February 2023

Technical Notes on the Dialysis Facility Quality of Patient Care Star Rating Methodology

Prepared By:

The Kidney Epidemiology and Cost Center
University of Michigan, School of Public Health
1415 Washington Heights, Suite 3645 SPH I
Ann Arbor, MI 48109-2029

CMS Contract Number: 75FCMC18D0041
Task Order Number: 75FCMC18F0001

SCHOOL OF PUBLIC HEALTH
KECC
UNIVERSITY OF MICHIGAN

Table of Contents

List of Figures

List of Tables

Glossary of Key Terms

The section provides a glossary of the key technical terminology (*italicized* throughout this report) for describing the *Star Rating* methodology.

Adjustment Factor: The ratio between the national observed event rate in the *evaluation period* and the national observed event rate in the *baseline period* for a given *standardized ratio measure*. This *adjustment factor* is applied so a standardized ratio measure is adjusted in the *evaluation period* to reflect the value it would take on in *the baseline period.*

Baseline Period: The time period, typically a calendar year, in which data are collected for the calculation of measure results used to define measure scoring criteria and *cutoff value*s for the *Star Rating* categories.

Cutoff Value: A *final score* value that determines the boundary between adjacent *Star Rating* categories. *Cutoff value*s are defined as the average of the highest score in the lower category and the lowest score in the higher category between two adjacent *Star Rating* categories.

Domain Score: A score summarizing a facility's performance on a related subset of clinical quality measures, calculated as the average of the individual *measure scores* for statistically correlated measures.

Domain Weight: The relative contribution that each *domain score* has in determining a facility's *final score* and *Star Rating*.

Evaluation Period: The time period, typically a calendar year, in which data are collected for the calculation of measure results and facility *Star Ratings*, as reported publicly on Dialysis Facility Care Compare.

Final Score: A continuous score calculated for each facility, which summarizes its performance on a set of clinical quality measures. It is a weighted average of the *domain scores* derived from the clinical quality measures included in the calculation of the *Star Rating*.

Measure Score: A standardized score derived from a specific *measure value*, which is calculated to have a mean of 0, a variance of 1, a minimum value of -2.58, and a maximum value of 2.58 (respectively corresponding to 0.5 and 99.5 percentiles of standard normal distribution).

Measure Value: The value achieved by a facility for a given clinical quality measure on its original scale, as reported on Dialysis Facility Care Compare. These values are standardized ratios or percentages.

Percentage Measure: A type of clinical quality measure with values representing the percent of patient-month observations in a facility meeting a certain criterion. *Measure value*s closer to 100% may represent higher or lower quality, depending on the definition of the criterion. For example, a higher percentage of patients utilizing catheters for hemodialysis represents lower quality of care, while a higher percentage of patients utilizing fistulas for hemodialysis represents higher quality of care.

Probit Transformation: The use of the probit function to standardize certain measure values into measure scores with mean 0, variance 1, and range -2.58-2.58. The probit function is the inverse of the cumulative distribution function, or the quantile function, of the standard normal distribution (mean 0, variance 1).

Rebaselining: The process by which, after a measure set modification, the *Star Rating* distribution is recalculated using the new measures to establish new *cutoff values*, so that the new overall facility *Star Rating* distribution is identical to the previous release's facility *Star Rating* distribution. This maintains the longitudinal continuity of the *Star Rating*s when measures are added, updated, or removed.

Resetting: The process by which the proportion of facilities in each star category is adjusted to current national performance levels and new *cutoff values* are subsequently defined. This process establishes a new *baseline period*, maintaining the discriminatory power of the *Star Rating*s for future *evaluation period*s.

Standardization: A process which transforms different *measure value*s to be on the same scale and in the same direction. After *standardization*, different measures are directly comparable.

Standardized Ratio Measure: A type of clinical quality measure with values representing the ratio of events (e.g. hospitalizations) observed in a facility to the estimated number of events expected in that facility given its characteristics and patient mix. *Measure value*s greater than one represent more observed events than expected, while *measure value*s less than one represent less observed events than expected.

Star Rating: A summary measure, on a scale from one to five, which represents a facility's overall quality of clinical care. Facilities with five stars are considered to deliver much above the national average quality of care and those with one star are considered to deliver care that is much below the national average.

Truncation: A statistical technique by which any *measure score*s exceeding a pre-specified upper or lower bound are set to equal the value of that upper or lower bound, respectively. This is done to limit the influence of extreme values on the calculation of final facility scores.

Truncated Z-Scores: A standardized score representing the number of standard deviations away from the mean, truncated at a maximum/minimum allowed value.

Executive Summary

The Centers for Medicare & Medicaid Services (CMS), through a contract with the University of Michigan Kidney Epidemiology and Cost Center (UM-KECC), developed the Dialysis Facility Quality of Patient Care *Star Rating*, hereinafter referred to as the *Star Rating*s, to rate the overall quality of care provided by dialysis facilities. The *Star Rating*s were first implemented in January 2015 to provide patients, their caregivers and other consumers with information to easily compare dialysis facility quality performance. Each facility is rated between one and five stars. Facilities with five stars are considered to deliver much above the national average quality of care and those with one star are considered to deliver care that is much below average.

Overview of the *Star Rating* Methodology

The *Star Rating*s are an aggregation of dialysis facility performance on a set of clinical quality measures reported on the Quarterly Dialysis Facility Care Compare on Medicare.gov (Dialysis Facility Care Compare). For a given *evaluation period*, facilities are rated according to fixed scoring criteria established using data from a *baseline period* of data collection. The methodology for this calculation can be summarized in the following five steps:

Step 1: Collect individual quality measure data and apply any necessary measure suppressions
Step 2: Standardize *measure value*s for data collected in the *baseline period* into *measure score*s
Step 3: Score *measure value*s for data collected in an *evaluation period* based on baseline standards
Step 4: Calculate *final score*s for facilities and define *final score* cutoffs in the *baseline period*
Step 5: Calculate *final score*s for facilities and apply *final score* cutoffs in the *evaluation period*

Survey of Patients' Experiences *Star Rating*

Results from the In-Center Hemodialysis Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Survey are reported as a separate Survey of Patients' Experiences *Star Rating* on Dialysis Facility Care Compare. The In-Center Hemodialysis Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Survey measure specifications and Survey of Patients' Experiences *Star Rating* technical notes are available at: https://ichcahps.org/

Overview of the Technical Notes

This technical report describes the methodology developed for the *Star Rating*s, highlighting updates beginning in October 2023 for Dialysis Facility Care Compare. Specifically, this technical report includes: (1) a history of updates to the *Star Rating*s (2) changes to the *Star Rating*s beginning in October 2023, (3) quality measures used in calculating the *Star Rating*s, (4) scoring of measures for inclusion in the *Star Rating*s, (5) aggregating individual *measure score*s into final facility scores, (6) translating of facility *final score*s into *Star Rating*s, and (7) processes for *rebaselining* or *resetting* the *Star Rating*s. The appendices include (a) detailed measure scoring guidelines for the individual clinical quality measures and (b) an illustrative example of the *Star Rating* calculation.

