure reliability. Moreover, we do not believe the Committee sufficiently addressed reliability for small facilities in particular, which is substantially lower than the overall IURs. In response to requests from NQF's Scientific Methods Panel, CMS provided IUR breakdown by tertiles of total patient-years. The IUR for those facilities falling within the lowest tertile (0-30.4 patient-years) was only 0.31. That is, *70 percent* of small facilities' ED30 scores can be attributed to noise and not signal. Again, KCP posits the measure specifications must indicate a minimum sample size or the measure overall be judged as not reliable for all facilities (the current specifications).
- **NQF 3565: Standardized ED Encounter Ratio (SEDR) for Dialysis Facilities.** Reliability testing for the SEDR yielded an overall IUR ranging from 0.62 to 0.63—a decrease from a previous version of the measure KCP reviewed 2017, then 0.64 to 0.72. We have significant concerns with a measure for which reliability has demonstrably decreased. And unlike the ED30, CMS has not provided stratification of SEDR reliability scores by facility size or tertiles, making it impossible to discern how widely reliability varies across the spectrum of facility sizes. Again, we are concerned that the reliability for small facilities may be

substantially lower than the overall IUR, as is the case with the ED30 and has been with the other standardized ratio measures.

Finally, we reiterate our concurrence with the SMP that the novel metric of reliability, termed the Profile-IUR (PIUR), is an inappropriate indicator of reliability for the QIP measures.[9] Per CMS, "The PIUR indicates the presence of outliers or heavier tails among the providers, which is not captured in the IUR itself. . . . [When] there are outlier providers, even measures with a low IUR can have a relatively high PIUR and can be very useful for identifying extreme providers."[10] We note that NQF's Scientific Methods Panel (SMP), none of whom were familiar with the PIUR, disagreed that it is an appropriate measure of reliability for measures in the ESRD Quality Incentive Program (QIP), which are used to distinguish performance among providers falling in the *middle* of the curve to determine penalties. The SMP concluded that the IUR is and remains the appropriate measure of reliability for this purpose.

KCP again thanks you for the opportunity to comment on this important work. If you have any questions, please do not hesitate to contact Lisa McGonigal, MD, MPH (lmcgon@msn.com or 203.530.9524).

Sincerely,

Kidney Care Partners:

Akebia Therapeutics, Inc.
American Kidney Fund, Inc.
American Nephrology Nurses Association
American Renal Associates
American Society of Nephrology
American Society of Pediatric Nephrology
Amgen, Inc.
Ardelyx
AstraZeneca
Atlantic Dialysis Management Services, LLC
Baxter International, Inc.
B. Braun Medical, Inc.
Cara Therapeutics, Inc.
Centers for Dialysis Care
DaVita, Inc.
Dialysis Patient Citizens, Inc.
DialyzeDirect
Fresenius Medical Care North America
Fresenius Medical Care Renal Therapies Group

[9] He K, Dahlerus C, Xia L, Li Y, Kalbfleisch JD. The profile inter-unit reliability. *Biometrics.* 2019 Oct 23. doi: 10.1111/biom.13167. [Epub ahead of print.]
[10] Kalbfleisch JD, He K, Xia L, Li Y. Does the inter-unit reliability (IUR) measure reliability? *Health Services and Outcomes Research Methodology.* 2018;18(3):215-225. Doi: 10.1007/s10742-018-0185-4.

Greenfield Health Systems
Kidney Care Council
National Renal Administrators Association
Nephrology Nursing Certification Commission
Renal Physicians Association
Renal Support Network
Rockwell Medical
Rogosin Institute
Satellite Healthcare, Inc.
U.S. Renal Care, Inc.
Vertex Pharmaceuticals
Vifor Pharma Ltd.

