Planned Changes to the DFC Star Rating Methodology

Introduction

This report describes planned changes to the DFC Star Rating Methodology informed by DFC Star Rating Technical Expert Panel (TEP) deliberations from April through December 2015 (TEP Summary Report and Minutes available here https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/TechnicalExpertPanels.html). Included are the TEP methodology recommendations, the rationale for possible changes to the DFC Star Rating Methodology described in this document, supporting analyses, and the final methodological recommendations. The current DFC Star Rating Methodology implemented in 2015 is available here https://dialysisdata.org/sites/default/files/content/Methodology/StarRatings.pdf for reference.

Two main themes emerged as the DFC Star Ratings TEP reviewed and discussed the Star Rating Methodology. First, the TEP discussed how to incorporate changes into the Star Ratings on facility performance over time. Initially, many TEP members preferred setting absolute thresholds for individual measures as a way to observe improvement over time. As this discussion continued during subsequent TEP calls, there was also concern expressed by some that there was no clear clinical evidence, or empirical evidence in the data, to support identification of absolute (externally derived) standards. As a result, consensus moved toward use of the methodology considered most accurate by the TEP while still allowing the Star Ratings to express facility improvement over time. These issues are addressed in the following section by defining individual measure scores in a baseline year. Second, the TEP discussed the need to evaluate and account for the two highly skewed measures (hypercalcemia and Kt/V) included in the methodology. This is addressed in subsequent sections through considering binary scoring and truncated z-scores.

Baseline Year for Scoring Measures and Rating Facilities

The DFC Star Rating TEP strongly emphasized the importance of tracking a facility's improvement or decline over time. Several options were considered for setting a baseline year, including fixing the baseline year a specific number of years in advance (e.g., 2 years); setting a baseline year and then rebasing when specific criteria are achieved; and, finally, conducting an annual review of the Star Rating distribution and then determining if rebaselining is needed. These options were presented to the TEP on a Post-TEP Conference call but the TEP members did not discuss these in detail.

In the original methodology currently implemented, the score given to a specific measure value is updated each year based on the relative distribution of the current data. Star Rating cutoffs are updated using the rank order of facility final scores relative to the population of dialysis facilities in that

reporting year using percentiles. The final score value for each Star Rating cutoff is redefined each year so that the percentage of facilities in each Star Rating category is held constant from year to year.

The methodology change described here in response to TEP recommendations is that CMS will use a baseline year to establish measure score thresholds. These thresholds will remain in effect in subsequent years until rebasing is required. We would fix scores associated with measure values in the baseline year regardless of the scoring method used. Thus, performance in a future year (the reporting year) is evaluated based on scores set in the baseline year. This allows facilities to demonstrate improvement on the measures over time. Similarly, defining the Star Rating thresholds/cutoffs in a baseline year allows for a final score in a future year having the same Star Rating as the final score in the baseline year.

Baselines for standardized ratio measures

The baseline year used for the SMR, SHR, and STrR (standardized ratio measures) will be the same as all other measures used in the rating. However, SMR, SHR, STrR represent ratios (observed events/expected events) based on expected events relative to the current year being reported. When calculating the ratio measures for the year being reported, we recommend an adjustment to these ratios. We would multiply the standardized ratio in the current year being reported by the ratio of observed event rates between the current year being reported and the baseline year. An example for STrR follows for the adjustment that would be given for data collected in 2014 if the baseline year is 2013:

	Year	
	2013	**2014**
Transfusions per patient year	0.433	0.408
Ratio of transfusion rates	0.941	

Since the transfusion event rate was higher in 2013 than in 2014, the expected events for the average facility are lower in 2014. By multiplying STrR in 2014 by a factor of 0.941 to create an adjusted STrR to use in the 2014 Star Rating, we are effectively measuring these facilities by 2013 criteria. That is, an average facility in 2014 (STrR of 1) would have been above average in 2013 (STrR ≅0.941) since there were more expected events in 2013.

This would effectively allow the current year measure results being reported to be scored relative to the baseline year.

Summary of DFC Star Rating Modifications - Update of Scores Relative to Baseline Year

- DFC Star Ratings will be based on measure scores developed using the 2014 measure scoring results. Star Rating cutoffs will be fixed in the baseline year based on the current methodology for determining Star Rating cutoffs. Fixing the scoring and Star Rating cutoffs in a baseline year allows the dialysis community to evaluate changes in performance over time.

