

**Overall Hospital Quality Star Rating on *Hospital Compare*
Public Input Request**

Prepared by: Yale New Haven Health Services Corporation/Center for Outcomes
Research & Evaluation (YNHHSC/CORE)

February 2019

Table of Contents

Executive Summary	4
Introduction	4
Background	4
Summary of Topics for Public Comment	5
1. Introductions	8
1.1. Background	8
1.2. Goal of Public Input Period	8
2. Instructions for Providing Feedback	10
3. February 2019 Methodology Updates	11
3.1. Background	11
3.2. Summary of Updates	12
3.3. February 2019 Methodology	13
3.4. Removal of Measures with Significant Negative Loadings	14
3.5. Use of Volume-based HAI Measure Weights	15
3.6. Update to Reporting Schedule	16
4. Potential Future Methodology Updates	17
4.1. Measure Grouping	17
4.2. Regrouping of Measures	22
4.3. Incorporating Precision of Measures	26
4.4. Period-to-Period Star Rating Shifts	29
4.5. Peer Grouping	34
4.6 Computational Update: Closed-Form Solution of LVM	36
5. Potential Long-Term Methodology Changes	37
5.1. Background	37
5.2. Explicit Approach	37
5.3. Clustering Alternative	39
5.4. Incorporation of Improvement	40
5.5. User-Customized Star Rating	41
Appendix A: Glossary of Terms	43
Appendix B: Eigenvalues and Scree Plots, Safety of Care Regrouping	44
Option 1: Retain PSI-90	44
Option 2: Switch to PSI components	45
Appendix C: Estimating Parameters in the Latent Variable Model for Star Rating Group Scores through a Closed Formed Solution	46

C.1. Overview.....	46
C.2. LVM and Log Weighted Likelihood.....	46
C.3. The EM Algorithm.....	47
C.4. Closed form maximization.....	48
C.5. Estimation.....	48

Executive Summary

Introduction

The purpose of this request for public comment is for CMS to gain feedback from a broad range of stakeholders (including technical experts, providers, patients, purchasers, and the public at large) on several potential updates to and future considerations for the methodology of the Overall Hospital Quality Star Rating on *Hospital Compare*.

CMS is asking for feedback from the public on several specific topics that address changes in hospitals' Overall Hospital Quality Star Ratings observed by some hospitals during July 2018 confidential reporting. CMS decided not to publicly report the July 2018 Overall Hospital Quality Star Ratings, in order to complete a more in-depth analysis and develop possible near-term methodology updates that are discussed in this request for public comment document. In addition, CMS would like input on its plans for some longer-term, potential future directions for the Overall Hospital Quality Star Ratings.

CMS understands that some material in this document is very technical in nature and may not be easy for all stakeholders to interpret. These select items have been included for public comment to ensure transparency with all aspects of the methodology, both technical and policy-oriented. CMS seeks guiding input from experts on these technical issues, even when they require specific knowledge of the approaches used or may not be easily communicated.

CMS believes that seeking public input on various aspects of the methodology will adhere to the project's guiding principles of wholesome transparency around major decisions and being as inclusive and responsive as possible to feedback from all stakeholders. CMS welcomes feedback from all stakeholders regarding the concepts under discussion, even if the technical content falls outside of one's area of expertise.

Background

To assess the overall performance of hospitals in the United States, CMS' Overall Hospital Quality Star Rating methodology combines results from a number of quality measures that are publicly reported on the *Hospital Compare* website. The methodology is described briefly, below, and is also explained in detail within the methodology report (Comprehensive Methodology Report (v3.0) posted on [QualityNet](#)).

- CMS first applies specified criteria to identify which measures will be used in the Overall Hospital Quality Star Rating. For example, CMS does not include measures that are reported by only a small number of hospitals, or measures where it is not clear if a higher or lower score indicates better quality (for example, payment measures in isolation are non-directional as it is not clear if spending more or less money is better or worse). Selection criteria can be found in the methodology report at the link provided above.
 - Currently, there are 57 measures on *Hospital Compare* meeting the criteria for inclusion.
- CMS then groups included measures into similar categories, called measure groups (such as Patient Experience, Mortality, or Safety of Care).
- CMS then calculates separate scores, called "measure group scores," for each hospital in each category using a method called latent variable modeling (LVM). LVM allows CMS to evaluate an underlying or "latent" aspect of quality. This latent trait is measured indirectly through the quality measures that are available and reported on *Hospital Compare*.

- Each measure within a group contributes to the measure group score. The contribution of each measure is based, roughly, on the number of patients that are accounted for by each measure, in addition to how related each of the measures are to each other in that group, in other words how consistent or correlated they are. This contribution is represented as a measure “loading,” and is computed based on the available data. A measure’s loading is the same across all hospitals.
- CMS next combines the measure group scores into one overall summary score for each hospital by calculating an average of the measure group scores. Each measure group contributes a fixed, pre-defined amount (or weight) to the overall hospital summary score. For example, Mortality and Safety of Care each account for 22% of the hospital summary score.
- Finally, CMS assigns hospitals to one of five star rating categories (from one star to five stars) based on the overall summary scores. CMS does this by comparing hospitals’ summary scores to each other and batching or “clustering” them into five groups.

Summary of Topics for Public Comment

In this public comment request, CMS is seeking feedback on several updates to this methodology that could be implemented in the near term, as well as additional topics for future exploration. These potential updates and future considerations are intended to address select stakeholder concerns about sensitivity of the Overall Hospital Quality Star Rating methodology to changes in the measures and underlying data.

Below is a summary of topics CMS is seeking feedback on regarding the Overall Hospital Quality Star Rating methodology:

- **Measure Grouping:** As individual measure specifications are updated, or measures are added or removed from programs that post data on *Hospital Compare* (including measures retired as part of the Meaningful Measure Initiative), CMS may need to reconsider the way that it groups measures and defines measure groups.
 - **CMS would like feedback from the public on a three-step approach to regrouping, which includes:**
 1. **Grouping measures based on clinical criteria;**
 2. **Using statistical tests to determine if an important latent quality trait is represented by the measures in the group; and**
 3. **Actively following measure groupings for consistency in how much each measure influences the measure group score over time.**
- **Incorporating Measure Precision:** CMS is considering changing the way that each measure’s and hospital’s scores precision are weighted within the statistical model. Right now, CMS uses, roughly, the number of patients that are part of each quality measure to determine the contribution or weight of that quality measure. This means that a hospital’s measure group score is based more on quality measures that have more of its own patients. For example, if a hospital only cares for 50 heart failure patients, but cares for thousands of pneumonia patients, the pneumonia measure would contribute more to that hospital’s group score. It also means that CMS is accounting for how *precise* each measure score is because the more patients that are measured, the less the measure score will randomly fluctuate or change. However, CMS has noticed that the amount that each measure contributes to the measure group score (the “loading”) is sometimes not balanced, and one measure may contribute much more (or have a higher loading) to the group score than another measure. CMS has also noticed that this imbalance appears to be related to both the approach used to account for measure precision and the approach used for measure grouping.

- **CMS is asking for feedback from the public regarding the importance of including measure precision in Overall Hospital Quality Star Rating, that is, whether the reliability of each measure should be accounted for in some way (currently, we use the measure’s denominator, which is often the number of patients), as well as alternative approaches to including precision that will support more balanced contributions of measures within a group.**
- **Period to Period Shifts:** Some stakeholders have expressed concern about larger-than-expected shifts in ratings from December 2017 public reporting to July 2018 confidential reporting, despite no updates to the methodology. It is important to note that some shifts in the Overall Hospital Quality Star Ratings are expected, as measure-level data and hospital-level performance change. In response, CMS looked into ways to temper the magnitude of shifts in the Overall Hospital Quality Star Ratings. One approach CMS is considering is a transition to reporting the Overall Hospital Quality Star Ratings once a year, rather than twice (as currently), so that changes in hospital ratings are more predictable based on changes in underlying measures. Other approaches to reduce shifts in this Overall Hospital Quality Star Rating could involve modifications to the methodology, such as combining data from both the current reporting period and from the closest prior reporting period (discussed below in Incorporation of Improvement).
 - **CMS would like feedback from the public regarding the benefits and drawbacks of refreshing the Overall Hospital Quality Star Rating only once a year.**
- **Peer Grouping:** Some hospital stakeholders have expressed interest in calculating and presenting the Overall Hospital Quality Star Rating results based on hospitals that “look like them,” which we refer to in this document as “peer grouping.” For example, safety-net hospitals could be grouped together to generate a star rating; teaching hospitals could be grouped together; and small/rural/Critical Access Hospitals could be grouped together. CMS could also use bed size as a peer grouping variable. CMS’s contractor (Yale-CORE) presented the option of peer grouping to a Technical Expert Panel (TEP), Provider Leadership Work Group, and Patient & Patient Advocate Work Group, and CMS has requested additional input from the public. Some stakeholders supported the concept, while others felt it would not be helpful and would be confusing, particularly to consumers and patients. Some stakeholders expressed concern with displaying two star ratings for a hospital (one overall based on all hospitals and another based on peer grouping) and believed it would be confusing for consumers and patients to interpret. In addition, there was a lack of consensus on which variables (for example bed size, safety-net, teaching hospitals, etc.) to use if peer grouping were implemented. CMS continues to receive interest from hospital stakeholders on this issue, and recently obtained updated feedback from the TEP and work groups via its contractor.
 - **CMS would like feedback from the public regarding the value of calculating the Overall Hospital Quality Star Rating based on peer groups of hospitals, and if so, how the information should be displayed. CMS would also like input on the most useful variables to use for peer grouping. CMS is also interested in feedback on whether there should be two star ratings generated – one overall rating based on all hospitals and a separate rating based on peer groupings – or just one star rating based on peer grouping.**
- **Closed Form Solution:** CMS has developed and evaluated a computational method (called the “[closed-form solution](#)”) that could replace the current approach (known as “[quadrature](#)”). The closed form solution computes substantially faster and produces the same results as quadrature, with the added advantage of modestly improved precision. This is a technical modification that CMS believes would improve the Overall Hospital Quality Star Rating statistical programming code for CMS, stakeholders, and the public.
 - **CMS would like feedback from the public regarding the benefits and drawbacks of this technical modification –replacing quadrature with the closed form solution.**

- CMS is asking for feedback on the following future considerations for the Overall Hospital Quality Star Rating methodology:
 - **Explicit Approach to Calculating Overall Hospital Quality Star Ratings:** Instead of using a statistical model to determine a hospital's measure group score, CMS could consider using a simplified, pre-defined approach that specifies or fixes the contributions or weights of each measure in a measure group. For example, CMS could decide to weight each measure within a measure group equally, or give more weight to a particular measure in a group.
 - **CMS would like feedback from the public on the advantages and disadvantages of an explicit approach to calculating Overall Hospital Quality Star Ratings, if CMS should consider this as a future direction, and feedback on how best to implement and maintain such an approach.**
 - **Alternatives to Clustering:** During initial development of the Overall Hospital Quality Star Rating, CMS considered input from the contractor's TEP and public on several approaches to assigning hospital star ratings, including approaches that involve pre-set cutoffs. In response to stakeholder feedback, CMS decided on and currently uses an approach that assigns the one- to five-star rating by comparing hospitals' overall summary scores to each other and batching or "clustering" them into five groups, based on how close the average, overall hospital summary scores are to each other. This is called "k-means clustering." Since implementation, stakeholders have expressed concern that clustering makes it difficult to predict a hospital's rating in future periods because the assignment of star ratings for any one hospital depends on the relationship of that hospital's summary score with the hospital summary scores of other hospitals.
 - **CMS would like input on whether it should consider alternatives to the current clustering method, and what should guide any future work with regard to clustering.**
 - **Incorporation of Improvement:** While the current Overall Hospital Quality Star Rating methodology captures improvement of hospitals in comparison to *other* hospitals, the methodology currently does not capture a hospital's improvement in comparison to its own prior performance. For example, CMS could average the hospital summary score from two different time periods by combining 50% of the prior reporting period with 50% of the current reporting period or 25% of the prior period with 75% of the current period.
 - **CMS would like feedback from the public on: the advantages and disadvantages of including improvement (including aligning with Dialysis Facility Compare Star Ratings); if CMS should consider this as a future direction; and feedback on how best to implement such an approach.**
 - **User-Customized Star Rating:** CMS is considering creating a user-customized Star Rating tool. Currently, the weights of each measure group are fixed (22% for each outcome group, 22% for patient experience, and 4% for each of the process measure groups), and this fixed approach may not reflect the values and preferences of patients and consumers. A user-customized approach would allow patients and consumers to express their preferences by setting the contribution or weight of each of the measure groups in the calculation of the hospital summary score and calculating star ratings for every hospital personalized to the user's values.
 - **CMS is seeking input about: whether it should consider introducing a user-customized tool; the usability, utility, and value of such a tool; as well as the benefits and drawbacks.**

1. Introductions

The Centers for Medicare & Medicaid Services (CMS) contracted with the Center for Outcomes Research and Evaluation (CORE) and the Lantana Consulting group, in collaboration with other contractors, to develop and refine the Overall Hospital Quality Star Rating on *Hospital Compare*. The goal of the Overall Hospital Quality Star Rating is to improve the usability, accessibility, and interpretability of CMS's hospital quality website, *Hospital Compare*, for patients and consumers. *Hospital Compare* is a website that includes information on over 100 quality measures from more than 4,000 hospitals. We seek public input on potential methodology updates and future topics of consideration for the Overall Hospital Quality Star Ratings.

CMS understands that some material in this document is very technical in nature and may not be easy for all stakeholders to interpret. These select items have been included for public comment given the technical nature of the methodology to ensure transparency. For these items, CMS seeks input from technical experts, even on issues that may not be easily communicated or that require specific knowledge of the approaches used. CMS is also seeking feedback on many less- or non-technical, policy-based topics as well.

CMS believes that seeking comment on both policy and technical aspects of the Overall Hospital Quality Star Rating methodology will adhere to its intent to be wholly transparent around major decisions and to be as inclusive and responsive as possible of feedback from all stakeholders (in accordance with the project's guiding principles). CMS welcomes feedback from all stakeholders regarding the concepts under discussion, even if the technical content falls outside of one's area of expertise.

1.1. Background

The primary objective of the Overall Hospital Quality Star Rating is to summarize information from the existing measures on *Hospital Compare* in a way that is useful and easy for patients and consumers to interpret. Consistent with other star ratings methodologies, each hospital is assigned a rating from one to five stars, reflecting the hospital's overall performance on selected quality measures. The Overall Hospital Quality Star Rating reflects efforts to report and improve quality from individual measures on *Hospital Compare*, and complements the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) Star Rating.

The guiding principles for the Overall Hospital Quality Star Rating methodology development are:

1. Alignment with *Hospital Compare*;
2. Transparency of methodological decisions; and
3. Being responsive to and inclusive of stakeholder input.

CMS and its contractors have been transparent and responsive to stakeholder input through convening two multi-stakeholder Technical Expert Panels (TEPs) (in 2014 and 2017), a Patient & Patient Advocate Work Group (2015), and a Provider Leadership Work Group (2017), as well as holding three public input periods, three National Provider Calls, nine listening sessions (all in 2018), and a hospital dry run. CMS and its contractors continue to maximize transparency by bringing the same topics outlined in this document to the current TEP, Patient & Patient Advocate Work Group, and Provider Leadership Work Group.