ATTACHMENT: MEDICARE ADVANTAGE ISSUES

Table 1: CMS Approach to MA Patients[1]

	NQF 0369: Standardized Mortality Ratio	NQF 1463: Standardized Hospitalization Ratio	NQF 2496: Standardized Readmission Ratio	NQF 2977: Standardized Fistula Rate	NQF 2978: Long-term Catheter Rate	NQF 2979: Standardized Transfusion Ratio	NQF 3565: Standardized ED Ratio (SEDR)	NQF 3566: ED within 30 Days (ED30)
Denominator Data Source(s)	Treatment History Files	Treatment History Files	Inpatient claims	CROWNWeb	CROWNWeb	Treatment History Files	Treatment History Files	Inpatient claims
Handling of MA Patients								
MA patients are included in all data sources, but their payment records are limited to inpatient claims; tracking by dialysis provider and treatment modality is available for all patients, including those with only partial or no Medicare coverage.	✓	✓	✓	✓	✓			
All time-at-risk for MA patients is included.	✓	✓						
Risk variable for proportion of MA months incorporated into risk model.		✓						
Risk variable to identify patients with MA coverage at time of index discharge.			✓					
Limited to Medicare patients enrolled in MA or who meet the criterion of being within 2 months after a month with either (a) $1200+ of Medicare-paid dialysis claims or (b) at least one Medicare inpatient claim.	✓	✓						
Binary Medicare coverage indicator: Patient-months where patient had at least 6 Medicare-covered months or 1 or more MA-covered month(s) in past 12 months. This indicator is used to determine the presence of prevalent comorbidities from Medicare claims in prior year.	✓	✓		✓				
Excludes patients with MA coverage						✓	✓	✓
Revision(s) to Identification of Prevalent Comorbidities Process to Account for MA								
Prevalent comorbidities data limited to inpatient claims.	✓	✓						
Past-year comorbidities limited to inpatient claims.			✓					
Past-year comorbidity data obtained from multiple Part A types (inpatient, home health, hospice, SNF claims) only.	✓	✓		✓				
Uses AHRQ CCS diagnosis categories to identify patient prevalent comorbidities.	✓	✓	✓					

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

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

State	N	Mean % (SD)
PR	44	44.2 (14.5)
RI	16	33.6 (18.5)
HI	31	27.8 (11.2)
OH	323	26.8 (11.4)
PA	307	25 (14.5)
AZ	121	24.6 (12.5)
CA	658	23.9 (16.6)
MN	119	23.5 (10.6)
OR	71	22.9 (15.3)
MI	211	22.4 (10.1)
TN	185	21 (8.9)
AL	176	19.8 (10.5)
FL	456	19.6 (10.3)
CO	125	18.7 (8.9)
WI	80	18.7 (11)
TX	675	18.6 (10.9)
NY	353	17.2 (7.6)
GA	296	17.2 (8.8)
NV	49	16.9 (9.7)
WV	45	16.6 (8.2)
KY	120	16.2 (6.7)
MO	165	15.2 (9.1)
NC	220	14.9 (8.6)
SC	150	14.4 (6.6)
IN	166	14.2 (8.1)
LA	175	14 (10)
NM	54	13.9 (12.2)
IL	317	13.2 (9.5)
MA	84	13.1 (11.8)
NJ	48	12.7 (4.9)
CT	179	12.7 (6.3)
VI	4	12.5 (25)
ID	43	12.1 (8.5)
UT	28	12.1 (8.9)
ME	17	11.6 (5.3)
WA	93	11 (8.5)
VA	189	10.9 (6.3)
AR	70	10.8 (6.4)
KS	57	9.3 (7.5)
IA	67	8.2 (6.6)
DC	86	7.8 (6.6)
MS	90	7.8 (5.1)
OK	21	7.7 (10.1)
NE	166	7.4 (9.7)
MD	38	7.2 (7)
ND	16	6.7 (4.9)
DE	28	6.2 (4.6)
VT	8	5.5 (2.8)
SD	27	5.3 (6)
NH	19	4.8 (3.3)
MT	15	3.6 (3.7)
AK	9	2.3 (3.2)
WY	10	2.2 (3.2)