1. History of Updates to the *Star Ratings*

The original *Star Ratings* were implemented in January 2015. The technical report for the original *Star Rating* methodology is available at:

https://dialysisdata.org/sites/default/files/content/Methodology/StarRatings.pdf

A technical expert panel was convened in April 2015. The primary recommendations from this panel were (1) to establish baseline criteria for scoring dialysis facilities, to monitor changes in facility performance over time, and (2) to update the method in which certain quality measures are standardized for inclusion in the *Star Ratings*. Based on these recommendations, an update to the *Star Rating* methodology was implemented in October 2016. An updated technical report highlighting these changes to the *Star Rating* methodology is available at:

https://dialysisdata.org/sites/default/files/content/Methodology/UpdatedDFCStarRatingMethodology.pdf

A second technical expert panel convened in February 2017. The panel made recommendations on the inclusion of candidate and updated quality measures in the calculation of the *Star Ratings*. Based on these recommendations, an update to *Star Rating* methodology was implemented in October 2018. An updated technical report highlighting these changes to the *Star Rating* methodology is available at:

https://dialysisdata.org/sites/default/files/content/Methodology/Updated_DFC_Star_Rating_Methodology_for_October_2018_Release.pdf

A third technical expert panel convened in June 2019. The panel made recommendations on *resetting* the *Star Ratings* to increase the utility of the rating for patients and consumers and on weighting the relative importance that certain clinical quality measures have in determining a facility's *Star Rating*. Deliberations from this panel are described in a comprehensive summary report, available at:

https://dialysisdata.org/sites/default/files/content/ESRD_Measures/2019_ESRD_DFC_Star_Rating_TEP_Summary_Report.pdf

Due to the impact of the novel coronavirus (COVID-19) pandemic on data reporting and End-Stage Renal Disease (ESRD) dialysis outcomes, methodological updates resulting from the June 2019 TEP deliberations were postponed. A fourth technical expert panel convened in March 2022. The panel made recommendations on the inclusion of two measures of transplant waitlisting and the establishment of a new *baseline period* against which to score facility performance in light of the COVID-19 pandemic. Deliberations from this panel are described in a comprehensive summary report, available at:

https://dialysisdata.org/content/esrd-measures

2. Summary of Methodology Updates for October 2023

In response to the recommendations of the 2022 TEP meeting, changes to the methodology, beginning with the October 2023 *Star Rating* release, are listed as follows:

1. The *Star Ratings* will include two measures of transplant waitlisting: (1) the Standardized Waitlisting Ratio (SWR) and (2) the Percentage of Prevalent Patients Waitlisted (PPPW). Based on the results from factor analysis, these two measures will comprise a new, fourth domain of care (see Development of Measure Domains).

2. The 2019 TEP recommended reduction of the *weight* for the third domain, comprised of dialysis adequacy and hypercalcemia measures, to 50% the weight of the other three domains. Thus, the *domain scores* for Domains 1, 2, and 4 will each constitute 2/7 of a facility's *final score*, while the *domain score* for Domain 3 will constitute 1/7 of a facility's *final score.*

   Facilities that provide only peritoneal dialysis services do not have *measure value*s for Domain 2, which is comprised of two measures of vascular access. These facilities are rated based on a weighted average of the other three *domain score*s, such that the *domain scores* for Domains 1 and 4 will constitute 2/5 of a facility's *final score*, while the *domain score* for Domain 3 will constitute 1/5 of a facility's *final score*.

3. Beginning with the October 2023 release, the *Star Rating* baseline distribution will be *reset*, such that 10% of facilities will receive 1-Star, 20% of facilities will receive 2-Stars, 40% of facilities will receive 3-Stars, 20% of facilities will receive 4-Stars, and 10% of facilities will receive 5-Stars. As a result, data collected for the October 2023 release will constitute a new *baseline period*. Future *evaluation periods* will use the criteria set by the October 2023 release, reflecting changes in facility performance over time since the October 2023 Release.

4. As part of an Extraordinary Circumstances Exception (ECE) in light of the COVID-19 pandemic, CMS has offered regulatory relief on quality measure reporting, waiving data submission requirements for the national ESRD patient registry and quality measure reporting system. On March 27, 2020, CMS released guidance describing the scope and duration of the ECE granted under each program. Under this guidance, providers were relieved of their obligation to report clinical data for the first two quarters of 2020. Additionally for claims-based measures, claims data from March 1- June 20, would be excluded from measure calculations. Additional details can be found at:

   https://www.cms.gov/files/document/guidance-memo-exceptions-and-extensions-quality-reporting-and-value-based-purchasing-programs.pdf

5. The standardized mortality, hospitalization, readmission, and transfusion ratio measures, will be appropriately risk-adjusted to mitigate the impact of COVID-19 on dialysis facility performance. Additional methodological details can be found at:

   https://dialysisdata.org/content/dfccmethodology

3. Measures Used in Calculating the *Star Ratings*

Table 1 reports the complete list of the current quality measures used in calculating the *Star Ratings*. Measures have undergone extensive review and most have attained endorsement by the National Quality Forum (NQF). The corresponding measure names are provided for reference. The full documentation for all endorsed measures can be viewed by entering the measure into the search toolbar at: http://www.qualityforum.org/QPS/