1.2. Goal of Public Input Period

CMS is seeking a wide range of stakeholder input on potential methodology updates as well as broader concepts for enhancing the Overall Hospital Quality Star Rating methodology. This request for public comment aims to

present technical and policy topics to gain feedback from the public, as well as to ensure transparency prior to implementation of any future modifications.

While we welcome public input and insight on any aspect of the Overall Hospital Quality Star Rating methodology, CMS would particularly appreciate comments on specific questions posed within this document. Please note CMS is simultaneously receiving input from its contractor's TEP and two work groups.

Specifically, this document:

1. Describes the process for providing feedback during the public input period ([Section 2](#))
2. Reviews February 2019 Methodology Updates ([Section 3](#))
3. Presents potential Overall Hospital Quality Star Rating methodology updates ([Section 4](#))
4. Presents broader topics and potential updates for future exploration ([Section 5](#))

We invite the public to comment on the Overall Hospital Quality Star Rating methodology. Feedback provided by stakeholders will inform any potential future Overall Hospital Quality Star Rating work by CMS.

2. Instructions for Providing Feedback

CMS requests that interested parties submit comments on the methodology under re-evaluation for the Overall Hospital Quality Star Rating. CMS asks that stakeholders provide comments regarding the near-term potential updates and future considerations for the Overall Hospital Quality Star Rating methodology. The public may also offer general suggestions.

- If you are providing comments on behalf of an organization, include the organization's name and contact information.
- If you are commenting as an individual, submit identifying or contact information.
- Comments are due by close of business **March 29, 2019**.
- Please do not include personal health information in your comments.
- Send your comments to cmsstarratings@yale.edu.

3. February 2019 Methodology Updates

This and following sections assume the reader is familiar with the current Overall Hospital Quality Star Rating methodology, as outlined briefly in [Section 3.3](#) below. Details of the methodology can be found in the Comprehensive Methodology Report (v3.0), available at qualitynet.org.

3.1. Background

The Overall Hospital Quality Star Ratings have been reported since July 2016 and were most recently refreshed in December 2017. While ratings were recalculated in July 2018 using updated data on *Hospital Compare* and were shared with hospitals during the Preview Period in May 2018, they were not publicly reported in July (as discussed below). Throughout this document, the “July 2018 Star Rating” refers to the unpublished results that were confidentially shared with hospitals in May 2018.

Throughout this document, CMS uses the term “refresh” to refer to the regularly scheduled update of each measure score on *Hospital Compare* reflecting the most recent available data. A measure refresh involves recalculation of hospital scores with new data and publication of the new scores on *Hospital Compare*. A measure refresh may also occasionally involve an update of measure specifications.

In July 2018, CMS observed changes in some hospitals’ ratings from December 2017 that were modest, though somewhat greater than expected given that there were no changes to the Overall Hospital Quality Star Rating methodology itself. CMS did not publicly report hospital star ratings in July 2018 to allow for time to better understand the observed changes.

To determine the cause of observed shifts, CMS first examined changes to the underlying individual measures within the Overall Hospital Quality Star Rating, then changes to measure groups (that is, measures that were added or deleted), and finally how those impacted the overall ratings. We found that there were several changes to individual measures, including data updates that occurred between December 2017 and July 2018:

- The CMS Patient Safety Indicator composite measure (PSI-90) in the Inpatient Quality Reporting (IQR) Program was updated in the following ways¹:
 - Converted to be used with ICD-10-coded claims data;
 - Refreshed with a completely new (non-overlapping) data period (from July 2014-September 2015 to October 2015-June 2017);
 - Transitioned to a new data collection period (from 15 to 21 months); and
 - Updated with new harm-based component weights;
- The severe sepsis and septic shock measure (SEP-1) was added to the Effectiveness of Care process measure group due to its introduction to *Hospital Compare*;
- The HCAHPS Pain Assessment measure (H-COMP-4) was removed from the Patient Experience measure;
- The Pneumonia 30-Day Readmission measure (READM-30-PN) was replaced with the new Pneumonia Excess Days in Acute Care measure (EDAC-30-PN) within the Readmission measure group (due to the addition of EDAC-30-PN to *Hospital Compare* and overlap between the two measures); and

¹ Department of Health and Human Services, Centers for Medicare & Medicaid Services. Fiscal Year 2017 Hospital Inpatient Prospective Payment Systems Final Rule. August 2016; <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2017-IPPS-Final-Rule-Home-Page-Items/FY2017-IPPS-Final-Rule-Data-Files.html?DLPage=1&DLEntries=10&DLSort=0&DLSortDir=ascending>. Accessed January 28, 2019.

- Many measures on *Hospital Compare* were refreshed according to their normal schedule, including many outcome measures that were last refreshed in July 2017.

Based on the measure updates listed above and findings from investigative analyses, CMS concluded that:

- Measure-level methodology updates and data refreshes can substantially impact the measure loadings, or measure contributions, a measure group score, and, in turn, a hospital's Overall Hospital Quality Star Rating.
 - Please note that loadings are not pre-determined by CMS; they are data-driven and empirically estimated each reporting period as part of the modeling procedure and so depend on the underlying data. The loadings are sensitive to two primary factors.
 1. Measures that are more consistent with each other have higher loadings. For example, if several measures all point consistently in one direction (such as, all hospitals perform well), these measures would be considered consistent or correlated and thus would receive a higher loading. Therefore, changes in the underlying measures that affect their relationship with other measures in the group can affect the measure loadings.
 2. Measures with larger denominators have higher loadings. Therefore, changes in the denominators of the underlying measures can affect the measures' loading.
- Changes in highly weighted outcome measure group scores can result in changes in the Overall Hospital Quality Star Rating that hospitals receive.

To convey these findings to stakeholders and the public, CMS hosted a series of nine listening sessions between September 6th and October 4th of 2018 with a broad array of stakeholders, including: patient advocates, Safety Net hospitals, academic and non-teaching hospitals, payer groups, and small, rural, and critical access hospitals. At the listening sessions, CMS presented analyses that demonstrated the impact of changes to individual measures on measure groups, particularly the effect of measure-level changes on the loadings within the in Safety of Care group.

CMS demonstrated that, as a result of changes in underlying performance on individual measures and measure loadings, the correlation between December 2017 and July 2018 Safety of Care group scores was both much lower than historically observed both in Safety of Care and in every other group between time periods. Furthermore, the Safety of Care group (as defined by the methodology) has a high weight in the overall summary score because it is an outcome group important to both providers and consumers; therefore, the change in underlying data translated into a greater change in hospitals' star ratings between periods than had been observed in the past.

CMS utilized the results of the listening sessions to gather stakeholder reactions and ideas regarding the sensitivity of the methodology to changes in the underlying data and has incorporated that feedback into re-evaluation analyses and this public comment document.

3.2. Summary of Updates

CMS sought to improve the consistency and predictability of the Overall Hospital Quality Star Rating by examining modifications to the methodology intended to reduce its sensitivity to substantial changes in the individual underlying measures and how those measures affect the measure groups. CMS has decided to include these updates in the February 2019 Overall Hospital Quality Star Rating release based on its analysis as well as prior feedback from stakeholders, as discussed below.

In specific response to concerns of select stakeholders regarding the July 2018 ratings, the first two methodology updates below have focused on ensuring the consistency of measure loadings within the Safety of Care measure group and improving the face validity of the methodology for February 2019.

- **Removal of measures with statistically significant negative loadings**, as these have an inverse relationship with other measures within their measure group, and was first observed within the Safety of Care measure group in July 2018. A negative loadings refers to a loading derived from the latent variable model that is negative, or below zero. Theoretically, stakeholders have suggested this could result in a hospital being penalized for performing well, although analyses have confirmed there is little to no impact. Removing statistically significant negatively loaded measures improves face validity.
 - Although measures with statistically significant negative measure loadings have little to no impact on the Overall Hospital Quality Star Ratings, CMS decided to remove measures with statistically significant negative loadings going forward to increase face validity and in response to stakeholder feedback.
- **Use of volume-based Healthcare Associated Infection (HAI) measure denominators** (device days, patient days, or number of procedures), rather than “predicted” infections, as weights for estimating the Safety of Care Latent Variable Model (LVM). This approach better represents the volume of the measure cohorts to captures the precision of the measure scores and is better aligned with the volume variables uses in other measure groups.
 - Please note that this update does not alter HAI measure score calculation, but rather utilizes a different variable, which most closely resembles volume, to weight HAI measures scores during the LVM calculation step.

In addition, CMS is considering updating the Overall Hospital Quality Star Rating reporting schedule so that ratings are refreshed once annually, rather than biannually. This would align changes in the Overall Hospital Quality Star Rating with refreshes of some individual measures and more clearly link changes in star ratings to changes in performance on the underlying measures.

These potential changes were discussed both with the TEP via a contractor and in previous public comment periods, receiving general support. Multiple stakeholders in both groups noted that the first two changes (removal of measures with statistically significant negative loadings and changes to the HAI measure denominator) are more technical and likely do not have a large practical impact, but make the results easier to interpret and more conceptually consistent. Sections [3.4-3.6](#) provide more details on these modifications and how they impact Overall Hospital Quality Star Rating.

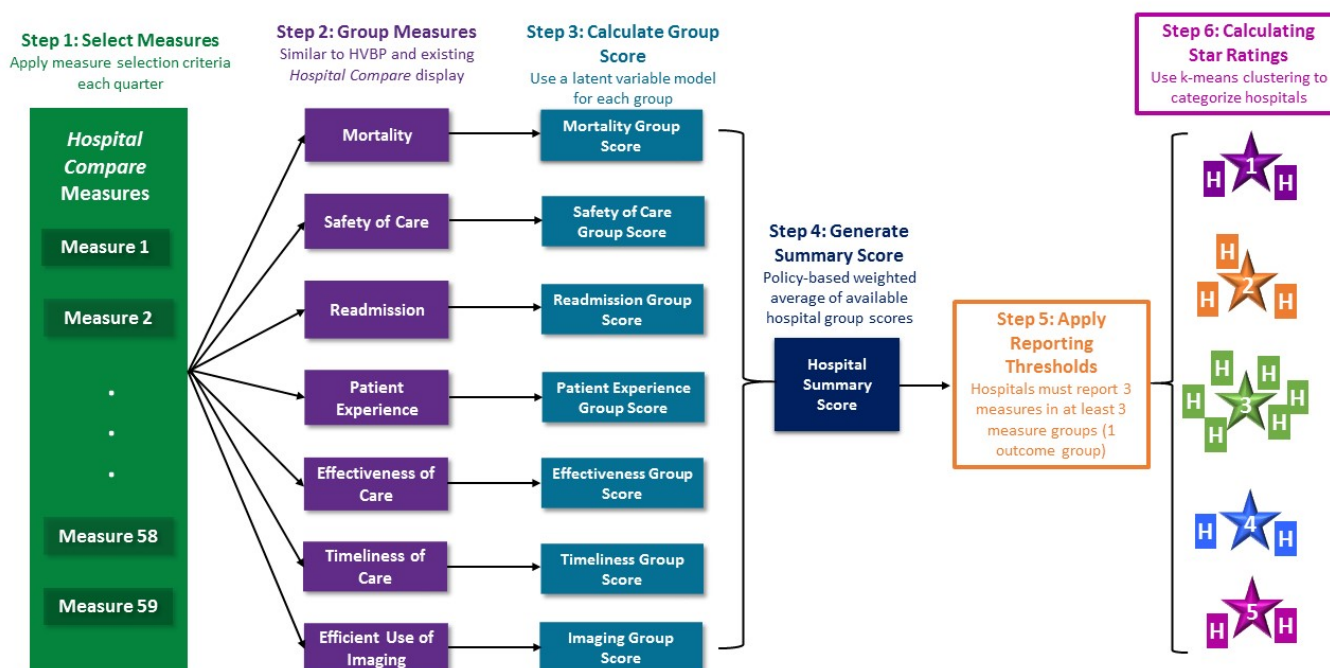
3.3. February 2019 Methodology

The methodology continues to be described by the six steps below and pictured within [Figure 1](#), with the two above methodology changes occurring in Step 3. Please refer to the February 2019 Quarterly Updates and Specifications Report on [QualityNet](#) for more details on the February 2019 methodology.

1. Selection and standardization of measures for inclusion in the Overall Hospital Quality Star Rating;
2. Assignment of measures to measure groups;
3. Calculation of latent variable model group scores;
 - a. Methodology Update: Volume-based HAI denominators (device days, patient days, or number of procedures) used for weighting to better account for measure sampling variation;
 - b. Methodology Update: Measures with statistically significant negative measure loadings are excluded from the final calculation of the latent variable model;
4. Calculation of hospital summary scores as a weighted average of measure group scores;

5. Application of public reporting thresholds for receiving a Star Rating; and
6. Application of clustering algorithm to translate a summary score into a Star Rating.

Figure 1: The Six Steps of the Current Overall Hospital Quality Star Rating Methodology



3.4. Removal of Measures with Significant Negative Loadings

As noted above, latent variable models (LVMs) are used to calculate a group score for each hospital in each measure group. Each model assumes that there is an underlying “quality signal” (known as a “latent variable”) for that measure group which represents, for each hospital, an unobserved factor which influences the measures in that group. The model calculates a ‘loading’ for each measure that represents how much the measure correlates with the latent variable; measures that are more correlated with other measures in the group receive higher loadings.

It is possible for some measures to have a negative loading, indicating that they are inversely correlated with other measures and the group score. If the loading is not statistically significant (that is, if the confidence interval includes zero) then this may just be noise; if it is significant, however, it indicates that the inverse relationship may be meaningful in the context of that group of measures (although the possibility of noise cannot be fully ruled out). For example, in July 2018, one measure (HAI-4) had a loading of -0.01, which was statistically significant less than zero. This means, this measure had a statistically significant negative loading.

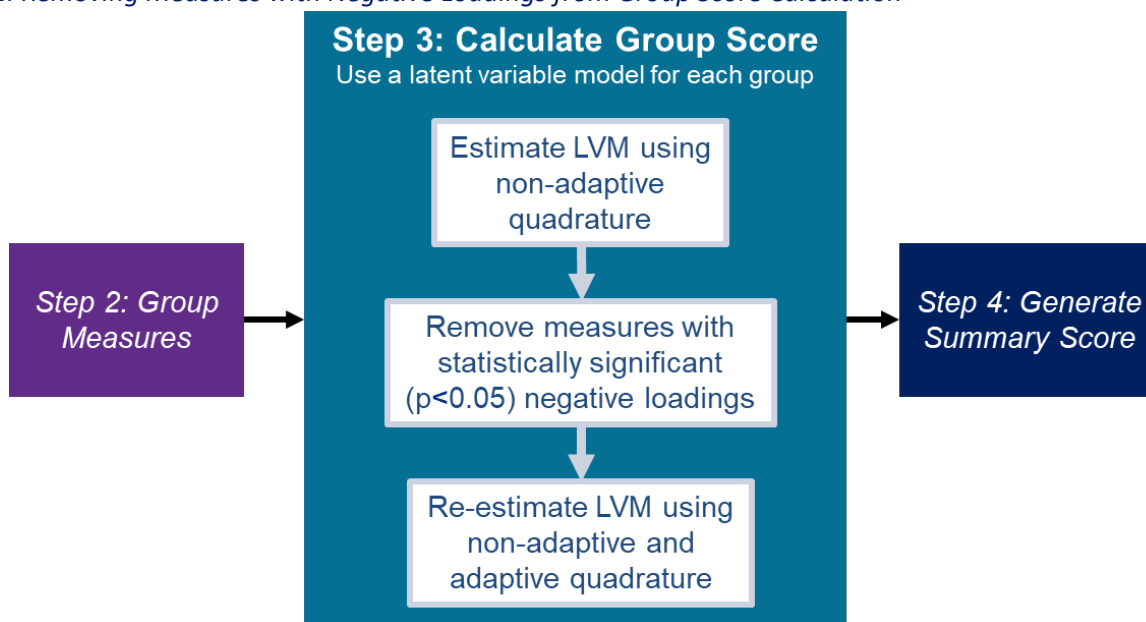
It should be noted that all previously observed negative loadings of Overall Hospital Quality Star Rating measures have been small in magnitude relative to measures with positive loadings, and have not had a substantial effect on group scores regardless of statistical significance. Furthermore, no measures had significant negative loadings before July 2018 (when a single measure had a significant loading of -0.04), nor did any measures in February 2019.