    - Adjust standardized (ratio) measures in the reporting year: Multiply the standardized ratio in the current year being reported by the ratio of observed event rates between the current year being reported and the baseline year. The baseline year for these measures will be the same as all other measures used in the rating.
- It is recommended that measure thresholds and Star Rating cutoffs will be rebased when:
    - New measures are added (and/or current measures are retired), or,
    - A TEP recommends that the baseline should be re-evaluated, or,
    - When there is shift in the Star Rating distribution that obscures differences between facilities because of shifting to extremes. Shifting to the extreme is defined as meeting one or more of the following criteria:
        - Greater than 50% of facilities achieve 4 or 5 stars or greater than 50% of facilities achieve 1 or 2 stars
        - Differences between 4 and 5 star facilities are not statistically significant for more than half of the individual measures
        - Differences between 1 and 2 star facilities are not statistically significant for more than half of the individual measures

Revision of Individual Measure Scoring

The discussion of the "number of thresholds" by the TEP motivated the evaluation of measure scoring. TEP discussion determined that the use of multiple thresholds (more continuous scoring) was generally preferable to using one threshold, particularly for metrics with a continuous distribution.

As discussed in the summary report, the TEP consensus was that skewed measures should be scored in such a way as to differentiate outlier performance. TEP members felt that it was important for the Star Ratings to include information about poor performing facilities, and to identify these outliers from the main group of facilities, particularly for highly skewed measures. For highly skewed measures, the TEP recommended considering use of a binary threshold to achieve this differentiation.

Binary Scoring

We first investigated using binary thresholds for Kt/V and hypercalcemia (the most skewed measures) while scoring the rest of the measures with the method currently implemented for the Star Rating (probit scoring). Applying binary scoring is akin to "pass/fail" scoring and ensures that the many facilities clustered at the top value on the measure are scored similarly. The few facilities that are outliers in the tail would get poor scores. This would clearly differentiate very poor performing facilities.

The TEP also acknowledged the significant difficulties in attempting to define absolute performance thresholds for individual measures. Analyses confirmed that there was insufficient empirical evidence to define single thresholds for the skewed measures using statistical methods. Additionally, there is insufficient clinical evidence and guidelines to set the individual measures thresholds. Therefore setting these cut-points would be determined arbitrarily.

Consideration of this method identified several additional concerns. For instance, the binary scoring method exaggerates differences across the threshold, while understating the differences on a single side of the threshold. Additionally, binary scoring can create small facility bias. Finally, analysis simulations showed this method might be less able to capture the "true" underlying facility quality from the "observed" facility quality (measure values that we observed).

Because of these limitations with binary scoring described above, we recommend applying z-scores (scoring using standard deviations from the mean) to score all the percentage measures included in the Star Ratings (Kt/V, hypercalcemia, catheter >90 days, fistula). For skewed measures, z-scoring methods score facilities clustered at the top similarly, while also identifying facilities in the tail. Z-scores also use more data points, providing greater precision in scores. Additionally, recognizing that some measures could become skewed through continuous improvement over time, using z-scores for all the percentage measures eliminates the need to make a decision on when to use different scoring methods. Finally, establishing a rule for truncating z-scores would eliminate the possibility that an outlier on a single measure would completely determine the Star Rating. Based on considerations described in the analysis, we propose an application of truncated z-scores for scoring all the percentage (non-standardized) measures.

The following example shows the steps for truncated z-scoring so that all scored measure values are between -2.58 and 2.58. (At the end of the document, there is a brief discussion on how to establish the truncation cutoff.)