Table 1: Quality Measures Used in Calculating the *Star Ratings*

Measure Name	Measure Abbreviation	Value Interpretation	Frequency of Update
Standardized Mortality Ratio for Dialysis Facilities	SMR	Lower is Better	Yearly
Standardized Hospitalization Ratio for Dialysis Facilities	SHR	Lower is Better	Yearly
Standardized Readmission Ratio for Dialysis Facilities	SRR	Lower is Better	Yearly
Standardized Transfusion Ratio for Dialysis Facilities	STrR	Lower is Better	Yearly
Hemodialysis Vascular Access: Standardized Fistula Rate	Fistula	Higher is Better	Quarterly
Hemodialysis Vascular Access: Long-Term Catheter Rate	Catheter	Lower is Better	Quarterly
Proportion of Patients with Hypercalcemia	Hypercalcemia	Lower is Better	Quarterly
Total Kt/V[1]	Total Kt/V	Higher is Better	Quarterly
Delivered Dose of Hemodialysis Above Minimum	Adult HD Kt/V	Higher is Better	Quarterly
Minimum spKt/V for Pediatric Hemodialysis Patients	Pediatric HD Kt/V	Higher is Better	Quarterly
Delivered Dose of Peritoneal Dialysis Above Minimum	Adult PD Kt/V	Higher is Better	Quarterly
Pediatric Peritoneal Dialysis Adequacy: Achievement of Target Kt/V	Pediatric PD Kt/V	Higher is Better	Quarterly
Percentage of Prevalent Patients Waitlisted	PPPW	Higher is Better	Quarterly
Standardized First Kidney Transplant Waitlist Ratio for Incident Dialysis Patients	SWR	Higher is Better	Yearly

[1] Four measures of dialysis adequacy are combined into a single measure (Total Kt/V) for the Star Ratings. Total Kt/V represents the percentage of dialysis patients eligible for the measure who had enough waste removed from their blood (Kt/V greater than or equal to a specified threshold). The measure is calculated by taking the average percentage of patients achieving Kt/V greater than the specified thresholds for each of four patient populations (adult hemodialysis, adult peritoneal dialysis, pediatric hemodialysis, and pediatric peritoneal dialysis), weighted by the number of patient-months of data available for each patient population. Including Total Kt/V, eight final quality measures are used to calculate the Star Ratings.

4. Measure Scoring for the *Star Ratings*

The clinical quality measures found on Care Compare have different distributions and scales. Therefore, the *measure values* for these individual measures must first be transformed in order to make them comparable in scale and direction. These transformations differ with respect to which period of data is being analyzed and what type of measure is being considered. As the current *Star Ratings* account for changes in dialysis facility performance over time, a *baseline period* is first established to set the criteria for scoring facilities. Facilities are scored using data collected during this *baseline period* to determine *cutoff values* for assigning *star ratings* in subsequent periods of data collection and evaluation (*evaluation period*).

Measure Scoring in a *Baseline Period*

Scoring facilities in an *evaluation period* against fixed thresholds established in a *baseline period* allows any facility that maintains or improves its performance on its quality measures to maintain or improve its *Star Rating*. Facilities are rated based on how their most recent performance compares to performance benchmarks in the *baseline period*. The *measure values* in the current *Star Rating* are either standardized ratios or percentages. In developing scores for the *baseline period*, different scoring methods are applied, and these methods are described below.

Percentage Measures

Percentage measures are scored using *truncated z-scores*. *Truncated z-scores* represent the number of standard deviations away from the mean, truncated at lower and upper bounds. During the *truncation* process, these measures are iteratively re-scored to ensure that the resulting distribution has mean 0 and variance 1. Highly skewed measures have the potential to have large z-scores for facilities with extreme *measure values*. These scores may exert too much influence on the *Star Ratings*. Limiting the range of scores via *truncation* ensures a facility's rating is not determined primarily by outlier performance on a single measure. A detailed description of this approach is provided in Appendix A: Detailed Measure Scoring Guidelines.

Standardized Ratio Measures

Standardized ratio measures are scored using percentile ranks and *probit transformations*. These measures are scored differently from *percentage measures*, as a unit change in a ratio measure is not equally spaced. For example, the quality difference between standardized mortality ratio *measure values* of 0.5 versus 1.0 is not the same as the quality difference between *measure values* of 1.0 versus 1.5. The former represents a two-fold difference, while the latter represents a difference in mortality that is only 1.5 times higher.

Probit scoring better accounts for these spacing differences than *truncated z-scores*, which assume equal spacing. Since the *probit transformation* maps percentile ranks for the *standardized ratio measures* to a distribution with mean 0 and variance 1, this type of scoring can also be easily combined with the *truncated z-scores* for the *percentage measures*. A detailed description of this approach is provided in Appendix A: Detailed Measure Scoring Guidelines.

Measure Scoring in an *Evaluation Period*

In order to compare current facility quality in an *evaluation period* to performance standards set in the *baseline period*, measures in the *evaluation period* are first mapped to values they would have received in the *baseline period* before scoring. The mapping and scoring processes are discussed separately for the percentage and *standardized ratio measure*s.

Percentage Measures

*Percentage measure*s in the *evaluation period* are mapped to the same score that the *measure value* would have been mapped to if it had been observed in the *baseline period*. *Measure scores* in the *evaluation period* are therefore calculated by subtracting the mean and dividing by the standard deviation of the measure in the *baseline period*. These *z-scores* are then truncated at the same values as truncated in the *baseline period* and re-standardized using the mean and the standard deviation of the *truncated z-scores* in the *baseline period*. A detailed example is given in Appendix A: Detailed Measure Scoring Guidelines.

Standardized Ratio Measures

The *standardized ratio measure*s represent observed/expected events in the *evaluation period*. We map the *standardized ratio measure*s in the *evaluation period* to the *baseline period* by multiplying them with an *adjustment factor*. The *adjustment factor*, which accounts for differences in population event rates between the *baseline period* and *evaluation period* data, is applied so that an adjusted *evaluation period* ratio value reflects the same value it would have had in the *baseline period*. The *adjustment factor* multiplied by the standardized ratio is the same for all facilities in the *evaluation period,* for that particular measure. For example, hospitalization rates were higher in 2019 than in 2018, so the expected number of events for the average facility is higher in 2019. The Standardized Hospitalization Ratio (SHR) in 2019 is then multiplied by an *adjustment factor* greater than one to calculate an adjusted SHR, so these facilities are effectively being measured by 2018, i.e., *baseline period* criteria. A detailed example is given in Appendix A: Detailed Measure Scoring Guidelines.