Based on feedback from stakeholders that negative measure loadings are counterintuitive and potentially inconsistent with policy communications, CMS has decided to remove measures with statistically significant

negative loadings beginning in February 2019. To date, this change only would have affected one measure in July 2018. Importantly, the Overall Hospital Quality Star Rating calculation methodology itself has been updated to automatically remove such measures in any future period, to ensure that this solution is consistent and data-driven for all future releases.

Measures with significant negative loadings will be removed as part of Step 3 of the methodology shown in [Figure 1](#), “Calculate Group Scores using LVM.” CMS will estimate each measure group LVM using non-adaptive quadrature, which produces an approximate solution. After this step, any measures with statistically significant negative loadings are removed. The model is then re-estimated in a two-step process using non-adaptive quadrature to re-estimate an approximate solution followed by adaptive quadrature to refine the accuracy of results. This is illustrated in [Figure 2](#) below. The final estimation of all groups is obtained using the adaptive quadrature step whether or not there was a measure with a significant negative loading.

Figure 2: Removing Measures with Negative Loadings from Group Score Calculation



3.5. Use of Volume-based HAI Measure Weights

All LVMs are estimated using weights to account for differences in sample size and measure precision across hospitals. This allows measures for which we have more precise estimates to contribute more to the model than measures for which we have less precise or reliable estimates. For most measures, the weights are the number of patients or admissions included in the measure denominator. However, not all measures are reported with the number of included patients.

The six HAI measures in the Safety of Care group are reported as standardized infection ratios (SIRs), defined as the number of observed infections (measure score numerator) over the number of predicted infections (measure score denominator). Predicted infections for each measure are based on statistical models of each patient’s likelihood of infection in an eligible health care encounter, summed across the eligible cohort of patients.

Previously, predicted infections were used to weight the HAI measures in the LVM. However, each HAI measure also has an alternative denominator reflecting the underlying volume (such as device days, number of procedures, or patient days) as listed in [Table 1](#) below. These data were not originally reported publicly but have recently become available on *Hospital Compare* and therefore available for use in Overall Hospital Quality Star Rating.

Table 1: HAI Measure Details

Measure (within the CMS IQR Program)	Cohort	Outcome	Volume-based Denominator
HAI-1	Patients with central line	Count of CLABSI events	Device days (central line)
HAI-2	Patients with urinary catheter	Count of CAUTI events	Device days (catheter)
HAI-3	Patients receiving colon surgery	Count of SSI events	Number of procedures
HAI-4	Patients receiving abdominal hysterectomy	Count of SSI events	Number of procedures
HAI-5	All patients	Count of MRSA infections	Total patient days
HAI-6	All patients	Count of C. diff infections	Total patient days

These volume-based weights are more consistent with those of other measures (for example, mortality measures which use the number of index admissions), and better capture differences in measure precision between hospitals. Using volume-based weights for HAI measures improved the consistency of loadings in the Safety of Care group based on historical data.

3.6. Update to Reporting Schedule

Originally, CMS intended to refresh Overall Hospital Quality Star Ratings every quarter along with some of the individual measures on *Hospital Compare*. Many of the heavily weighted outcome measures, however, are refreshed annually at the same time (for example, July every year), which results in substantial changes to the dataset on *Hospital Compare* one time a year. In parallel, CMS transitioned Overall Hospital Quality Star Ratings to a biannual schedule.

However, some stakeholders have expressed concern that biannual Star Rating refreshes may not be well aligned with the annual refresh of most underlying outcome measures. As a result, changes in rating for hospitals near cutoffs may be very sensitive to modest changes in individual measure scores outside the major annual refresh schedule. Therefore, CMS is considering changing the publication of Overall Hospital Quality Star Ratings to an annual schedule. Under this potential plan, star ratings would be published once a year using data that hospitals previewed in the previous quarter. This would ensure better alignment between measure scores and the Overall Hospital Quality Star Rating refresh schedules and is intended to make hospitals' changes in rating more predictable based on their performance on individual measures. CMS is seeking public input on an annual Overall Hospital Quality Star Rating publication schedule.

4. Potential Future Methodology Updates

CMS is committed to building upon and improving the existing Overall Hospital Quality Star Rating methodology through continuous evaluation and refinement. CMS has received the following input about the methodology:

- Recent stakeholder concerns that the methodology is overly sensitive to subtle changes in the underlying data; and
- Interest from select stakeholders in a methodology that is:
 - More consistent between periods (select stakeholders raised concerns that shifts between periods were greater than expected for some hospitals and that these shifts can be challenging to interpret or explain given modest changes in individual scores.);
 - More balanced in emphasis on individual measures; and
 - More predictable for future periods.

In addition to the updates described in [Section 3](#) for February 2019, CMS is considering several methodology updates that could be designed, evaluated, and presented to a wide range of stakeholders for feedback in time for potential near-term implementation. These potential updates are summarized here and are further discussed in the subsequent sections.

- **Measure Grouping:** CMS is considering updating the criteria used to define measure groups and evaluated the possibility of regrouping some measures; notably, by partitioning Safety of Care into two separate groups, each with its own LVM.
- **Measure Precision:** CMS is considering other methods to account for measure precision, other than denominator weighting. CMS identified two alternative approaches: removal of denominator weighting altogether or weighting based on the precision of measure scores (for measures with that information).
- **Period-to-Period Shifts:** CMS is considering mitigating between-period shifts by using a summary score based on performance from both the current and previous period.
 - CMS presents sample analyses in [Section 4.4.2](#) incorporating data from the current period and the period six months prior to illustrate the concept.

4.1. Measure Grouping

4.1.1. Background

Originally, the seven Overall Hospital Quality Star Rating measure groups (Mortality, Readmission, Safety of Care, Patient Experience, Process Effectiveness, Timeliness of Care, and Efficiency of Medical Imaging) were created based on clinical coherence, measure type, and underlying latent traits of quality.² These seven groups were vetted through multiple stakeholder groups and public input.

The objective of the LVM approach is to capture one underlying construct of healthcare quality for each hospital and in each measure group by estimating a group score reflecting common performance across the group's measures. LVM assumes each measure reflects information about an underlying, unobserved dimension of quality. In developing Overall Hospital Quality Star Rating, CMS used factor analysis to assess the degree to which

² Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (YNHHSC/CORE). Overall Hospital Quality Star Ratings on Hospital Compare Methodology Report (v3.0). December 2017; <https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228775957165>. Accessed January 28, 2019.

a dominant underlying factor exists for each measure group. [Factor analysis](#) is a widely used statistical analysis that investigates the relationship between measures or concepts.

Given recent and upcoming changes to measures reported on *Hospital Compare*, such as the retirement of many measures as part of the Meaningful Measures Initiative³, CMS believes this is an opportune time to examine and improve Overall Hospital Quality Star Rating grouping criteria. CMS, therefore, is asking for public input on two possible options, summarized here and discussed in more detail below, for improving the grouping of measures in the Overall Hospital Quality Star Rating:

- Creation of additional criteria to evaluate measure groups; and
- Examining alternative measure groupings (“regrouping”) that may improve model performance and actionability.

4.1.2. Criteria for Evaluating Measure Groups

In order to create a more robust approach to grouping that can accommodate changes in the underlying measure set as measures change and hospital scores evolve, CMS is considering a more explicit approach for composing measure groups. This includes both a clinical rationale and empirical criteria for checking the existence of a dominant quality factor. The potential approach to regrouping is based on three criteria:

Criterion 1. Initial Clinical Grouping: After applying existing measure exclusion criteria, measures would be initially grouped based on clinical coherence. In the near term, Overall Hospital Quality Star Ratings would retain the clinical focus of current measure groups until the composition of the available measures changes.

Criterion 2. Confirmatory Factor Analysis: Each clinical group would be assessed using factor analysis to ensure that a dominant underlying quality measure is present (one dominant factor), using several empirical tools (detailed below):

- a. Ratio of the first to second [eigenvalue](#) (In factor analysis, an “eigenvalue” is the amount of variation across measure scores captured by each one of a set of underlying factors.), compared to the ratio of the second eigenvalue to any other.
- b. Qualitative assessment of shape of the eigenvalue scree plot.

Criterion 3. Ongoing Active Monitoring: Measure groups would be periodically re-assessed to confirm that measure loadings are balanced within each group and relatively consistent over time, in order to ensure the usability of information for patients and providers.

Criterion 1: Initial Clinical Grouping

This part will be explored more in the next section ([Section 4.2](#)). CMS will refer to any changes made in this step as “regrouping” throughout this document.

Criterion 2: Confirmatory Factor Analysis

Factor analysis was evaluated during the initial creation of measure groups but has not been re-assessed with every subsequent Overall Hospital Quality Star Rating publication. Factor analysis is a way to examine if a group of measures can be explained by a single common underlying factor. As hospitals and measures evolve, the

³ Department of Health and Human Services, Centers for Medicare & Medicaid Services. Fiscal Year 2018 Hospital Inpatient Prospective Payment Systems Final Rule. August 2017; <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2018-IPPS-Final-Rule-Home-Page-Items/FY2018-IPPS-Final-Rule-Data-Files.html>. Accessed January 28, 2019.

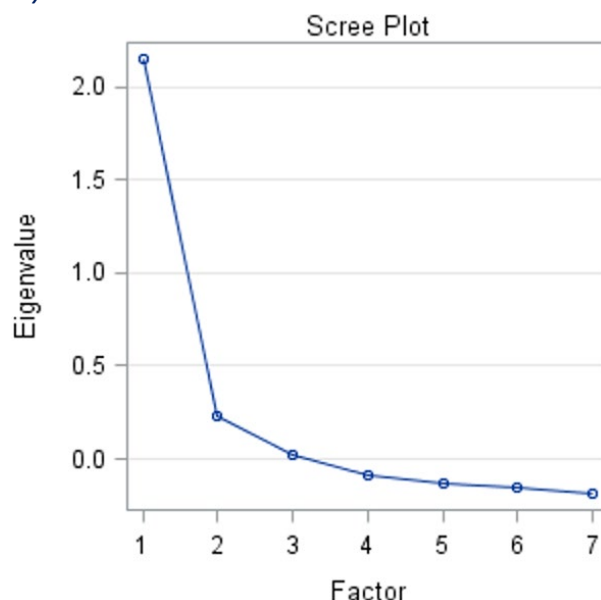
underlying relationships between measures will change; including these empiric criteria going forward would allow CMS to confirm that the existing groups are adequately capturing relationships between measures.

In accordance with scientific literature, CMS decided to use the criterion of “ratio of first to second eigenvalue in weighted factor analysis greater than 3”^{4,5} as a guide for the dominance of one factor. A larger first eigenvalue is desired for LVM because it indicates that a single underlying factor is strongly associated with all measures in the group. (Please note that the weights used in factor analysis are hospital-specific, not measure-specific like in the LVM, so using factor analysis with or without weights is for exploratory purpose only.)

In addition to the guidance of an eigenvalue ratio greater than 3, CMS would qualitatively evaluate the Scree plot generated by weighted factor analysis: in a group with one strong factor, there should be a sharp turn in the plot; that is, the first point should lie substantially out from the others. Please note that these criteria are intended to serve as guidance rather than hard cut points.

[Figure 3](#) below shows the Scree plot for Mortality in July 2018, as an example of a well-constructed group with a strong underlying factor. The first eigenvalue is 2.14 and the second is 0.225, a ratio of 9.55 (much greater than 3). Visually this can be seen in the Scree plot, where the first point is much greater than the remaining points.

Figure 3: Scree Plot, Mortality, July 2018

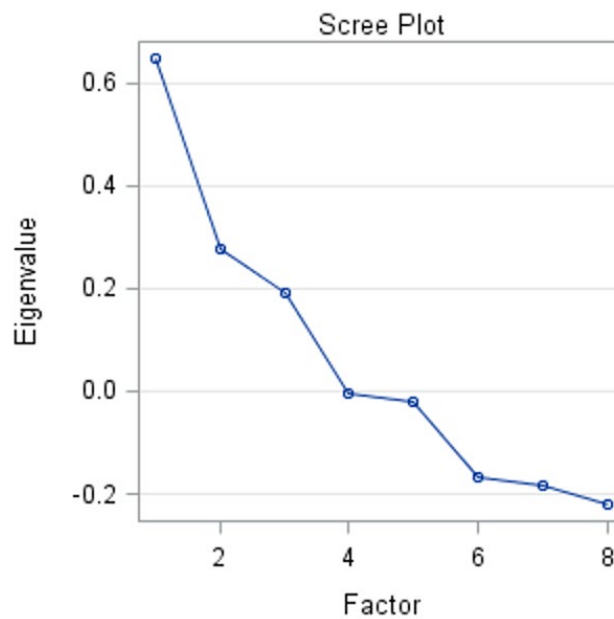


In contrast, the Safety of Care group, while meeting statistical criteria for a dominant factor, is relatively weaker in construction than mortality. This can be seen in the Scree plot for the Safety of Care group for July 2018, shown below ([Figure 4](#)). The first eigenvalue for Safety of Care is 0.647 and the second eigenvalue is 0.276, a ratio of 2.34, which is below the ideal value of 3. Furthermore, in contrast to the Mortality Scree Plot shown above, the Scree plot for Safety of Care does not have a prominent turning point at the second eigenvalue. These criteria are meant to act as guidance and not as explicit cut-offs; therefore, CMS is using this analysis to inform additional re-evaluation of the Safety of Care group.

⁴ Gorsuch RL (1983). Factor analysis: second edition. Hillsdale, NJ: Lawrence Erlbaum

⁵ Lord, F. M. (1980). Applications of item response theory to practical testing programs. Mahwah, NJ: Lawrence Erlbaum Associates.