Scoring Steps

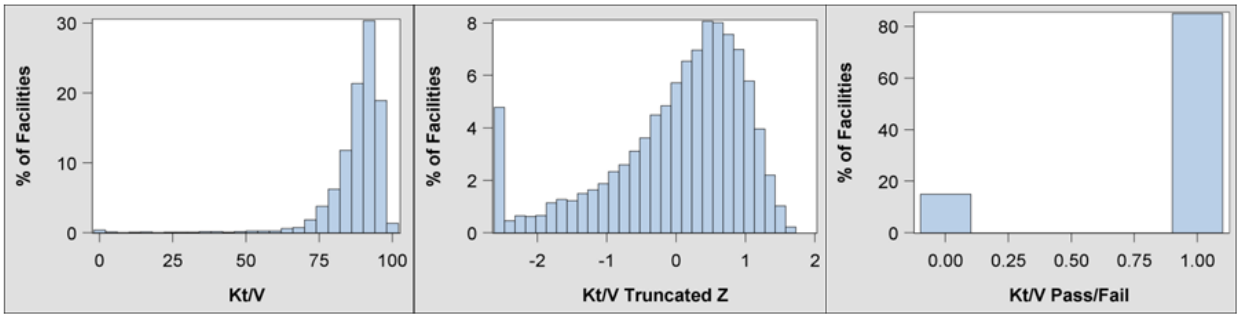
1) Percentage measures realigned so that the highest value (100) represents care much above average and the lowest value (0) represents care that is much below average.
   Purpose: Scored measures need to have same directionality before they are combined.
2) Calculate z-scores of realigned measures. All scored measures now have mean 0 and variance 1 at this step.
   Purpose: A more variable measure will have more influence on the Star Rating than a less variable measure if both measures are equally weighted. Variance stabilization therefore helps ensure measures are given equal influence if equally weighted in the rating.
3) Perform winsorization (iterative truncation) of z-scores, so that all measure scores are between -2.58 and 2.58 and still have a standardized mean and variance. Winsorisation is the transformation of statistics by limiting extreme values in the data to reduce the effect of possibly spurious/false outliers.
   Purpose: Highly skewed measures have the potential to result in large z-scores for facilities in the tail of the measure. These large scores may exert too much influence on the Star Rating. Limiting the range of the scores through winsorization ensures that Star Ratings are not determined by outlier performance on a single measure.

In addition to measure alignment, variance stabilization, and outlier adjustment, the truncated z-scoring method has other advantages. It also scores measures continuously giving the methodology more power

to detect true differences between facilities where they exist. Additionally, it is responsive to the concern about the scoring of skewed measures used in the currently implemented methodology. It ensures that many facilities clustered at one extreme value on a measure are scored similarly. It also ensures facilities in the tails of these skewed measures are recognized as having lower performance on these measures.

For example, as shown in the second panel of the figure, the truncated z-scores differentiate the facilities in the tail of the distribution, allowing many low scores. The third panel of the figure is an example of what a pass/fail score distribution might look like if 85% of facilities pass a threshold on the measure. Comparing this distribution to the original scores (first panel), we notice that using pass/fail scoring risks overstating the differences of facilities near the pass/fail threshold while understating the differences between facilities on the same side of the threshold.

Figure 1: Scoring Kt/V Examples



Recommended Scoring of Standardized Ratio Measures

For the four percentage measures (Kt/V, hypercalcemia, catheter >90 days, fistula) it may be reasonable to assume 1 percentage point higher to be associated with the same increase in quality. However, ratio measures are considered separately since the range and the distance between measure values is less straightforward than the percentage measures. For instance, ratio measures have a minimum value of 0, an expected value of 1, and a maximum value of (maximum number of events/ expected events). Thus, a difference between ratios of 0.1 and 1 may not be the same as the difference between ratios of 1 and 1.9. Since the probit function maps percentiles to a standard normal distribution with mean 0 and variance 1, this type of scoring can be easily combined with the percentage measures that are scored with truncated z-scores (which also have mean 0 and variance 1). For this reason, it is recommended to retain the probit scoring technique for ratio measures. The probit scoring will be on the same scale as the measures scored with z-scores (with mean = 0, variance = 1).

Scoring Steps

1) Calculate percentiles of the realigned measure values.
   Purpose: Percentiles/100 are the necessary input to the probit function
2) Percentiles realigned so that highest value (100) represents care much above average and the lowest value (0) represents care much below average.
   Purpose: Scored measures need to have the same directionality before they are combined.

3) Apply the probit function to the percentiles: probit score = Φ^{-1}(*percentile ÷ 100)* where Φ is the cumulative distribution function of the standard normal distribution. All scored measures now have mean 0 and variance 1 at this step.
Purpose: This function scores the standardized measures on a normal distribution based on ranks. This solution avoids trying to interpret the meaning of distance between ratio values. Additionally, the probit function puts these scores on the same scale as the scored percentage measures (scored with winsorized z-scores).