5. Combining *Measure Scores* into *Final Scores*

Development of Measure Domains

As some clinical quality measures are clinically more closely related than others, measures are grouped into *domains* in a data-driven manner using factor analysis.[2] Factor analysis is used to define domains where more highly correlated measures are grouped within a domain and less correlated measures are assigned to different domains. The standardized mortality, hospitalization, readmission, and transfusion ratios form one domain, the hemodialysis vascular access standardized fistula and long-term catheter rates form a second domain, and the total Kt/V and hypercalcemia measures form a third domain. Beginning with the October 2023 release, the Percentage of Prevalent Patients Waitlisted (PPPW) and Standardized First Kidney Transplant Waitlist Ratio for Incident Dialysis Patients (SWR) will form a new, fourth domain. Weighting *domain scores*, rather than *measure scores*, to calculate a facility's *final score* avoids overweighting particular measures that may represent a similar aspect of quality as other measures. Measure domains are re-established whenever *resetting* or *rebaselining* is carried out.

Calculating *Domain Scores* and *Final Scores*

Calculated *measure scores* are combined to determine a facility's *final score* as follow. A facility's *measure scores* are first averaged within each of the four domains to calculate *domain scores*. Facilities are then given a *final score* by taking a weighted average of the four *domain scores*. Beginning with the October 2023 release, Domains 1, 2, 3, and 4 will constitute 2/7, 2/7, 1/7, and 2/7 of a facility's *final score*, respectively. Facilities are eligible to receive a *final score* if they have at least one *measure value* in each domain. Note that facilities providing only peritoneal dialysis do not have *measure values* for the Hemodialysis Vascular Access measure domain. These facilities are rated based on a weighted average of the other *domain scores*, where Domains 1, 3, and 4 constitute 2/5, 1/5, and 2/5 of their *final scores*, respectively.

Missing Values

With the exception of facilities that only provide peritoneal dialysis, facilities are eligible to receive a rating if they have at least one non-missing *measure value* in each *domain*. Missing measures for eligible facilities are imputed by the mean score for that measure in the *evaluation period*.[3] This imputation method ensures one measure does not exert too much influence on the *domain score*, and in turn, does not overly influence the *final score* used to determine the *Star Rating*. For example, consider a facility which had the maximum *measure score* of 2.58 for one measure and missing values all other measures in that domain. It would not be appropriate to assume that domain should be given the maximum score of 2.58 based on the one observed measure in that domain. By imputing average scores for the other measures, we instead give the domain a submaximal above average score. The example facility is still above average for this domain, but the *domain score* will not be based solely on the one observed *measure score*, thus limiting the influence

[2] Searle, S. R., Casella, G., & McCulloch, C. E. (2009). Variance components (Vol. 391). John Wiley & Sons.
[3] Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

of that measure.

6. Translating Facility *Final Score*s into *Star Rating*s

Defining *Final Score* Cutoffs in a *Baseline Period*

Final scores in the *baseline period* are calculated, and this score distribution is used to define the *Star Rating* categories in all subsequent *evaluation period*s. Specifically, the distribution of stars in the *baseline period* is pre-specified, such that the lowest scoring 10% of facilities receive 1 star, the next 20% receive 2 stars, the next 40% receive 3 stars, the next 20% receive 4 stars, and the highest 10% receive 5 stars. *Star Rating cutoff values* are calculated as the average of the highest score in the lower category and the lowest score in the higher category between two adjacent star categories. The same *baseline period* and *cutoff value*s are used in subsequent *Star Rating* releases until a new *baseline period* and *cutoff value*s are established.

Assigning *Star Rating*s in an *Evaluation Period*

The *final score cutoff values* that are defined in the *baseline period* are used to assign ratings to facilities in each subsequent *evaluation period*. The table below shows the average *measure values* for facilities within each star category in a given *evaluation period*. Better *measure values* and *final scores* correspond to higher star categories. Further, if the population of facilities improves in their measure performance from the year in which the cutoffs are established, more facilities could receive higher ratings compared to the *baseline period*, as they are being compared to performance measured in an earlier historical time period. Note, this table uses data from the October 2020 *Star Rating* release as an *evaluation period*, with data from the October 2019 *Star Rating* release as the *baseline period* to illustrate this example; all of the periods are prior to the COVID-19 pandemic that started in spring 2020.

Table 2: Mean *Measure Value*s and *Final Scores* within each *Star Rating* Category*

Measure	★ N = 549 (7.8%)	★★ N = 1,153 (16.4%)	★★★ N = 2,923 (41.6%)	★★★★ N = 1,615 (23.0%)	★★★★★ N = 785 (11.2%)
SMR	1.19	1.10	1.01	0.92	0.84
SHR	1.25	1.10	1.00	0.89	0.79
SRR	1.20	1.10	1.00	0.91	0.83
STrR	1.65	1.23	0.97	0.77	0.59
Fistula	48.60	56.44	62.40	68.25	73.48
Catheter	24.49	16.47	12.49	9.49	7.25
Hypercalcemia	5.25	2.32	1.59	1.26	1.00
Total Kt/V	91.52	95.45	96.74	97.59	97.85
SWR	0.59	0.68	0.87	1.18	1.85
PPPW	10.73	13.24	16.50	20.28	27.58
Final Score	-0.81	-0.37	0.02	0.39	0.76

** October 2020 Star Rating release data used for the evaluation period, October 2019 release data used for the baseline period*

7. *Rebaselining* and *Resetting* the *Star Ratings*

Rebaselining

Data releases may incorporate new quality measures on different aspects of care, update current measure definitions, or retire certain measures that no longer provide actionable information in the calculation the *Star Ratings*. As the *Star Rating* measure set changes, one cannot directly compare current facility scores to the cutoffs established previously using the *baseline period* results. In order to maintain the longitudinal continuity of *Star Ratings*, the *Star Rating* release under a modified measure set will use the previous release's *Star Rating* distribution to *rebaseline* the *Star Ratings*. The current release will use the new measure specifications applied retrospectively to the prior release data to establish a new set of *final score* cutoffs. The cutoffs will reproduce the facility *Star Rating* distribution from the prior release using the prior measures and methodology. These cutoffs will be applied to all subsequent *Star Rating* releases. Thus, the prior release serves as an *evaluation period* for the former measure set and as the *baseline period* for the new measure set. For example, the October 2018 *Star Rating* release featured new, replaced, and updated measures. Therefore, it was not appropriate to directly compare this *evaluation period's* data to the original *baseline period* criteria. Instead, the new measure set was applied to the April 2018 release, and then the April 2018 *Star Rating* distribution was used to establish a new set of cutoffs for the October 2018 release.