Figure 4: Scree Plot, Safety of Care, July 2018

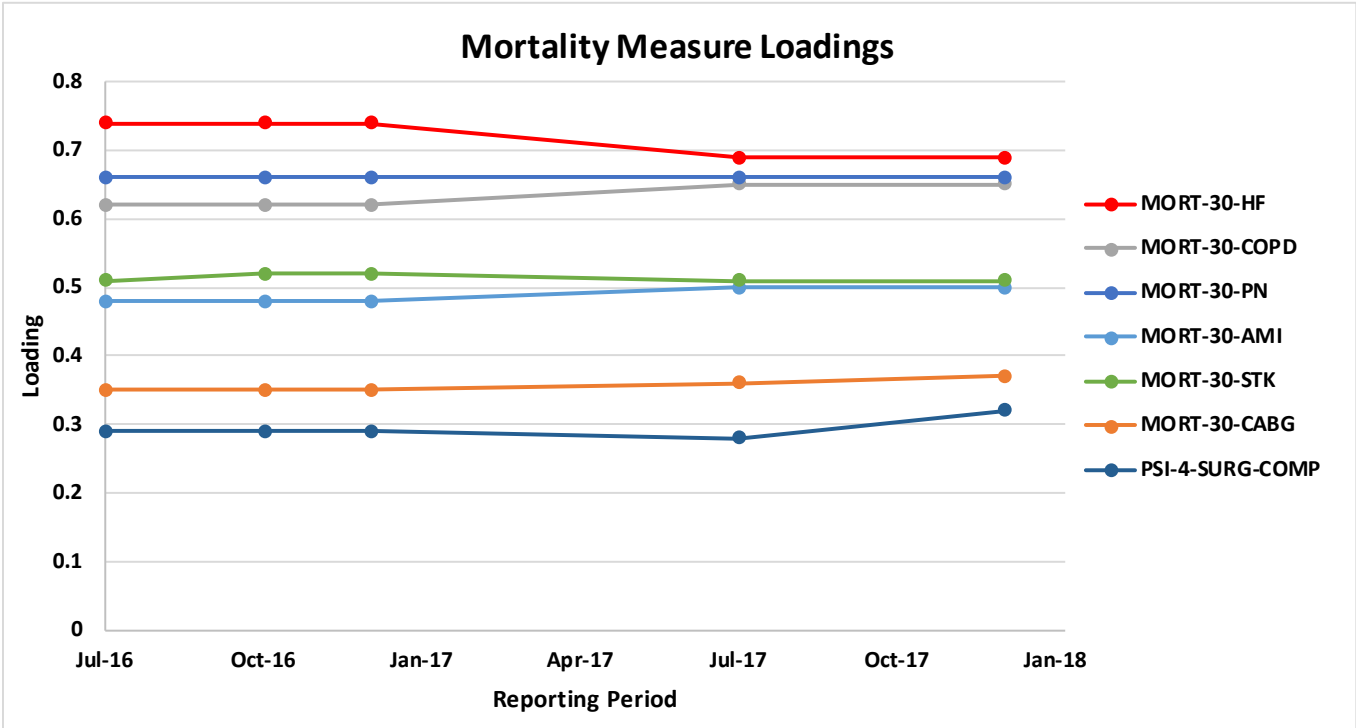


Criterion 3: Ongoing Monitoring of Loadings for Balance, Consistency, and Predictability

This part of the measure group assessment process is a qualitative assessment of the results of LVM in each group. While LVM is an empirical method designed to best summarize the available information, some stakeholders have given feedback that it may be overly sensitive to subtle changes in the underlying data. This criterion is intended to ensure that loadings are reasonably balanced within periods and reasonably consistent between periods; as a result, hospitals would see more predictability in their rating based on their measure score performance.

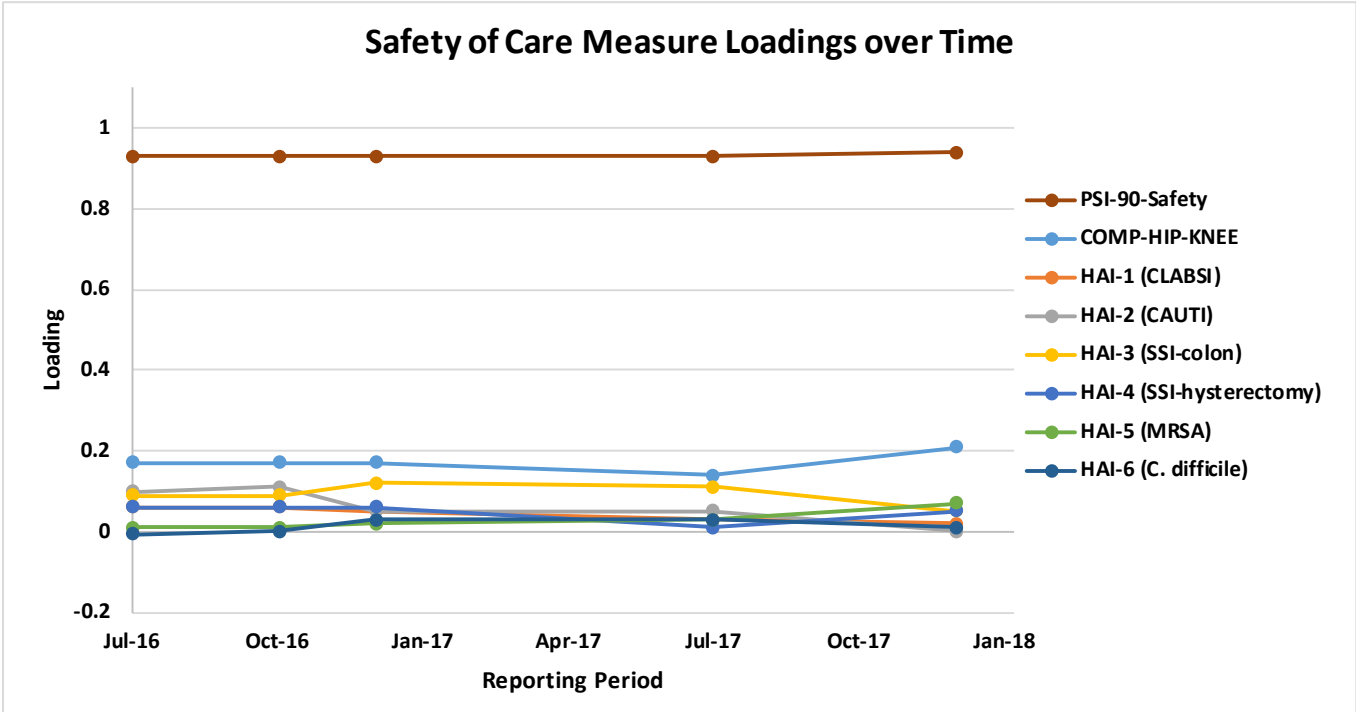
As an example, the loadings of measures in the Mortality measure group over time are shown in [Figure 5](#) below. While there is variation in the loading of different measures, they are still reasonably similar (ranging from 0.3 to 0.75). Additionally, the relative position of all measures is quite consistent across each period, with no loading shifting by more than 0.05 between periods.

Figure 5: Measure Loadings over Time, Mortality



In contrast, the pattern observed for the Safety of Care group is much different ([Figure 6](#) below). While the loadings are still reasonably consistent over time, they are not very well balanced. In particular, the PSI-90 (Patient Safety Indicator composite) measure historically has a much larger loading than other measures (greater than 0.90, while others are no greater than 0.20). Some stakeholders have expressed concern that this places too much emphasis on the PSI-90 composite measure while not emphasizing the other measures enough, particularly the HAI measures.

Figure 6: Measure Loadings over Time, Safety of Care



Stakeholder Feedback

Via the contractor, TEP members were supportive of the grouping criteria presented above. Several members commented that the contractor could explore other options for groupings, always with the intent of making information clear to consumers. One TEP member suggested exploring the second latent factor further to ensure it was not also measuring a signal of quality.

Questions for the Public:

- We would like to use a three-step approach (clinical coherence, confirmatory factor analysis, and ongoing monitoring) to define measure groups. Is this approach reasonable?
- Should CMS use the balance and consistency of loadings as a factor in evaluating grouping?

4.2. Regrouping of Measures

Based upon the initial grouping analyses presented above, CMS identified the current grouping of measures in Safety of Care as potentially contributing to challenges in consistency and predictability. In addition to other methodological updates, CMS is also considering regrouping measures, particularly within Safety of Care, in the near term to address stakeholder concerns. Previously, measures have been added or removed from programs that feed into *Hospital Compare* and therefore Overall Hospital Quality Star Rating, but otherwise the measure groups have not been altered. As performance evolves, these groupings may require re-specification to ensure groups are coherent; CMS would like feedback from stakeholders on this possibility.

CMS defined the current measure groups based primarily on clinical coherence and utility for consumers, so that measures are grouped with other measures relating to a similar domain or aspect of quality in a conceptually meaningful way. CMS used factor analysis to assess the empirical coherence of each group and found a single common factor for each group during the initial Overall Hospital Quality Star Rating development process. These groups were vetted extensively through a TEP, the Patient & Patient Advocate Work Group, and a previous Public Comment period.

CMS recently analyzed the Safety of Care group using the criteria above. It was found that this group had less consistent loadings, suggesting that the strength of the underlying latent variable may be weaker in the Safety of Care group compared to other measure groups. CMS hypothesized that model performance may be improved by subdividing measures in the Safety of Care group into two separate measure groups, each of which may have a single stronger factor. This potential method also functions with the possibility, suggested by some stakeholders, of replacing the PSI-90 composite measure with the individual PSI component measures; this change would also require a careful evaluation of group composition and may provide other options for regrouping.

CMS considered several alternative measure groupings for the Safety of Care measures, guided by clinical relevance and factor analysis. CMS assessed the eight current Safety of Care measures and determined that they could be clinically partitioned into *surgical safety* and *non-surgical* or *medical safety* groups as shown in [Table 2](#) below. (PSI-90 was assigned to the Surgical division because most [eight of ten] component measures are surgery-specific, as shown in [Table 3](#).)

Table 2: Safety of Care Measure Descriptions

Clinical Division	Measure	Description
Surgical	Comp-Hip-Knee	Complication rate, total hip or knee arthroplasty
	PSI-90	Patient Safety Indicator composite
	HAI-3	Surgical site infections (SSI) – colon surgery
	HAI-4	Surgical site infections (SSI) – hysterectomy
Medical	HAI-1	Central line-associated bloodstream infections (CLABSI)
	HAI-2	Catheter-associated urinary tract infections (CAUTI)
	HAI-5	MRSA infections
	HAI-6	C. diff infections

Alternatively, the PSI-90 measure could be divided into its ten component measures, which could also be assigned to either surgical or medical safety as shown in [Table 3](#) below (with the HAI and Comp-Hip-Knee measures maintaining the same partition as in [Table 2](#) above).

Table 3: PSI Component Measure Descriptions

Clinical Division	Measure	Description
Surgical PSI-90 components	PSI-06	Iatrogenic Pneumothorax rate
	PSI-09	Perioperative Hemorrhage or Hematoma rate
	PSI-10	Postoperative Acute Kidney Injury rate
	PSI-11	Postoperative Respiratory Failure rate
	PSI-12	Perioperative Pulmonary Embolism (PE) or Deep Vein Thrombosis (DVT) rate
	PSI-13	Postoperative Sepsis rate
	PSI-14	Postoperative Wound Dehiscence rate
	PSI-15	Unrecognized Abdominopelvic Accidental Puncture/Laceration rate
Medical PSI-90 components	PSI-03	Pressure Ulcer rate
	PSI-08	In-Hospital Fall with Hip Fracture rate

Using these clinical partitions provides two options to divide Safety of Care into two groups, summarized in [Table 4](#) below (one of which retains the PSI-90 composite, and one of which replaces PSI-90 with the ten PSI components).

Table 4: Options for Partitioned Safety of Care Groups

Groupings	Option 1: Retain use of PSI-90	Option 2: Switch to PSI components
Medical Safety group	HAI-1, HAI-2 HAI-5, HAI-6	HAI-1, HAI-2 HAI-5, HAI-6 PSI-3, PSI-8
Surgical Safety group	Comp-Hip-Knee HAI-3, HAI-4 PSI-90	Comp-Hip-Knee HAI-3, HAI-4 PSI components: 6, 9—15

CMS did not consider removal of any measures from the Safety of Care group(s) beyond any finalized for removal from various CMS quality programs that feed into *Hospital Compare* to ensure alignment with an original principal of Overall Hospital Quality Star Rating: to include as many measures as possible.

CMS evaluated the potential groupings using the criteria identified above:

- Ratio of first to second eigenvalue (weighted factor analysis);
- Qualitative Scree plot comparison; and
- Measure loading balance and consistency.

Option 1 (retaining PSI-90) resulted in an eigenvalue ratio of 51 (very strong) in the Medical Safety group but 1.5 in Surgical Safety, which is lower than eigenvalue ratio of 2.34 for the existing grouping for Safety of Care. The results for Option 1 can be seen in the Scree plots, in which there is a sharp difference in eigenvalues for the Medical Safety group, but not the Surgical Safety group ([Appendix B](#)). Measure loadings for these options, as shown in [Tables 5](#) and [Table 6](#) below, were fairly stable in both groups over time and reasonably balanced in the Medical Safety group, but less well balanced in Surgical Safety group.

Table 5: Loadings, Medical Safety group, Option 1 (retain PSI-90)

Measure	Jul. 2016	Dec. 2016	Jul. 2017	Dec. 2017	Jul. 2018
HAI-1	0.49	0.54	0.50	0.48	0.47
HAI-2	0.33	0.28	0.24	0.23	0.26
HAI-5	0.27	0.30	0.32	0.24	0.22
HAI-6	0.06	0.08	0.07	0.06	0.10

Table 6: Loadings, Surgical Safety group, Option 1 (retain PSI-90)

Measure	Jul. 2016	Dec. 2016	Jul. 2017	Dec. 2017	Jul. 2018
COMP-Hip-Knee	0.17	0.17	0.14	0.21	0.20
HAI-3	0.09	0.10	0.10	0.05	0.05
HAI-4	0.06	0.06	0.003	0.05	0.04
PSI-90	0.94	0.94	0.94	0.94	0.90

Option 2 (using PSI components instead of PSI-90) improved the eigenvalue ratio, in particular for the Medical Safety group, in comparison with the existing grouping for Safety of Care (eigenvalue ratio of 2.34). The ratio of 6.6 in Medical Safety is a strong indicator of a single factor for the group; this can be seen in the sharp “elbow” of the Scree plot at the second eigenvalue ([Appendix B, Figure B1](#)). The ratio of the Surgical Safety group is 2.4, comparable to the existing grouping.

Loadings for the groups in this option are shown in [Tables 7](#) and [Table 8](#) below. The loadings for the measures in the Medical Safety group are fairly consistent over time but not very well balanced, with PSI-3 dominating the group. The loadings in the Surgical Safety group appear more balanced and also fairly consistent over time, with the exception of December 2017, in which large changes were observed.

Table 7: Loadings, Medical Safety Group, Option 2 (PSI components)

Measure	Jul. 2016	Dec. 2016	Jul. 2017	Dec. 2017	Jul. 2018
HAI_1	0.06	0.05	0.03	0.02	0.03
HAI_2	0.05	0.03	0.05	0.02	0.04
HAI_5	0.09	0.06	0.05	0.05	0.03
HAI_6	0.01	0.03	0.04	0.02	0.02
PSI-3	0.69	0.69	0.69	0.59	0.42
PSI-8	-0.003	-0.003	-0.003	-0.01	0.03

In [Table 8](#) below, changes in loading from the previous period greater than 0.2 in either direction are denoted with an asterisk (*). Changes of this magnitude were only observed in the Surgical Safety group using PSI components and not in any other potential groups.