A note on determining the range of probit scores and truncated z-scores:

To create probit scores, we input a "percentile/100" into the probit function (inverse cumulative distribution function for standard normal). This produces the normal quantile associated with the input percentile. Minimum and maximum values of probit scores are determined by precision of the percentile inputed into the probit function. For instance, if we split the facilities into 199 distinct percentiles, then the minimum probit score is Φ^{-1}(1/200) = -2.58 and the maximum probit score is Φ^{-1}(1/99/200) = 2.58.

We recommend that the probit scores for standardized measures and the truncated z-scores for percentage based measures have the same range of values. We recommend first scoring standardized measure percentiles into 199 distinct groups as described above to determine the range of probit scores as ±2.58. These values are then chosen as the cutoff to truncate the z-scores.

Summary of Recommendations

Baseline Year:

- Star Ratings will be based on measure thresholds developed using the 2014 measure scoring results.
    - Fixing these thresholds in a baseline year allows the dialysis community to evaluate changes in performance over time.

A new baseline will be established when the Star Rating distribution becomes ineffective at communicating differences in outcomes between facilities due to shifting to the extreme and/or when individual measures are added or removed.

Measure Scoring:

- Apply truncated z-scores for all the percentage (non-standardized ratio) measures included in the Star Ratings.
    - Using truncated z-scores is appropriate for all the percentage measures and will handle subsequent measure shifts and skewness that could develop over time.
- Retain the probit scoring technique for the standardized (ratio) measures.
    - The probit scoring will be on the same scale as the measures scored with truncated z-scores (with mean = 0, variance = 1) allowing for the combining of all measures when calculating a final facility score.

Additional Analyses for the Star Rating Methodology Report: Comparing Current and Recommended Methodology Changes

Table 1: 2014 New method with 2013 as baseline year

Measure	★	★★	★★★	★★★★	★★★★★
# of facilities	364	874	2239	1473	922
Final Score	-0.97	-0.42	0.05	0.45	0.83
SMR	1.38	1.13	1.04	0.95	0.84
SHR	1.31	1.15	1.03	0.92	0.75
STrR	1.51	1.29	1.06	0.86	0.65
Kt/V	79.89	87.23	90.08	92.71	94.03
Hypercalcemia	4.54	3.63	2.33	1.39	0.99
Fistula	49.14	57.23	63.02	68.23	75.11
Catheter	21.22	14.96	10.39	7.45	5.36

*Preliminary calculations and results

Table 2: 2013 Facility Star Ratings using current and new scoring method with 2013 DFC data.

Current **New**

Cell Counts = # of facilities	**1**	**2**	**3**	**4**	**5**	**Total**
1	520	49	0	0	0	569 9.98
2	49	955	135	0	0	1139 19.99
3	0	136	2024	121	0	2281 40.02
4	0	0	121	943	76	1140 20.00
5	0	0	0	76	494	570 10.00
Total	569 9.98	1140 20.00	2280 40.01	1140 20.00	570 10.00	5699 100.00

*Preliminary calculations and results

Here we apply the new scoring method with CY 2013 DFC Star Rating results used as the baseline year. The correlation of final scores across the two years is similar to the correlation between the current scoring method applied with 2013 DFC data and compared to the new scoring method using 2013 data.

However, applying the new scoring method to 2014 allowed for improvement from the 2013 baseline year, where more facilities improved and received higher scores in 2014.

Table 3 shows for example, that there are 5% more facilities in both the 4 and 5 star categories. Overall, 2,072 (37%) facilities move up in the star ratings in 2014, while 679 (12%) facilities move down in 2014. Additionally, 301 (5%) facilities move up 2 or more star rating categories in 2014 while 55 (1%) facilities move down by 2 or more star rating categories in 2014.

Table 3: New method Star Rating with baseline year of 2013: Ratings for 2013 and 2014 tabulated against each other.

2013 **2014**

Cell Counts = # of facilities	**1**	**2**	**3**	**4**	**5**	**Total**
1	252	210	78	7	1	548 9.80
2	69	416	559	60	9	1113 19.91
3	12	181	1231	676	146	2246 40.17
4	1	12	239	547	326	1125 20.12
5	0	0	30	135	394	559 10.00
Total	334 5.97	819 14.65	2137 38.22	1425 25.49	876 15.67	5591 100.00

*Preliminary calculations and results