Resetting

As the *Star Ratings* account for changes in dialysis facility quality of care over time, continued improvement may lead to progressive shifts in facility performance relative to the historical standards set in the *baseline period*. If progressive national improvement in facility performance occurs, the *Star Rating* distribution may become compressed due to overall high achievement relative to historic standards that may not reflect current care and outcomes. The *Star Ratings* then may not differentiate facility-level performance in a way that provides current, actionable information to patients and other consumers. In order to maintain the discriminatory power of the *Star Ratings*, the distribution will periodically be *reset* to update scoring cutoffs and reflect current performance. The purpose of the reset is to capture the full range of facility performance and to increase the effectiveness of the reporting program. For a release in which the *Star Rating* distribution will be *reset*, facility *final scores* for this release will be calculated using the scoring methodology for a *baseline period*. As a result, the *reset* defines new baseline scoring cutoffs for facilities to be rated in the future *evaluation periods* and sets the proportion of facilities in each star category such that 10%, 20%, 40%, 20%, and 10% of facilities would receive 1, 2, 3, 4, and 5 stars, respectively, for the *reset Star Rating* release. The October 2023 *Star Ratings* will be reset as follows. First, the January 2023 release data will be used to establish a new *baseline period*; namely, to compute *final scores* used to rate facilities so that 10%, 20%, 40%, 20%, and 10% of facilities will receive 1, 2, 3, 4, and 5 stars, respectively. From this *Star Rating* distribution, new *Star Rating cutoff values* will be determined. Future releases, starting with October 2023, will allow the *Star Rating* distribution to shift from the distribution established in the *baseline period*, reflecting longitudinal changes in facility performance based on the new established cutoffs.

Appendix A: Detailed Measure Scoring Guidelines

This section gives detailed examples of how to score *standardized ratio measure*s and *percentage measures* in the baseline and evaluation periods, respectively.

Detailed Example of Scoring the *Standardized Ratio Measure*s

Baseline Period

To calculate *probit* scores for each of the *standardized ratio measure*s in the *baseline period*, the measures are first realigned so that higher values indicate better performance. This is due to the fact that higher ratio *measure values* indicate poorer performance on these measures. We then calculate 199 percentile ranks based on their *measure values*, separately. These percentile ranks range from 0.5 to 99.5, in increments of 0.5. The percentile ranks for each measure are then realigned so that the highest value is 99.5 (representing the best possible care quality) and the lowest value is 0.5 (representing the worst possible care quality). The realigned percentile ranks are then divided by 100 and mapped to *probit* scores, where:

$$\textit{Probit}\ \text{Score} = \phi^{-1}(\text{Percentile Rank} / 100)$$

The *probit* function, ϕ^{-1}, is the inverse cumulative distribution function for the standard normal distribution. This produces the normal quantile associated with the input scaled percentile rank/100. The minimum and maximum values of the *probit* scores are determined by the minimum and maximum percentile ranks input into the *probit* function. As the *Star Rating*s use percentiles ranks ranging from 0.5 to 99.5 in increments of 0.5, the associated minimum *probit* score is: $\phi^{-1}(0.5/100) = -2.58$, and the maximum *probit* score is: $\phi^{-1}(99.5/100) = 2.58$. After *probit transformation*, the resulting *measure scores* have mean 0 and variance 1. For example, consider an observed *measure value* of 1.09 for a given *standardized ratio measure*. Here a value greater than 1 indicates slightly worse than expected performance. The measure is first realigned so that higher values indicate better performance. We then rank all facilities based on their realigned *measure values*. After ranking all facilities, the *measure value* of 1.09 is found to be in the 47.5th percentile. A *probit-transformed measure score* associated with this *measure value* is then calculated to be -0.06. The score of -0.06 is slightly below average (0) in the *baseline period*. Figure 1 below shows an example of the overall distribution of *measure values* for Standardized Mortality Ratio (SMR) on the left (where lower values are better) and an example distribution of the *probit-transformed measure scores* for Standardized Mortality Ratio (SMR) on the right (where higher scores are better).

Evaluation Period

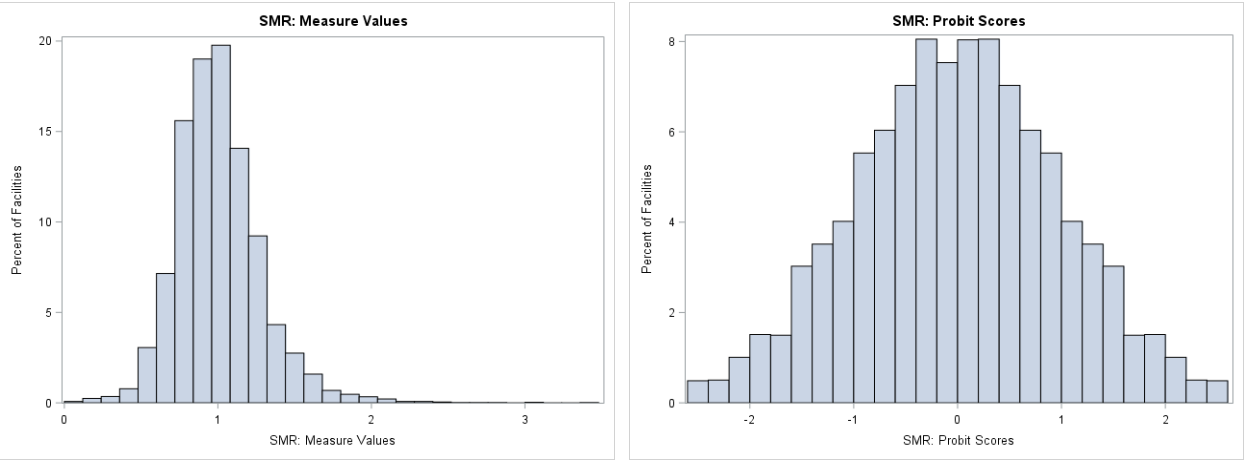
Evaluation period measure values for the *standardized ratio measure*s are first multiplied by an *adjustment factor* to calculate individual facility adjusted ratios. Each adjusted ratio is mapped to the same percentile rank that the ratio would have been mapped to if it had been observed in the *baseline period*. The cutoffs used for the percentile ranks are determined by the best *measure value* within each percentile rank in the