Table 8: Loadings, Surgical Safety Group, Option 2 (PSI components)

Measure	Jul. 2016	Dec. 2016	Jul. 2017	Dec. 2017	Jul. 2018
COMP_HIP_KNEE	0.49	0.49	0.42	0.98*	0.43*
HAI_3	0.10	0.16	0.15	-0.03	0.11
HAI_4	0.09	0.11	0.03	0.01	0.06
PSI-6	0.11	0.11	0.13	0.02	0.16
PSI-9	0.19	0.19	0.20	-0.003*	0.10
PSI-10	0.20	0.20	0.23	0.02*	0.26*
PSI-11	0.26	0.27	0.31	0.06*	0.24
PSI-12	0.37	0.36	0.32	0.11*	0.22
PSI-13	0.14	0.14	0.15	0.06	0.25
PSI-14	0.02	0.02	0.02	-0.01	0.03
PSI-15	0.12	0.13	0.15	-0.01	0.14

***Denotes changes in measure loading from the previous period greater than 0.2 in either direction**

If CMS begins a process of regrouping measures that substantially changes the composition of measure groups, further input from the public and stakeholders would be needed to evaluate the new groups and new group weights. Importantly, the measure groups themselves are currently weighted to create each hospital's summary score, according to the importance of each group to "overall" quality; the current methodology gives a base weight of 22% to each of the three outcome groups (Mortality, Readmission, and Safety of Care) and to Patient Experience, and 4% to each of the process groups (Timeliness, Effectiveness, and Imaging Efficiency). These weights have been vetted extensively with technical experts, patient advocates, and the public to reflect a broad range of input.

Stakeholder Feedback

TEP members were generally not supportive of either of the re-grouping options, as they did not achieve the grouping criteria, or the goal of more balanced measure loadings. TEP members suggested focusing on other areas for reevaluation, such as statistical modeling and user-customized Overall Hospital Quality Star Ratings to address the measure loadings rather than the groups themselves. While Provider Leadership Work Group members were comfortable with the current measure groupings, they acknowledged the eventual removal of several measures and the need to reconsider the measure groups.

Questions for the Public:

- Is the current grouping or one of the potential alternative groupings of the Safety of Care measures most suitable for the Overall Hospital Quality Star Rating based on previously mentioned criteria?

4.3. Incorporating Precision of Measures

4.3.1. Background

The current Overall Hospital Quality Star Rating methodology uses denominator weighting in order to account for differences in measure score precision, so that hospitals and measures with a larger denominator are more heavily weighted in each LVM. This ensures that hospitals are scored more heavily on measures including more patients and get more weight when estimating loadings. This approach is consistent with the approach used for many aggregated individual measures to ensure that more precise estimates are given more emphasis, given that denominators are generally correlated with precision. For a sample mean, the inverse square of the standard error equals the sample size divided by the population variance ($1/SE^2 = n/\sigma^2$)—that is, the inverse squared standard error is proportional to the denominator size.

Recent assessment of the Safety of Care measure group, however, revealed that while denominator weighting may reflect sample size differences, it may also contribute to the imbalance of measure loadings and worse model fit. While the exact cause of this effect is unknown, the different measures in Safety of Care, unlike other groups, use different types of denominators which have skewed denominator distributions; as such, this may contribute to worse model fit and overwhelm potential benefits (for example, some HAI measures discussed previously in [Section 3.5](#) use patient days, while the Mortality measures use the number of admissions).

CMS has sought to quantify the benefits and disadvantages of denominator weighting and evaluated other alternative approaches for incorporating measure score precision into the Overall Hospital Quality Star Rating including: weighting by the logarithm of the denominator, confidence interval-based weighting, or removing weighting altogether.

CMS surveyed the current Overall Hospital Quality Star Rating measures and found that those in the outcome groups (Mortality, Readmission, and Safety of Care) include some adjustment for precision by accounting for volume in the score itself, while the measures in the four remaining groups (Patient Experience, Effectiveness, Timeliness, and Imaging Efficiency) have no such adjustment. This suggests that some information in denominator weighting is already accounted for by individual measures within outcome measure groups.

The measures in the remaining four groups (Patient Experience, Effectiveness, Timeliness, and Imaging Efficiency) do not utilize risk-adjustment models and do not have confidence intervals available, meaning that volume-based weighting is the only option to account for precision of measures in these groups.

CMS explored denominator, confidence interval, and no weighting in each relevant measure group (Mortality, Readmission, and Safety of Care) by comparing model fit statistics and evaluating measure loadings for consistency and balance. Results are shown below for the Safety of Care group as an illustrative example. Note that after these analyses were completed, CMS added the additional approach – weighting by the logarithm of the denominator. CMS explored this option by evaluating Safety of Care loadings.

4.3.2. Analyses

Model Fit Statistics

CMS measured the [weighted mean square error \(MSE\)](#) using data from December 2017, July 2018, and February 2019, for each of the three options (denominator weighting, confidence interval weighting, and no weighting). Results are shown in [Table 9](#) below. A lower MSE indicates a better fit when using the same data but is not comparable when using different data sets. Please note that MSE is only one metric that may be used to compare performance of different models and has its own limitations; it is only one possible indicator to consider when choosing a model.

Table 9: Weighted Mean Square Error (MSE) by Weighting Option, Safety of Care

Period	Denominator Weighting (current)	Confidence Interval Weighting ($1/CI^2$)	No Weighting
Dec. 2017	0.57259	0.55325	0.7526
Jul. 2018	0.58696	0.55240	0.74295
Feb. 2019	0.58071	0.53858	0.74028

Within each period, MSE is smallest when using confidence interval weighting, suggesting that this model is the best fit for the data in the Safety of Care group. Denominator weighting produced a fit that was slightly worse than confidence interval weighting but still reasonably close (as expected given the correlation between denominator size and precision). At the same time, the unweighted models had substantially greater MSE, suggesting the model fit least well; this in turn suggests that accounting for precision contributes valuable information to the model. Use of confidence interval weighting within the Safety of Care measure group also appears to improve the stability and consistency of model performance during simulation analyses.

Loadings

[Table 10](#) below shows measure loadings for the Safety of Care group in July 2018 and February 2019 using each of the three options.

Table 10: Loadings by Weighting Option, Safety of Care

Measures	Denominator Weighting (Current) July 2018	Denominator Weighting (Current) February 2019	Confidence Interval Weighting ($1/CI^2$) July 2018	Confidence Interval Weighting ($1/CI^2$) February 2019	No Weighting July 2018	No Weighting February 2019
Comp-Hip-Knee	0.20	0.20	Sig. Neg.	0.13	0.04	0.06
HAI-1	0.02	0.01	0.33	0.32	0.52	0.62
HAI-2	0.003	0.01	0.35	0.33	0.40	0.38
HAI-3	0.05	0.05	0.35	0.31	0.24	0.19
HAI-4	0.04	0.07	0.18	0.19	0.29	0.25
HAI-5	0.05	0.04	0.23	0.21	0.30	0.37
HAI-6	0.02	0.03	0.34	0.36	0.14	0.09
PSI-90	0.88	0.90	0.14	0.17	0.11	0.09

Notably, none of the options completely resolves concerns about unbalanced loadings, although denominator weighting has the largest disparity between the highest loading and the remaining measures. In particular, both confidence interval weighting and no weighting produced higher loadings for the six HAI measures, while reducing the loadings of hip-knee complications and PSI-90; this indicates a better balance of measures' influence on the group score. Both confidence interval weighting and no weighting also produced loading estimates that were more consistent between the two quarters, potentially indicating better predictability for hospitals.

In addition to Safety of Care, the measures in Mortality and Readmission include confidence intervals for scores that can be used for this purpose. Loadings for Mortality were generally very stable and well-balanced regardless of the choice of weight, and in fact were quite similar; this is likely because the technical specifications of all measures are very similar. Results in Readmission were similar, but marginally less consistent; this could be because the group is a combination of 30-day readmission and excess days in acute care (EDAC) measures.

More recently, CMS explored the option of weighting by the logarithm of the denominator. Log transformation is a common approach for rescaling distributions that are skewed. We hypothesized that applying it to the denominators that are highly asymmetric might improve the stability and regularity of the loadings. [Table 11](#) shows the measure loadings for February 2019 data, comparing the results of the current denominator approach with those of the log transformation of the denominator.

Table 11. Comparison of February 2019 Measure Loadings for Safety of Care Group Using Log Transformation of the Denominator

Measure	Denominator (Current)	Log transformation [log(denominator)]
COMP-HIP-KNEE	0.20	0.10
HAI-1	0.01	0.53
HAI-2	0.01	0.37
HAI-3	0.05	0.21
HAI-4	0.07	0.28
HAI-5	0.04	0.36
HAI-6	0.03	0.08
PSI-90-SAFETY	0.90	0.13

4.3.3. Measure Precision Options: Advantages & Disadvantages

CMS has summarized its assessment of advantages and disadvantages of different weighting options in [Table 12](#) below.

Table 12: Advantages and Disadvantages of Weighting Options

Weighting Option	Advantages	Disadvantages
Denominator Weighting	<ul style="list-style-type: none"> • Current approach • Accounts for precision of measurements • Hospitals with more patients more heavily influence loadings and group scores • All measures have available denominators 	<ul style="list-style-type: none"> • May not produce desired effect in some groups due to denominator distributions • Some measures do not use patient- or admission-level denominators and may perform differently as a result
Confidence Interval Weighting	<ul style="list-style-type: none"> • Serves the conceptual purpose of denominator weighting (accounting for measure precision) • Hospitals scored more heavily on measures with more patients • Best represents the concept of statistical precision 	<ul style="list-style-type: none"> • Using confidence intervals as a proxy for variance, which would be preferred • Not useable in all groups, as most process group measures do not have confidence intervals

Weighting Option	Advantages	Disadvantages
		<ul style="list-style-type: none"> Implementation will substantially affect hospital ratings
No Weighting (equal weighting)	<ul style="list-style-type: none"> Avoids potential redundant inclusion of denominator information 	<ul style="list-style-type: none"> Does not account for precision of measurements Does not score hospitals more heavily on measures with more patients Implementation will substantially affect hospital ratings
Log (denominator) weighting for non-volume denominators, otherwise denominators	<ul style="list-style-type: none"> Retains relationship with precision Improves consistency of weights that are highly skewed. 	<ul style="list-style-type: none"> Mixed weighting scheme Not intuitive; other transformations could serve the same purpose

Each option has advantages and disadvantages. CMS believes that incorporating measure precision in the Overall Hospital Quality Star Rating is conceptually important but would like to gain public feedback on this matter.

Stakeholder Feedback

In general, TEP members were in agreement that accounting for measure precision was important. TEP members were also in agreement that a statistically sound method that resulted in more balanced measure loadings that are consistent over time would be beneficial. Most TEP members favored the confidence interval method but agreed that the log transformation method would be mathematically appropriate. Provider Leadership Work Group members supported investigating an approach to balance measure loadings despite the expected shifts in star ratings if this update were to be implemented.

Questions for the Public:

- Do you have any concerns about changing the methodology to use a combination of denominator weighting and log (denominator) weighting, based on the type of measure?
- Do you have any concerns about applying a change to the weighting approach across all measure groups (where data are available) vs. applying the change only to measure groups that meet specific criteria?
- Are there other approaches that CMS should consider?

4.4. Period-to-Period Star Rating Shifts

4.4.1. Background

Based on stakeholder feedback regarding larger shifts in ratings in July 2018 than some expected, CMS chose to evaluate methods that could make the Overall Hospital Quality Star Rating more stable between refreshes. Stakeholders were particularly concerned that such large shifts were observed in a six-month period and indicated it can be difficult to explain these changes in rating despite observing relatively modest changes in performance on individual measures.

CMS studied historical Overall Hospital Quality Star Rating shifts and found that more hospitals shifted by at least two stars in July 2018 than in previous periods. CMS attributes these shifts to changes in individual measures,

including the annual refresh of many important outcome measures and the methodology update to PSI-90 (as discussed previously in [Section 3.1](#)).

CMS also noted that, historically, there have been more substantial Overall Hospital Quality Star Rating shifts in July refreshes than in December refreshes (after accounting for the one-time effects of methodology updates by re-calculating all periods with the most recent methodology). This coincides with the reporting schedule of individual measures, many of which are refreshed only in July every year. However, there still were some substantial changes in December, despite actual changes in measure scores being generally minor (as many highly-weighted measures are not refreshed and others that are refreshed often have overlapping data periods). In a previous TEP meeting, panelists suggested that this indicates some changes in Overall Hospital Quality Star Ratings are due to the effect of subtle changes in the underlying data that result in hospitals, particularly those with borderline scores, falling into a different star category.

Based on this observation as well as feedback from a previous TEP meeting, CMS is considering a transition to an annual refresh schedule for the Overall Hospital Quality Star Rating (discussed previously in this document but added here since it is also applicable to addressing period-to-period shifts). This has the advantage of ensuring that every measure refreshes before each Overall Hospital Quality Star Rating calculation, and that changes in hospitals' ratings can be more clearly attributed to observed changes in performance on the underlying measures.

However, given the sensitivity of the methodology to subtle changes in individual measure scores, CMS believes that some stakeholders may still experience substantial changes in rating. As such, CMS additionally considered using a rolling average of summary score information as a policy-based approach to attenuating period-to-period changes.

4.4.2. Weighted Average Summary Scores

CMS would like to gain public input on a potential option that would reduce period-to-period changes in the Overall Hospital Quality Star Rating by incorporating data from an older period.

Background

Several stakeholders have asked CMS to consider updating the Overall Hospital Quality Star Rating methodology so that changes in performance are incorporated gradually, rather than in a single period when measures are added, updated, or refreshed. CMS considered operationalizing this by using data from both the current and immediately prior reporting period of *Hospital Compare*. This approach would systematically introduce more consistency in scores and reduce variability between periods while also allowing hospitals more time to adapt to new or changed measure scores (although most measures are already refreshed with overlapping data periods).

Other star ratings, such as that on Nursing Home Compare, have adopted a weighting scheme for a component of their rating (the inspection surveys) in which a nursing home's total weighted score for that component is calculated by weighting more recent surveys more than previous surveys.⁶

However, CMS notes that many individual measures already include some overlapping data between *Hospital Compare* refreshes due to the amount of data required for each performance period, meaning that the above approach is already partially incorporated into the methodology. For example, readmission measure scores are refreshed annually but have three years of data contributing to the hospital scores so each refresh has two years

⁶ Centers for Medicare & Medicaid Services. Design for Nursing Home Compare Five-Star Quality Rating System: Technical User's Guide. July 2018; <https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/CertificationandCompliance/downloads/usersguide.pdf>. Accessed January 28, 2019

of data from the previous refresh and one year of new data. In addition, other stakeholders expressed concern that older data may be outdated and less reflective of current performance and have advocated for CMS to use the most recent available data.

Analyses

CMS assessed how combining a weighted average of older and current data would affect the Overall Hospital Quality Star Rating by reviewing hospital star rating reclassifications under three conditions: the current method (using only the most recent data); a 75%-25% method (with new data receiving 75% of the weight and data from the previous period the other 25%); and a 50%-50% method. CMS applied this weighting scheme to hospitals' overall summary scores based on measures available for that period.