baseline period. Below is an example using October 2020 release data (calendar year 2019 data) as the *evaluation period*, adjusted to October 2019 release data (calendar year 2018 data) event rates, which is the *baseline period*:

$$\text{Standardized Hospitalization Ratio Adjustment} = \frac{\text{Evaluation Period Hospitalization Rate}}{\text{Baseline Period Hospitalization Rate}} = \frac{1.88}{1.86} = 1.01$$

Since hospitalization rates were higher in 2019 than in 2018, the expected number of events for the average facility is higher in 2019. By multiplying Standardized Hospitalization Ratio (SHR) in 2019 by a factor of 1.01 to calculate an adjusted SHR, these facilities are effectively being measured by 2018, i.e., *baseline period* criteria. This is interpreted as how the facility performed in the *evaluation period* relative to the typical facility in the *baseline period*. Next, to map the *standardized ratio measure values* in an *evaluation period* to the percentile ranks defined in the *baseline period*, percentile rank cutoffs must be established. There are 199 distinct percentile ranks for each ratio measure calculated using *baseline period* data. Thus, it is possible for a range of adjusted *standardized ratio measure values* to fall between the *measure values* associated with each percentile rank. The cutoffs are determined by taking the best *measure value* within each percentile rank in the *baseline period*. For any *measure value* in the *evaluation period* that falls between the percentile rank cutoffs in the *baseline period*, the *evaluation period measure value* will be "rounded up" to the higher of the two percentile rank values. A higher percentile rank indicates better performance. For example, suppose we are considering a measure for which a higher ratio indicates poorer performance on the measure. If the lowest value receiving a ratio measure percentile rank of 47.5 in the *baseline period* is 1.092 and the highest value receiving the next higher percentile rank value of 48.0 is 1.089, then the ratio measure in a future year (after applying the *adjustment factor*) of 1.090 would be given a percentile rank of 48.0. These mapped "percentile ranks" are then input into the *probit* function to determine the *measure scores* for the *evaluation period*.

Figure 1: Scoring Standardized Mortality Ratio (SMR), October 2019 *Baseline Period*

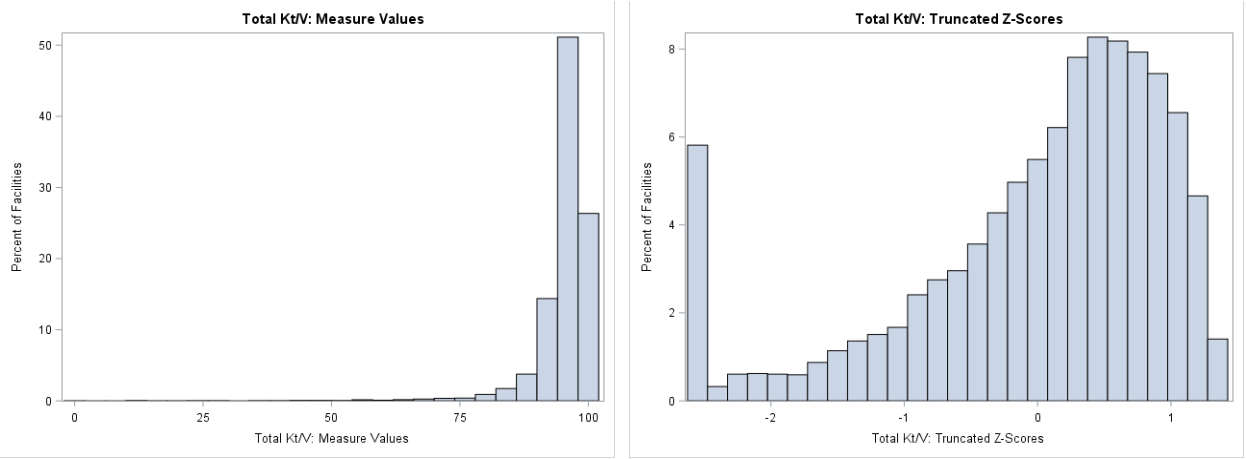


Detailed Example of Scoring *Percentage Measures*

Baseline Period

Percentage measure values in the *baseline period* are first realigned so that the highest possible value, 100%, represents the highest performance and the lowest possible value, 0%, represents the lowest performance. This is to ensure scored measures have the same directionality before they are combined. Z-scores are then calculated. All z-scored measures now have mean 0 and variance 1 at this step. The z-scores are then truncated at upper and lower bounds for each measure. Specifically, the maximum and minimum *probit* scores (± 2.58) are chosen to be the maximum and minimum values for the *truncated z-scores*. An iterative *truncation* procedure is carried out where the *measure value*s for each *percentage measure* are truncated and re-standardized until the final *truncated z-scores* have a mean of 0, a variance of 1, and a maximum possible range of -2.58 to 2.58. Thus, the *probit* scores for the *standardized ratio measure*s and the *truncated z-scores* for the *percentage measure*s have the same range of values when scoring. It should be noted that highly skewed measures may not have a maximum value of 2.58 after the *truncated z-scores* are calculated. The figure below shows an example distribution of the *measure values* for Total Kt/V on the left and an example distribution of the *truncated z-scores* for Total Kt/V on the right:

Figure 2: Scoring Total Kt/V, October 2019 *Baseline Period**



**The Total Kt/V measure is calculated by taking the average percentage of patients achieving Kt/V greater than the specified thresholds for each of four patient populations: adult hemodialysis, adult peritoneal dialysis, pediatric hemodialysis, and pediatric peritoneal dialysis, weighted by the number of patient-months of data available for each patient population.*

Evaluation Period

Here we show how *truncated z-scores* are defined in the *baseline period* and applied in an *evaluation period*. Table 3 shows how scoring is defined in an example *baseline period*. In the first row, we display Total Kt/V

summary statistics from the *baseline period*. In the second row, z-scores are obtained by subtracting each *measure value* by its mean (91.69) and dividing by its standard deviation (6.91). In the third row, initial *truncated z-scores* are calculated by truncating the z-score at a lower bound (-1.80), but no *truncation* is needed for the upper bound since the maximum score from the second row is already below 2.58. Finally, in the fourth row, the initial Total Kt/V *truncated z-score* is re-standardized by subtracting each value by its mean (0.07) and dividing by its standard deviation (0.72). Note that the *truncation* bounds in row 3 are chosen by an iterative algorithm that ensures the re-standardized measure lies within -2.58 and 2.58. The summary statistics in this table are then used to calculate the scores in the *evaluation period*.