For example, suppose Hospital A reports measures to *Hospital Compare* every quarter. A summary score for Hospital A in July is calculated using the current Overall Hospital Quality Star Rating methodology (calculating measure group scores using LVM and taking a weighted average of group scores) and data published on *Hospital Compare* in July. Hospital A also receives a summary score in December, using the same methodology but with refreshed December data. Under the current methodology, only the December summary score is used to assign December star ratings. In any of the weighted methodology approaches, the December and July summary scores would be averaged together to assign December star ratings. The following July would then include the new July score with the old December score, and so on.

[Table 13](#) below shows the changes in Overall Hospital Quality Star Rating that would have been observed in July 2017, December 2017, and July 2018 using each of the schemes (no weighting, 75% new-25% old, and 50% new-50% old).

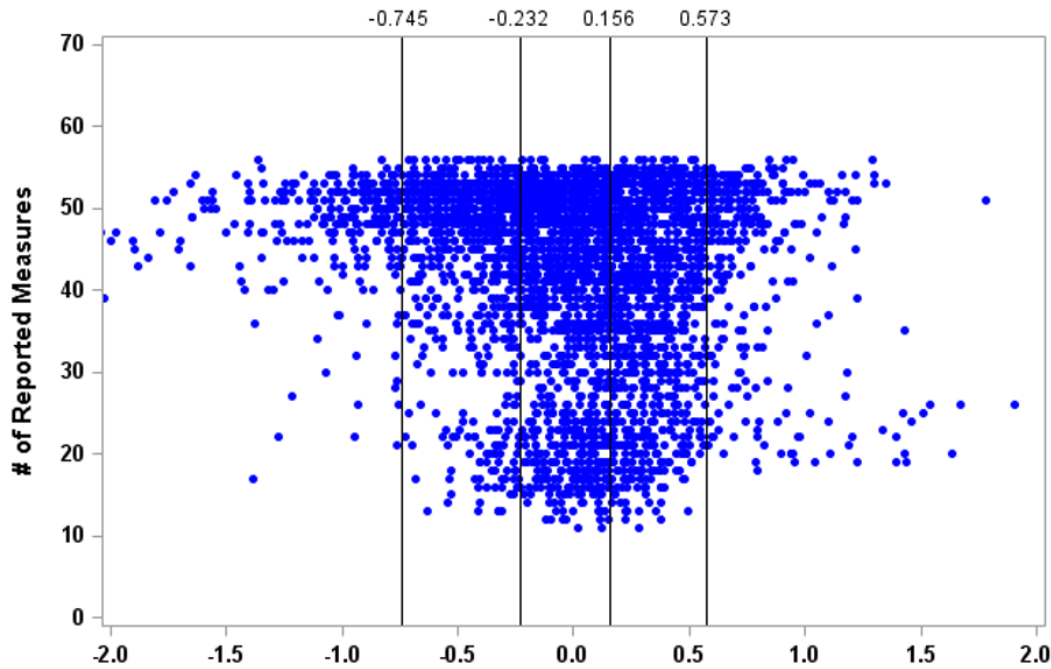
Table 13: Shifts in Star Rating by Weighting Option

Weighting Scheme	Change in Star Rating:	Dec.2016 – Jul. 2017 (n=3,556)	Jul. 2017 – Dec. 2017 (n=3,600)	Dec. 2017 – Jul. 2018 (n=3,630)
No weighting (current)	-2 Star or more	20 (0.56%)	17 (0.47%)	73 (2.0%)
	-1 Star to +1 Star	3526 (99.2%)	3531 (98.1%)	3478 (95.8%)
	+2 Star or more	10 (0.28%)	52 (1.4%)	79 (2.2%)
Weighting: 75% new, 25% old	-2 Star or more	4 (0.11%)	4 (0.11%)	23 (0.63%)
	-1 Star to +1 Star	3551 (99.9%)	3582 (99.5%)	3581 (98.6%)
	+2 Star or more	1 (0.03%)	14 (0.39%)	26 (0.72%)
Weighting: 50% new, 50% old	-2 Star or more	1 (0.03%)	1 (0.03%)	6 (0.17%)
	-1 Star to +1 Star	3554 (99.9%)	3584 (99.6%)	3615 (99.6%)
	+2 Star or more	1 (0.03%)	15 (0.42%)	9 (0.25%)

Reading down each column shows the reclassification that would have been observed when incorporating previous data at a progressively higher weight. Notably, incorporating previous data at higher weights reduces major reclassification (shifts of two or more stars) within each period. Among hospitals experiencing changes, the changes were progressively more limited, with a greater number of hospitals either receiving the same rating or changing by only one star in either direction.

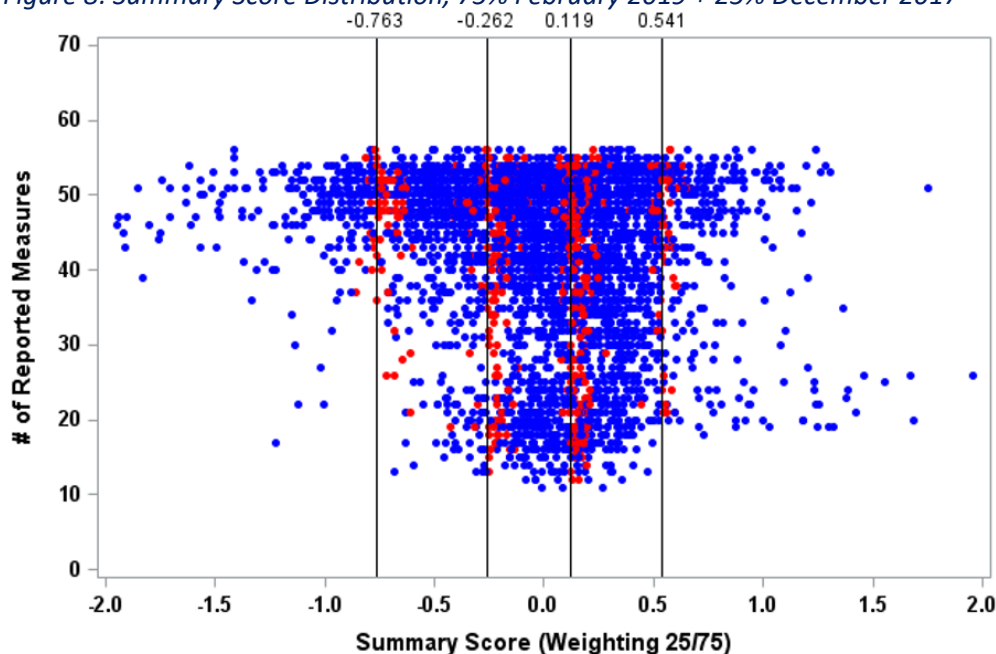
[Figure 7](#) and [Figure 8](#) below illustrate these shifts using February 2019 Overall Hospital Quality Star Ratings. [Figure 7](#) below shows the overall distribution when using only February 2019 summary scores, with the vertical lines indicating star rating cut points.

Figure 7: Summary Score Distribution, February 2019



[Figure 8](#) below shows what the distribution would look like if incorporating December 2017 data as 25% of the summary score, with hospitals receiving a different rating indicated in red. Five hundred and fifty-one hospitals (15%) would have received a different rating as a result. (Please note that the red dots indicate hospitals with different February 2019 ratings when using the weighting scheme compared to no weighting, not hospitals that changed since December). This illustrates that hospitals most likely to be affected are those near the cutoff points between star rating categories, which holds to some extent using other variations of this weighting scheme.

Figure 8: Summary Score Distribution, 75% February 2019 + 25% December 2017



These observations suggest that using a weighted summary score would result in hospitals receiving new ratings closer to their previous rating than they would using only the most recent data, with a greater effect for using more data from the previous period. This can be observed particularly among borderline hospitals in the figure above, which would be expected given that subtle differences in hospital summary scores may determine a hospital's star rating in either direction. This would make the Overall Hospital Quality Star Ratings less sensitive to changes in individual measures to some degree; however, the degree to which reduced sensitivity is desirable is unclear as some stakeholders have also previously indicated that the Overall Hospital Quality Star Ratings should reflect the most recent data.

Stakeholder Feedback

All three stakeholder groups (TEP, Provider Leadership Work Group, and Patient & Patient Advocate Work Group) were not in favor of this approach; all groups agreed that it was more important to use the most current data rather than including older data; Patient & Patient Advocate Work Group members further noted that using data from previous periods could be misleading to consumers, who value having the most current information. TEP members suggested alternative ways to reduce period-to-period shifts: one TEP member suggested exploring "partial" star ratings, such as 4.5 stars; another TEP member suggested using three star categories rather than five. In addition, one TEP member inquired about moving to annual updates of the Overall Hospital Quality Star Rating; another TEP member agreed with this approach.

Questions for the Public

- What are possible benefits and drawbacks to increasing stability by limiting change in this way?
- Should the Overall Hospital Quality Star Rating methodology be modified to incorporate data from previous periods through a time averaged approach?
- Are there other approaches to this CMS should consider?

4.5. Peer Grouping

4.5.1 Background

Some hospital stakeholders have expressed interest in calculating and presenting Overall Hospital Quality Star Rating results based on hospitals that “look like them,” which we refer to in this document as “peer grouping.” For example, safety-net hospitals could be grouped together to generate a star rating, teaching hospitals could be grouped together, and small/rural/Critical Access Hospitals could be grouped together) or CMS could consider use of bed size to distinguish.

Recently, CMS implemented peer grouping within the Hospital Readmission Reduction Program (HRRP).⁷ In HRRP, CMS calculates a penalty threshold relative to other hospitals within a peer group. Specifically, CMS stratifies hospitals into five peer groups (quintiles) based on hospitals’ proportion of dual-eligible patients. CMS then uses the median “excess readmission ratio” for hospitals within a peer group as the threshold for determining payment penalty on each readmission measure in the program (Please visit [CMS website](#) for more detail on HRRP methodology).

A similar approach could be used in the Overall Hospital Quality Star Rating methodology to allow for direct comparisons of performance on star ratings between hospitals within a peer group for a particular hospital characteristic (proportion of dual-eligible patients, or another feasible variable such as teaching hospitals, critical access hospitals, or number of measures reported). This could involve calculating the Overall Hospital Quality Star Rating for a hospital based on its peer group assignment. This could be done at different steps within the methodology, for example, at the k-means clustering step for hospitals within a peer group for a particular hospital characteristic.

TEP, providers, patients, and the public have provided preliminary input on the option of peer grouping. Some stakeholders supported the concept, while others felt it would not be helpful and would be confusing, particularly to consumers and patients. In addition, there was a lack of consensus on which variables to use if peer grouping were implemented.

CMS continues to receive interest from hospital stakeholders on this issue, and recently obtained updated feedback from certain stakeholders. CMS is interested in receiving additional public input on this topic. Past and recent feedback are outlined below.

4.5.2 Prior Stakeholder Feedback

Public Input

During the previous Overall Hospital Quality Star Rating public input period from August to September of 2017, CMS received feedback from 22 individual commenters on peer grouping. Most comments, representing hospitals and hospital associations, were in support of peer grouping by similar types of hospitals. However, there was no consensus on what variables to group by, and many candidate variables were not feasible due to the information not being consistently available for all hospitals. Those who were not in favor of peer grouping noted the complexity or confusion it would add, and that it would conflict with the original goal of a simple summary rating for consumers.

⁷ Department of Health and Human Services, Centers for Medicare & Medicaid Services. Fiscal Year 2019 Hospital Inpatient Prospective Payment Systems Final Rule. August 2018; <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2019-IPPS-Final-Rule-Home-Page-Items/FY2019-IPPS-Final-Rule-Regulations.html>. Accessed January 28, 2019.

Patient & Patient Advocate Work Group

The Patient & Patient Advocate Work Group universally did not support peer-grouped Overall Hospital Quality Star Ratings for *Hospital Compare* based on the belief that it would be both confusing, potentially misleading, and not meaningful to consumers. The group advocated for not changing the single summary star rating, but supported the idea that if CMS institutes peer grouping, it should be supplemental to the Overall Hospital Quality Star Rating. Patient & Patient Advocate Work group members were interested in a filtering function on *Hospital Compare* but one that allows consumers to identify hospitals by location and healthcare network, rather than hospital characteristics.

Technical Expert Panel

The TEP expressed mixed reactions to the topic of peer grouping. Some agreed it would be unhelpful and confusing for patients, while others felt it was important to acknowledge differences in hospitals. There was no consensus on what variable to use for peer grouping.

When asked specifically about using an approach similar to HRRP, TEP consensus was unsupportive of dual-eligible proportion as a stratification variable, due to the potential to set different standards of care for different populations. TEP members also felt that addressing differences among hospitals should be done through measure-level risk adjustment rather than peer grouping.

When asked about peer grouping to allow for comparison on other hospital characteristics (such as bed size, or teaching status), some TEP members were supportive of the idea of a web-based tool that would allow for comparisons between hospitals within the same peer group. However, some TEP members emphasized that clarity to support consumer decision-making should be a top priority for the Overall Hospital Quality Star Ratings; one member pointed out that patients who use the Overall Hospital Quality Star Ratings, use them to make choices between hospitals available to them (based on proximity, or insurance coverage), not hospitals like each other.

Provider Leadership Work Group

The Provider Leadership Work Group has consistently supported peer grouping, but did not provide any consensus support for the variables analyzed or for any stratification methodology.

Questions for the Public:

1. Would it be valuable to calculate Overall Hospital Quality Star Ratings among peer groups? How should the information be displayed? If CMS decides to move forward with this feature, which stakeholders do you believe would use the information and how would they use it?
2. Among the feasible variables that could be used for peer grouping (specialty, number of measures reported, teaching status, number of beds, critical access hospital, proportion of dual eligible patients), which would be most useful? Descriptions for each mentioned variable are included below.
 - a. Proportion of dual-eligible describes the proportion of patients eligible for both Medicare and Medicaid. Dual-eligible could be used to peer group hospitals with similar proportions of dual-eligible patients by quintile, for example.
 - b. Teaching hospitals are those that have one or more accredited residency programs or have an intern or resident to bed ratio of 0.25 or higher. Teaching and non-teaching hospitals may differ in mission, financial considerations, and services. Teaching status could be used to peer group teaching and non-teaching hospitals.
 - c. Number of beds at a hospital is a proxy for hospital size. Smaller hospitals may have fewer services and resources while larger hospitals tend to be in urban areas and may serve disadvantaged populations.

- d. Hospitals that report more measures may not be directly comparable to hospitals that report fewer measures. Number of measures reported could be used to group hospitals by quartile, for example.
- e. Certain rural hospitals can qualify as critical access designation for CMS purposes to indicate lack of proximity to other hospitals for prospective patients. Hospitals could be grouped as either critical access or non-critical access.
- f. Specialty hospitals are those that primarily or exclusively engage in the care and treatment of patients with cardiac conditions, orthopedic conditions, conditions requiring surgical procedures, or other specialized services. Hospitals could be grouped and compared as specialty or non-specialty.

4.6 Computational Update: Closed-Form Solution of LVM

Currently, the Overall Hospital Quality Star Rating methodology uses an approach known as quadrature to solve the mathematical equations of the latent variable models and calculate hospitals' measure group scores. This approach produces accurate and precise solutions, but can take a long time to compute.