Table 3: Defining Scores for Total Kt/V in the *Baseline Period*

Variable	Mean	SD	Minimum	Maximum
Total Kt/V *Measure Value*	91.69	6.91	12.44	100.00
Total Kt/V Z-Score	0.00	1.00	-11.47	1.20
Initial Total Kt/V Truncated Z-Score	0.07	0.72	-1.80	1.20
Final Total Kt/V Truncated Z-Score (Re-Standardized)	0.00	1.00	-2.58	1.57

Table 4 shows how scoring is defined in an example *evaluation period*. The first row reports Total Kt/V and its summary statistics. In the second row, the z-score is obtained by subtracting each Total Kt/V value by the *baseline period* mean (91.69) and dividing by the *baseline period* standard deviation (6.91) in Table 3. In the third row, initial *truncated z-scores* are formed by truncating the z-score at the lower bound (-1.80) used in the *baseline period*. Finally, in the fourth row, the initial Total Kt/V *truncated z-score* is re-standardized by subtracting each value by the mean (0.07) and dividing by the standard deviation (0.72) of the initial *truncated z-scores* in the *baseline period*. Using the mean and standard deviation from the *baseline period*, the Total Kt/V values are scored by criteria defined in the *baseline period*. Note that the mean of the re-standardized score in Table 4 is greater than 0, indicating improvement in the population average of Total Kt/V from the *baseline period*.

Table 4: Defining Scores for Total Kt/V in the *Evaluation Period*

Variable	Mean	SD	Minimum	Maximum
Total Kt/V	94.64	6.44	18.31	100
Total Kt/V "Z-Score"	0.43	0.93	-10.62	1.20
Initial Total Kt/V Truncated Z-Score	0.48	0.63	-1.80	1.20
Final Total Kt/V Truncated Z-Score (Re-Standardized)	0.58	0.89	-2.58	1.57

Development of Measure Domains

Measure domains are re-analyzed whenever the *Star Ratings* are *reset* or *rebaselined*. For example, the measure domains were re-analyzed using October 2020 release data under an expanded measure set to inform on the addition of the transplant waitlisting measures (SWR and PPPW). Results supported the creation of four measure domains to be used beginning with the October 2023 Star Rating release. Factor analysis is performed to statistically group measures that are more related or measure similar aspects of care. Results from factor analysis, called loadings, express how correlated the individual measures are with each calculated domain of care. The factor analysis loadings are presented in the table below. As shown, four *standardized ratio measures* (Standardized Hospitalization Ratio (SHR), Standardized Mortality Ratio (SMR), Standardized Readmission Ratio (SRR), and Standardized Transfusion Ratio (STrR)) formed the first domain, the Standardized Fistula Rate and Long-Term Catheter Rate measures form the second domain, the Total Kt/V and Hypercalcemia measures form the third domain, and the Percentage of Prevalent Patients Waitlisted (PPPW) and Standardized Waitlisting Ratio (SWR) form the fourth domain.

Table 5: Measure Domain Results from Factor Analysis, October 2020 Release Data[†]

Measure	Domain 1		Domain 2		Domain 3		Domain 4	
SMR	26	*	4		19		14	
SHR	63	*	13		9		5	
SRR	49	*	6		-4		-4	
STrR	43	*	6		13		6	
Fistula	9		55	*	8		17	
Catheter	12		58	*	16		5	
Hypercalcemia	5		27		33	*	-7	
Total Kt/V	18		29		36	*	-13	
SWR	3		5		-1		53	*
PPPW	5		7		-7		57	*

[†]*Values represent how correlated the individual measures are with each calculated domain of care. These correlation values range from 0 to 1 and are multiplied by 100 and rounded to the nearest whole number for display purposes.*

**The measures retained within each domain.*

Appendix B: An Example *Star Rating* Calculation

This section illustrates the current methodology, beginning with the October 2023 release. The calculation is carried out for two sample facilities: *Facility A,* which provides a combination of in-center hemodialysis, home hemodialysis, and/or peritoneal dialysis, and *Facility B*, which provides only peritoneal dialysis. This contrasts scoring between peritoneal dialysis-only facilities and all other facilities. October 2019 release data are used for the *baseline period* and October 2020 release data are used for the *evaluation period*.

Step 1: Apply Measure Suppressions to the Facilities

There are a number of valid reasons why a measure may be suppressed from a facility's *Star Rating* calculation: (1) the facility did not have enough patients or eligible observations to meet the measure-specific reporting threshold, (2) the facility was not open long enough to supply sufficient measure data, (3) the facility did not provide a particular treatment modality or service the patient population specific to the measure, or (4) the facility was granted suppression by the Centers for Medicare & Medicaid Services for another reason (e.g. the facility was affected by a natural disaster). Additional information on measure suppression can be found in the Data Dictionary for dialysis facility data on Medicare.gov:

https://data.cms.gov/provider-data/topics/dialysis-facilities

For (1) – (3), *measure value*s that are suppressed are set to missing when calculating the *Star Rating*. For (4), all measures are calculated and the *Star Rating* are released prior to any measure- or facility-specific suppression requests. If the suppression request is approved, the facility will not have those data displayed on Care Compare. Thus, a facility will still contribute a *measure value* to the calculation of the *Star Rating*, but that facility's *star rating* may be suppressed if it doesn't meet the criteria for inclusion. For this example, both *Facility A* and *Facility B* are facilities that were not suppressed.