CMS recently developed a different approach for solving these equations that can be incorporated into the statistical program (SAS 9.3) that calculates the Overall Hospital Quality Star Rating results. This methodology uses a "closed-form solution" to more quickly solve the equations, and eliminates the need for the computationally time-consuming quadrature approach. Utilizing this new approach means that the star rating results can be calculated much faster, which increases its usefulness for: producing results for public reporting; quality control; ongoing methodology evaluation; and re-creation by the public (CMS makes the code and datasets necessary for replicating the Overall Hospital Quality Star Ratings freely available). In addition, the improved efficiency allows the software to produce more precise and stable results than what was feasible using the quadrature approach.

The mathematical details of this new solution method are technically complex; those interested in learning more may refer to [Appendix C](#) which presents the specifications in depth. For those among the public with experience in programming or mathematics, CMS is interested in any feedback on this approach from a technical perspective.

CMS seeks input from the public in general on the conceptual merits of making this update. CMS analyses have shown that the new algorithm modestly improves precision of results but does not have a major substantive impact; this change would be a technical modification that greatly improves the usability of the code with at most a trivial effect on results.

Stakeholder Feedback

Few TEP members had input on this technical change. One TEP member agreed this approach was more suitable than the quadrature approach that is currently used.

Question for the Public:

- Should CMS use a "closed-form solution" or make technical changes like this potential solution and consider opportunities for such changes in the future?

5. Potential Long-Term Methodology Changes

5.1. Background

The Overall Hospital Quality Star Rating has continued to perform in alignment with its initial principles and has received substantial support from many stakeholders. However, several parts of the methodology may be suitable for substantial redesign to ensure the ratings reflect the quality information available on *Hospital Compare* and meet the needs of healthcare providers and consumers.

CMS has identified several topics to consider for guiding future work, all of which reflect stakeholder input. These are summarized here and discussed in greater detail further below:

- Replacing LVM with an explicit approach (such as an average of measure scores) to group score calculation;
- Using an alternative approach to clustering;
- Incorporating facilities' improvement into their scores; and
- User-customized ratings.

These topics are considered long-term considerations in that the scope of such changes are being considered for reporting in 2020 and beyond. CMS is seeking input on these topics to guide the direction of future work. Please also note that these topics are presented in isolation but are not necessarily incompatible with each other or with other parts of the current methodology.

5.2. Explicit Approach

Background

Latent variable modeling offers several advantages in summarizing measure groups' information (as summarized in the Comprehensive Methodology Report v3.0)⁸:

- Used for other composite measures in healthcare quality literature⁹;
- Accounts for consistency of performance by giving more importance to measures that are correlated within a group;
- Accounts for missing measures by accounting for all available information, meaning hospitals with varying amounts of information can be accommodated in the model;
- Accounts for sampling variance and differences in precision of measure scores; and
- Easily accommodates changes to *Hospital Compare* over time.

However, some stakeholders have given feedback that LVM is not an intuitive or easy-to-understand methodology, and have suggested a less complex or more explicit approach, like those currently used in other CMS star rating methodologies, such as Medicare Part C & D Star Ratings (see example below). While data-driven aspects of the LVM may reduce arbitrariness, the approach introduces inherent uncertainty into the process:

⁸ Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (YNHHSC/CORE). Overall Hospital Quality Star Ratings on *Hospital Compare* Methodology Report (v3.0). December 2017; <https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228775957165>. Accessed January 28, 2019.

⁹ Schwartz, M., Restuccia, J. D., & Rosen, A. K. (2015). Composite Measures of Health Care Provider Performance: A Description of Approaches. *The Milbank quarterly*, 93(4), 788-825.

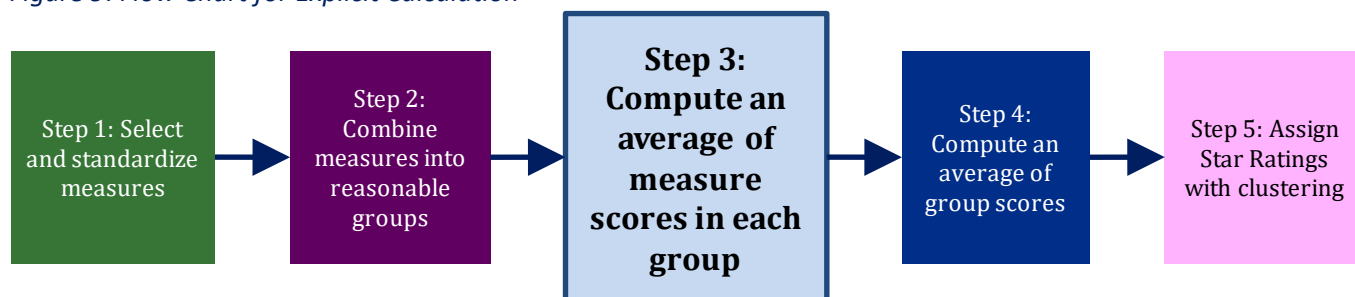
loadings are determined empirically based on the available data and may change over time, making it less transparent how changes in individual scores will translate into hospital star ratings.

CMS would like input from the public about alternative approaches to LVM that assign explicit (though arbitrary) weights to each measure in each group, independently of the performance distribution or relationships between measures.

Example

An explicit approach could be implemented in different ways. CMS considered an example in which the current methodology is unchanged, except at the group score calculation step. Instead of latent variable modeling, CMS would assign weights to each measure in each group, then calculate each hospital's group score as a weighted arithmetic average of its measure scores. This is illustrated in [Figure 10](#) below. Note that the Medicare Part C & D ratings use this approach.¹⁰

Figure 9: Flow Chart for Explicit Calculation



As an example of how the calculation may work, CMS created an example using a mortality group of three measures. Each measure is assigned a weight that is the same for all hospitals. In the simple case, each measure receives the same weight; however, a system could also be used in which each measure gets a different weight. Each hospital's group score is then the sum of the products of the measure weight with the measure score, as shown in [Table 14](#) below. In this example, Hospital A and B receive the same summary score when using equal weighting for measures. Using an example of differently weighted measures results in a lower score for hospital A and a higher score for hospital B, due to the relative performance and weighting of measures.

¹⁰ Centers for Medicare & Medicaid Services. Medicare 2019 Part C & D Star Ratings Technical Notes. September 2018; <https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/Downloads/2019-Technical-Notes.pdf>. Accessed January 28, 2019.

Table 14: Example of Explicit Group Score Calculation, Equal Weights vs. Different Weights

Measure	Measures have equal weights			Measures have different weights		
	Measure weight	Hospital A standardized measure scores	Hospital B standardized measure scores	Measure weight	Hospital A standardized measure scores	Hospital B standardized measure scores
MORT-AMI	1/3	0.2	1.5	0.45	0.2	1.5
MORT-HF	1/3	-0.7	0.2	0.35	-0.7	0.2
MORT-PN	1/3	1.5	-0.7	0.2	1.5	-0.7
Group score	--	0.333 (1/3)* (0.2-0.7+1.5)	0.333 (1/3)* (1.5+0.2-0.7)	--	0.145 [(0.45*0.2)– (0.35*0.7)+(0.2*1.5)]	0.605 [(0.45*1.5)+(0.35*0.2)– (0.2*0.7)]

An advantage of LVM that would be lost is that it allows the data to empirically estimate loadings based on the correlations between measures for each refresh. Therefore, the LVM approach may be more feasible to maintain over time. Using pre-specified measure weights would require broad stakeholder agreement on which measures to weight more heavily, and this consensus might be difficult to achieve. In the example above, each hospital has the same measure scores but for different measures; because one hospital did better on higher-weighted measures, however, it has a notably higher summary score.

Stakeholder Feedback

Many TEP members felt this approach warranted further evaluation and consideration. TEP members noted that simplifying the methodology was beneficial for transparency and stakeholder understanding. However, other TEP members noted gaining consensus on measure contribution weights would be difficult, and that the best methodology should be used, regardless of complexity. These TEP members suggested more clear explanations around the methodology for stakeholders rather than simplifying it. Provider Leadership Work Group members were similarly interested in investigating the explicit approach; they also noted the benefit of a simplified methodology for better hospital understanding but acknowledged the challenge of establishing measure contributions.

Questions for the Public:

- What are the advantages and disadvantages of a more explicit approach to calculating Overall Hospital Quality Star Ratings?
- Is the explicit approach a worthwhile change in approach and direction to consider further?
- How could such an approach be best operationalized or sustained?

5.3. Clustering Alternative

Background

Currently the Overall Hospital Quality Star Rating methodology uses k-means clustering to assign each hospital to a discrete star rating category from the continuous distribution of summary scores. K-means clustering groups hospitals so that a hospital's score is closer to the average score of its own category than to that of any other category (that is, any 3-star hospital is more like an average 3-star hospital than it is an average 2- or 4-star hospital, and so on).

CMS originally used this approach to identify empiric rather than arbitrary cut points, accommodate changes in the underlying distribution of scores, and provide a comparative assessment for consumers.

However, some stakeholders have expressed concerns about k-means clustering, including:

- It limits hospitals' ability to predict cut points in future periods, or
- It results in star rating assignments that seem arbitrary for hospitals with borderline scores.

CMS seeks input from the public as to what alternatives might exist for grouping hospitals into star rating categories and to how to address these stakeholder concerns.

Questions for the Public:

- Should CMS consider potential alternatives to k-means clustering in more detail?
 - If so, what sort of change should CMS consider?
- What other considerations should guide future CMS work regarding clustering?

5.4. Incorporation of Improvement

Background

The Overall Hospital Quality Star Rating methodology is inherently comparative, due to the use of LVM and k-means clustering, and a hospital's performance is determined by its measure scores relative to those of other hospitals. As such, the Overall Hospital Quality Star Rating currently captures a hospital's improvement in measure scores in excess of other hospitals' improvement, but not necessarily relative to its own prior performance.

Some stakeholders have expressed interest in modifying the Overall Hospital Quality Star Rating methodology to account for a hospital's absolute improvement on measure scores compared to its performance in the prior period. However, at what step in the methodology or the degree to which improvement should be incorporated remains to be determined.

Stakeholder Feedback

In general, the Provider Leadership Work Group and the Patient & Patient Advocate Work Group did not support incorporating improvement into the Overall Hospital Quality Star Rating methodology. They felt that incorporating improvement based on data from previous years would not provide consumers with the most current data for decision-making. One Provider Leadership Work Group member expressed they wanted consumers to know if an organization had improved or not; another member suggested using an icon in the display of information to indicate improvement. Members of the Patient & Patient Advocate Work Group suggested alternative options for display, such as displaying historical trend information using icons, and making it optional for users to view this information. Patient & Patient Advocate Work Group members agreed that considerations need to be made whether trend information is appropriate as hospitals may change star ratings due to changes in their measure performance as well as changes relative to other hospital performance. This topic was not addressed with the TEP.

Questions for the Public:

- Should CMS consider incorporating improvement in future iterations of the Overall Hospital Quality Star Rating?
- What are conceptual benefits and risks of incorporating absolute score improvement into the Overall Hospital Quality Star Rating?
- How should CMS operationalize this topic?

5.5. User-Customized Star Rating

Background

In alignment with the consumer and patient focus of the Overall Hospital Quality Star Rating, CMS has considered the creation of a user-customizable star rating tool. This concept has been discussed in prior TEP meetings and work groups with generally positive response.

Currently, measure group weights are fixed (22% for the outcome groups and Patient Experience, 4% for the three process measure groups). This allows hospitals to be compared fairly, with the same emphasis given to each measure group across hospitals. However, some stakeholders have suggested that these weights may not match the priorities, preferences or values of all patients or consumers.

User-customized star ratings would allow *Hospital Compare* users to interactively set the weights of measure groups that are used to calculate hospital summary scores, and display ratings clustered based on those customized summary scores. This would allow users to prioritize domains of care that are more important to them and compare hospitals on the basis of that preference. The tool could provide a set of pre-determined default weights as a starting point for users who do not want to set their own weights. In addition, due to computational limitations, a limited number of possible combinations of group weight would be available.

For example, the tool could ask users to rate each measure group as 1 (not very important), 2 (somewhat important), or 3 (very important). With seven groups, there would be 3^7 or approximately 2,200 ways to calculate summary scores and as many possible groupings of ratings, all of which would be pre-calculated to allow for rapid display of results. The tool would use the user's selected weights to determine a summary score, as in the example in [Table 15](#) below.

Table 15. Example of User-Customized Measure Group Contributions

Group	Hospital score	User A's importance	User A's summary score	User B's importance	User B's summary score
Mortality	1.4	3-Very	$(3/17)*1.4$	1-Not very	$(1/14)*1.4$
Readmission	0.2	2-Somewhat	$(2/17)*0.2$	1-Not very	$(1/14)*0.2$
Safety of Care	0.7	3-Very	$(3/17)*0.7$	2-Somewhat	$(2/14)*0.7$
Patient Experience	1.2	3-Very	$(3/17)*1.2$	3-Very	$(3/14)*1.2$
Effectiveness	-0.2	2-Somewhat	$(2/17)*(-0.2)$	1-Not very	$(1/14)*(-0.2)$
Timeliness	0.5	3-Very	$(3/17)*0.5$	3-Very	$(3/14)*0.5$
Imaging Efficiency	0.0	1-Not very	$(1/17)*0.0$	3-Very	$(3/14)*0.0$
Total	n/a	17	0.671	14	0.465

In this example, User A's priorities led to a summary score of 0.671 while User B's priorities led to a summary score of 0.465 for the same hospital. Depending upon how other hospitals performed on the measures and the clustering of results, this facility may or may not receive a different star rating when User A and User B are choosing a hospital. However, the ratings they see will be aligned with their own priorities to a greater degree than a uniform set of weights might be.

The disadvantage of this approach is that without a uniform set of weights, hospitals may not be able to receive feedback and reports for the Overall Hospital Quality Star Rating as they do using the current methodology. Furthermore, while the Overall Hospital Quality Star Rating is intended primarily for consumers, some hospitals use their rating for quality improvement, and the lack of a uniform set of weights may diminish the utility of the

star ratings for this use. Hospitals could, however, continue to use the Overall Hospital Quality Star Rating for quality improvement by setting the weights to be consistent with their local quality strategies.

Stakeholder Feedback

In general, TEP members expressed interest and support for a user-customizable tool. TEP members cautioned that any tool should be thoroughly user tested to avoid confusing consumers. TEP members suggested ways to allow for customization, including allowing for setting of group weights, or selecting specific measures included in the rating to better allow for consumers to pin-point the type of care they were researching. One TEP member noted that providing the default Overall Hospital Quality Star Rating alongside the user-customized rating was important. Provider Leadership Work Group members expressed interest in the concept but had questions about how the user-customized star ratings would be operationalized. In contrast, Patient & Patient Advocate Work Group members expressed a mixed reaction to the concept of user-customized star rating; while some members felt this feature would be useful to consumers, others felt that personalization would add a level of complexity that may be confusing and burdensome to consumers. Some members suggested adding filters to *Hospital Compare*, allowing users to filter by hospital characteristics and location, as an alternative.

Questions for the Public:

- Should CMS consider introducing user-customization to the Overall Hospital Quality Star Rating?
- What is the usability, utility, and validity of such a tool?
- What are potential benefits and drawbacks to such a tool?
- How could CMS incorporate such a tool into the existing Overall Hospital Quality Star Rating methodology?

Appendix A: Glossary of Terms

Table A1: Glossary of Terms

Term	Definition
Closed-form solution	An alternative calculation approach to quadrature for solving LVM equations
Confidence interval	A metric of a measure score's precision; a smaller confidence interval indicates more precision and less uncertainty about the score
Dual-eligible patients	CMS defines proportion of dual-eligible patients as: the proportion of Medicare fee-for-service (FFS) and managed care stays where the patient was dually eligible for Medicare and full-benefit Medicaid ¹¹
Eigenvalue	In factor analysis: a number indicating the amount of variation attributable to a particular underlying factor
Factor analysis	A method to assess the presence and strength of underlying factors explaining variation in measures in a group
Harm-based weights	Used in PSI-90 measure. Components are weighted based on relative total harm, so measures of more harmful conditions are given more influence on the score
Loadings	Empirical estimates from LVM representing the contribution of each individual measure; a higher loading indicates measures that are more correlated with each other and with the underlying aspect of quality
Preview period	Time shortly before public release in which facilities can privately view their results
Quadrature (adaptive and non-adaptive)	The calculation approach used to estimate measure group scores in LVM.
Refresh	The update of measure scores on <i>Hospital Compare</i> to reflect newly available data
Scree plot	In factor analysis: a plot of eigenvalues used to qualitatively assess factor strength
Volume-based weights	Weighting of measures based on the volume of care giving rise to the measurement, so measures or hospitals with more volume get more influence
Weighted mean square error	The mean square error is the average of the errors, that is, the average squared difference between the estimated values and what is estimated (observed). A weighted mean square error multiplies the square error by the weight for each hospital, which is the same weight used in the latent variable model.

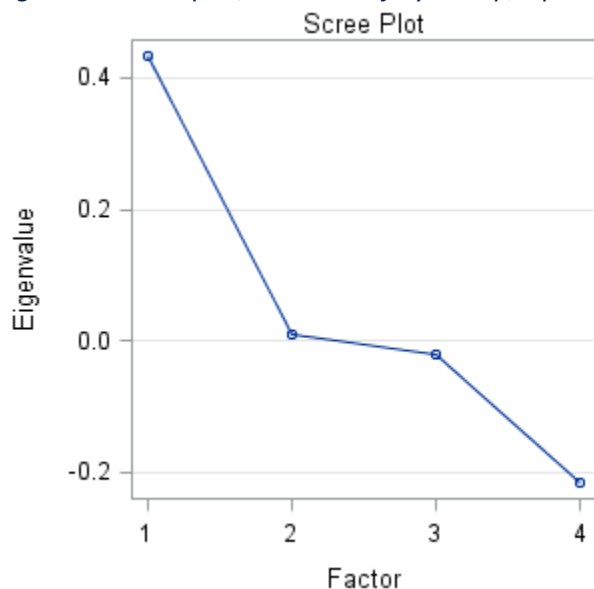
¹¹ Centers for Medicare & Medicaid Services. Hospital Readmissions Reduction Program (HRRP). https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Medicare_Beneficiaries_Dual_Eligibles_At_a_Glance.pdf. Accessed January 29, 2019.

Appendix B: Eigenvalues and Scree Plots, Safety of Care Regrouping

Option 1: Retain PSI-90

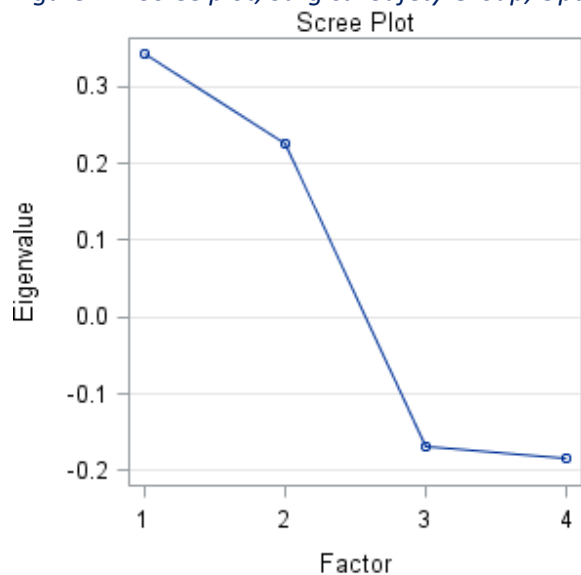
In the Medical safety group, the first two eigenvalues were 0.433 and 0.00855, a ratio of 51. The scree plot is shown in [Figure B1](#) below.

Figure B1: Scree plot, Medical Safety Group, Option 1 (retain PSI-90)



In the Surgical safety group, the first eigenvalues were 0.343 and 0.227, a ratio of 1.5. The Scree plot is shown in [Figure B2](#) below.

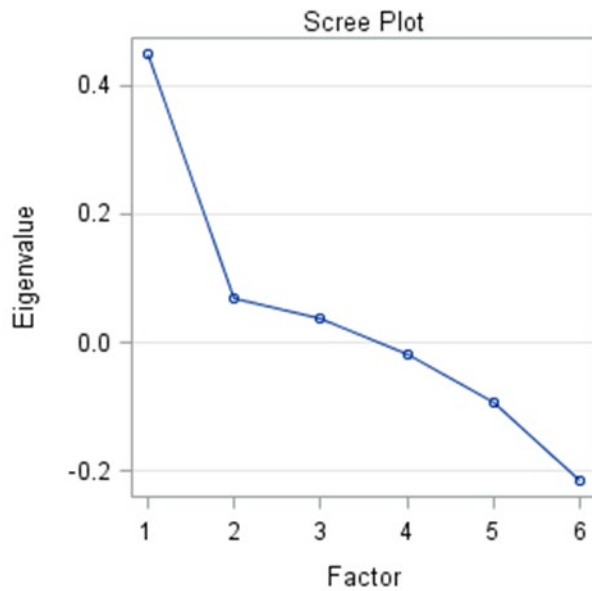
Figure B2: Scree plot, Surgical Safety Group, Option 1 (retain PSI-90)



Option 2: Switch to PSI components

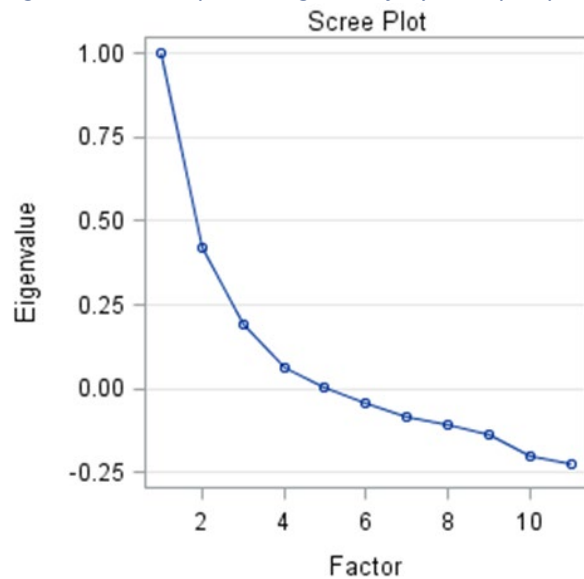
In the Medical safety group, the first two eigenvalues were 0.449 and 0.068, a ratio of 6.6. The scree plot is shown in [Figure B3](#) below.

Figure B3: Scree plot, Medical Safety Group, Option 2 (switch to PSI components)



In the Surgical safety group, the first eigenvalues were 1.00 and 0.419, a ratio of 2.4. The Scree plot is shown in [Figure B4](#) below.

Figure B4: Scree plot, Surgical Safety Group, Option 2 (switch to PSI components)



Appendix C: Estimating Parameters in the Latent Variable Model for Star Rating Group Scores through a Closed Form Solution

C.1. Overview

The Overall Hospital Quality Star Ratings methodology entails estimating latent variable models (LVMs) for each measure group in order to compute a group score for each hospital in that group. From the beginning, these LVMs have been estimated using Gaussian quadrature to maximize likelihood. This document describes two alternatives to quadrature for estimating the LVMs: an estimation approach of an EM (“expectation-maximization”) algorithm¹² and a closed form approach of maximizing log weighted likelihood (LWL). Both methods are faster, more accurate and easier to converge than the current Gaussian quadrature;¹³ going forward, we recommend that the second method based on a closed form be used to estimate Overall Hospital Quality Star Rating group scores.

Briefly, the EM consists of two iterative steps, each of which has a closed form expression; the steps are iterated until successive expectation values differ by less than some threshold value. The other approach is based on closed form expression for the LWL which we derived; this closed form can be maximized directly, without quadrature. The main difference between the current quadrature method and these two methods is that the former involves numerically integrating the latent variable and the latter two completely avoid numerical integration. Numerical integration is not only computationally intensive, but it risks convergence failure that is not a risk with either of the alternative approaches. The EM and the closed form maximization could be implemented in SAS through IML coding and PROC NLMIXED respectively.¹⁴

In this document, we first review the LVM, and display the LWL of the LVM. We proceed to describe the closed form expressions from the EM calculation and finally derive the closed form expression of LWL that can be maximized without quadrature.

C.2. LVM and Log Weighted Likelihood

The LVM is currently specified as follows--for j^{th} measure (type j) of hospital h , Y_{jh} , omitting group label, we use the following latent variable model:

$$Y_{jh} = \mu_j + \gamma_j \alpha_h + e_{jh},$$

where μ_j is the intercept for type j measure, γ_j is the measure loading of the unit normal latent variable α_h for hospital h , and e_{jh} is the error term that has a normal distribution with mean 0 and variance σ_j^2 . Measures indexed by j 's within a hospital in a given group share a same latent variable that represents the quality performance of the hospital.

¹² McLachlan G and Krishnan T (2008). The EM Algorithm and Extensions. 2nd Edition. John Wiley and Sons, Inc.

¹³ Pinheiro JC and Bates DM (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics* 4:12–35.

¹⁴ SAS Institute Inc. SAS/STAT® 9.2 User's Guide, Section Edition. April 2010.

<http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#titlepage.htm>. Accessed January 2019.

The objective in estimating the LVM is to obtain estimates that maximize the logarithm of weighted likelihood (LWL) of the LVM, which is given as:

$$\text{LWL} = \sum_h \log(\int \prod_j f(Y_{jh}|\alpha_h)^{w_{jh}} f(\alpha_h) d\alpha_h), \quad (1)$$

where $f(Y_{jh}|\alpha_h)$ denotes the density for the j^{th} measures of h^{th} hospital Y_{jh} conditional on the hospital h specific latent variable α_h , $f(\alpha_h)$ denotes the density for the latent variable α_h and w_{jh} is the weight for the measure Y_{jh} . Currently, the weight w_{jh} is specified as the denominator volume. Terms within the integrals of (1) involve the latent variable α_h , which poses the major computational challenge to any software. We currently use SAS PROC NLMIXED to perform numerical integration through the quadrature method. LWL is a marginal likelihood with the latent variable integrated out, and the integration often either has no closed form expression or is difficult to derive. SAS PROC NLMIXED provides a numerical quadrature method for calculating the integral regardless whether the integral has a closed form expression or not.

C.3. The EM Algorithm

The EM algorithm is a standard method for obtaining estimates for a LWL such as given in (1) by using the joint likelihood without integrating the latent variable. The joint weighted likelihood is same as the term within the logarithm in (1) without the integral; thus the EM algorithm avoids evaluation of the integral by replacing the latent variable with its closed form expectation in the log joint weighted likelihood (LJWL):

$$\sum_h \log(\prod_j f(Y_{jh}|\alpha_h)^{w_{jh}} f(\alpha_h)), \quad (2)$$

which is the same as (1) except for the absence of integration. The proof for this specification can be found in McLachlan and Krishnan (2008) that maximizing (2) through the EM results in the same estimates as maximizing (1) through numerical integration. That is, in order to maximize (1), we instead find parameters which maximize (2). The EM method maximizes (2) by iterating between the following two steps. In the E-step (“expectation” step) of EM, the latent variable α_h in (2) is substituted by its conditional expectation which has a closed-form expression:

$$\widetilde{\alpha}_h = E(\alpha_h|Y_{jh}, \forall j) = \frac{\sum_j w_{jh}(Y_{jh} - \mu_j)\gamma_j/\sigma_j^2}{1 + \sum_j w_{jh}\gamma_j^2/\sigma_j^2}. \quad (3)$$

The $\widetilde{\alpha}_h$ in (3) is also the group score estimate that has the variance estimate:

$$\text{Var}(\widetilde{\alpha}_h|Y_{jh}, \forall j) = (1 + \sum_j \frac{w_{jh}\gamma_j^2}{\sigma_j^2})^{-1}.$$

In the M-step (“maximization step”) of EM, the closed form estimates for the parameter of loadings γ_j , mean μ_j and error variance σ_j^2 are obtained by maximizing the (2) with α_h substituted by $E(\alpha_h|Y_{jh}, \forall j)$ in (3). Specifically, a score equation is obtained by taking first derivative of (2) in which $\widetilde{\alpha}_h$ substituted α_h with respect to each of the parameters, then the estimate is obtained as the solution to the score equation. Each solutions has a closed form expression. These two steps are iterated until success values of the expectation differ by some small threshold value.

C.4. Closed form maximization

The application of the EM algorithm allows us to obtain the closed form expression for (1) that is integral free and can be maximized with respect to the parameters without using quadrature. With some derivation it can be shown that (1) is proportional to:

$$\log\left(1 + \sum_j \frac{w_{jh} \gamma_j^2}{\sigma_j^2}\right) + \sum_j w_{jh} \log(\sigma_j^2) + \sum_j \left(\frac{w_{jh} (Y_{jh} - \mu_j)^2}{\sigma_j^2} - \frac{\left(\sum_j \frac{w_{jh} (Y_{jh} - \mu_j) \gamma_j}{\sigma_j^2} \right)^2}{1 + \sum_j \frac{w_{jh} \gamma_j^2}{\sigma_j^2}} \right) \quad (4)$$

Equation (4) is an integral free expression that is derived from (1). This closed form expression of LWL as (4) can be maximized directly, without quadrature by using for example SAS PROC NLMIXED.

C.5. Estimation

The EM algorithm and the closed form maximization both afford several advantages. First, each requires much less computational time than the quadrature method, namely in a few seconds rather than hours. Subsequent evaluation of the EM algorithm and closed form maximization also indicates higher precision, i.e., converging at the tolerance of 10^{-8} in comparison to the quadrature approach that converges at a tolerance between 10^{-3} and 10^{-4} , based on SAS setting. In addition, we have found that hospital scores estimated using the two approaches are close with estimates of loadings differing only in the third or fourth digit after the decimal place and that the estimated group scores differing below 10^{-5} .

Because the objective function is still the weighted likelihood of the LVM, local maxima may still exist and both the EM algorithm and closed form maximization of (4) requires initial values to ensure optimization. Though the EM algorithm is the most direct solution, it is challenging to implement in standard software packages, while the closed form maximization can be implemented directly in SAS or other software; for this reason we are proposing to use the closed form maximization to estimate the LVMs for Star Ratings.

Closed form maximization without quadrature of (4) through SAS PROC NLMIXED gives same results as the EM. No major discernable computational or analytic costs have been identified to using the EM or the closed form maximization approach.