Step 2: Define Scores in a *Baseline Period*

1. *Standardized Ratio Measure*s: Apply *probit transformation* to each measure
    a. Generate 199 percentile ranks for each measure (0.5 to 99.5)
    b. Generate *probit* scores where the score = ϕ^{-1}(percentile rank / 100)
2. *Percentage Measure*s: Apply iterative truncated Z-score algorithm to each realigned measure
    a. Let the measure of interest be m and first standardize m to get a z-score, z, by subtracting the mean of m and dividing by its standard deviation
    b. Iteratively truncate *z* to get *truncated z-scores*, *t*, and standardize t to get *measure scores, w*. This process stops, and *truncation* bounds are found, when *w* has a mean of 0, a variance of 1, and minimum and maximum possible values of at least -2.58 and at most 2.58, respectively
3. Impute eligible facility's missing values with the national average for that measure

The facility *measure values* and *measure scores* calculated for the *baseline period* are reported in Table 6. Here, *measure value* refers to the actual performance value for the clinical quality measure as reported on Care Compare. *Measure score* refers to the transformed *measure value*s for each individual metric, after applying Step 2, which are used to calculate a facility's *final score* and subsequent *Star Rating*.

Table 6: *Baseline Period Measure Value*s and Standardized *Measure Score*s for two example facilities

Measure	Facility A		Facility B	
	Measure Value	Standardized Score	Measure Value	Standardized Score
SMR	0.93	0.21	Missing*	0.00
SHR	0.56	1.75	Missing*	0.00
SRR	0.45	1.81	0.72	1.02
STRR	1.48	-0.82	Missing*	-0.01
Fistula	92.05	2.58	-	-
Catheter	3.10	1.46	-	-
Hypercalcemia	0.55	0.74	0.00	1.05
Total Kt/V	97.40	0.44	74.78	-2.58
SWR	0.00	-1.70	3.38	2.05
PPPW	20.58	0.20	43.21	2.35

** A facility missing values for SMR, SHR, and STrR was chosen to demonstrate the scoring rules and missing imputation (Step 3, Part 3)*

Step 3: Score Values in *Evaluation Period* Based on *Baseline Period* Standards

1. *Standardized Ratio Measure*s
    a. Apply *adjustment factor* to *evaluation period measure value*s.
    b. Assign *probit* scores in the *evaluation period* using bounds defined in the *baseline period*
2. *Percentage Measure*s
    a. Standardize *evaluation period measure value*s by subtracting the *baseline period mean* and dividing by *the baseline period* standard deviation
    b. Truncate standardized *measure score*s at *truncation* bounds from *baseline period*
    c. Re-standardize truncated scores by subtracting the *baseline period mean* and dividing by the *baseline period* standard deviation
3. Impute eligible facility's missing values with the national average for that measure

The example *evaluation period measure values* and standardized *measure scores* are reported in Table 7.

Table 7: *Evaluation Period Measure Value*s and Standardized *Measure Score*s for two example facilities

Measure	Facility A		Facility B	
	Measure Value	Standardized Score	Measure Value	Standardized Score
SMR	0.81	0.72	3.17	-2.58
SHR	0.76	0.88	2.21	-2.58
SRR	0.92	0.27	0.25	2.33
STrR	1.98	-1.25	Missing*	0.11
Fistula	94.20	2.58	.	-0.04
Catheter	1.49	1.72	.	-0.08
Hypercalcemia	0.26	0.90	0.34	0.86
Total Kt/V	93.78	-0.80	77.01	-2.58
SWR	0.38	-0.77	3.13	1.96
PPPW	12.21	-0.60	41.50	2.19

** A facility missing values for STrR was chosen to demonstrate missing imputation (Step 3, Part 3)*

Step 4: Define *Final Score* Cutoffs in *Baseline Period*

1. Determine which facilities will be rated in the *baseline period* based on the suppression criteria outlined in Step 1
2. Score the facility in the *baseline period*
   a. Average standardized *measure score*s within each domain to obtain *domain score*s
   b. Average *domain score*s with specified weights to obtain a *final score*
3. Define *Star Rating*s in *baseline period* based on the *Star Rating* proportions reported for the *baseline period* data
4. Define the *Star Rating* cutoffs as the average of the highest score in the lower category and the lowest score in the higher category between two adjacent *Star Rating* categories

For our example facilities, the *baseline period domain score*s and *final scores* are reported in Table 8 below; the *Star Rating* cutoffs are reported in Table 9. Note that the column Cutoff between 1-Star and 2-Stars is defined to be the average score between the highest scoring facility in the 1-Star category and the lowest scoring facility in the 2-Star category. Cutoffs for subsequent columns are defined similarly.

Table 8: *Baseline Period Domain Score*s and *Final Score*s

Measure	Facility A	Facility B
Domain 1	0.74	0.25
Domain 2	2.02	.
Domain 3	0.59	-0.77
Domain 4	-0.75	2.20
Final Score	0.66	0.83

Table 9: *Cutoff Value*s for *Star Rating* Categories

Cutoff	Cutoff between 1-Star & 2-Stars	Cutoff between 2-Stars & 3-Stars	Cutoff between 3-Stars & 4-Stars	Cutoff between 4-Stars & 5-Stars
Value	-0.57	-0.21	0.24	0.56

Step 5: Apply *Final Score* Cutoffs in *Evaluation Period*

1. Determine which facilities will be rated in the *evaluation period* based on the suppression criteria
2. Score the facility in the *evaluation period*
    a. Average standardized *measure score*s within each domain to obtain *domain score*s
    b. Average *domain score*s with specified weights to obtain a *final score*
3. Translate *final scores* to *Star Rating*s using the *Star Rating* cutoffs defined in the *baseline period*

The example *evaluation period domain scores* and *final scores* are reported in Table 10 below. Using the cutoffs reported in Table 9, both the *Facility A* and *Facility B* would be assigned 4-Stars.

Table 10: *Evaluation Period Domain Score*s & *Final Score*s

Measure	Facility A	Facility B
Domain 1	0.15	-0.68
Domain 2	2.15	.
Domain 3	0.05	-0.86
Domain 4	-0.69	2.08
Final Score	0.47	0.39
Star Rating	4-Star	4-Star