Overall Hospital Quality Star Rating on *Hospital Compare*
Methodology Report (v3.0)

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

December 2017

Table of Contents

List of Tables

List of Figures

Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE) Project Team

Arjun K. Venkatesh, MD, MBA, MHS*	Project Lead
Susannah M. Bernheim, MD, MHS	Project Director
Li Qin, PhD	Lead Analyst
Haikun Bao, PhD	Supporting Analyst
Jaymie Simoes, MPH	Project Manager
Megan Wing, MA	Research Associate
Erica Norton, BS	Research Associate
Grace Glennon, MS	Research Associate
Rushi Shah, BS	Research Assistant II
Jeph Herrin, PhD*	Statistical Consultant
Haiqun Lin, MD, PhD	Statistical Consultant
Zhenqiu Lin, PhD	Analytics Director
Harlan M. Krumholz, MD, SM*	Principal Investigator

*Yale School of Medicine

Acknowledgements

This work is a collaborative effort, and the authors gratefully acknowledge and thank our many colleagues and collaborators for their thoughtful and instructive input. Acknowledgment of input does not imply endorsement of the methodology and policy decisions.

We would like to acknowledge and thank Lein Han, PhD, Kristie Baus, MS, RN, Reena Duseja, MD, and Cynthia Tourison at the Centers for Medicare & Medicaid Services (CMS) for their contributions to this work.

We would also like to acknowledge and thank the contribution and work from Kit Cooper, BS and Dawn Beard, RN from the Lantana Consulting Group, as well as Rachel Singer, PhD, MPH, MPA from NORC at the University of Chicago.

Additionally, we would like to acknowledge Hospital Consumer Assessment of Healthcare Providers Systems (HCAHPS), Centers for Disease Control & Prevention (CDC), Mathematica Policy Research (MPR), and statistical consultant from Harvard University, Sharon-Lise Normand, PhD.

We would also like to thank Lisa Suter, MD, Elizabeth Drye, MD, SM, Lori Geary, MPH, and Tamara Mohammed, MHA, CHE from CORE.

Furthermore, we would like to thank the National Partnership for Women and Families (NPWF) for their collaboration in improving patient and consumer outreach and involvement.

Finally, we would like to recognize the commitment and contributions of the members of our original and newly-convenedTechnical Expert Panel (TEP)s, our Patient & Patient Advocate Work Group, and our Provider Leadership Work Group (PLWG).

I. Executive Summary

This report presents the methodology for the Overall Hospital Quality Star Rating, developed by the Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE) under contract to the Centers for Medicare & Medicaid Services (CMS). This report describes CMS's approach to construct a methodology for generating an Overall Hospital Quality Star Rating for each eligible hospital publicly reporting quality information on *Hospital Compare*. This report represents the updated methodology, implemented in December 2017.

Overview of Project Objective

CMS contracted with CORE to work in collaboration with other contractors to develop a methodology for the Overall Hospital Quality Star Rating on *Hospital Compare*. *Hospital Compare* includes information on over 100 quality measures and more than 4,000 hospitals. The primary objective of the Overall Hospital Quality Star Rating project is to develop a statistically sound methodology for summarizing information from the existing measures on *Hospital Compare* in a way that is useful and easy to interpret for patients and consumers. Consistent with other CMS Star Rating programs, this methodology assigns each hospital between one and five stars, reflecting the hospital's overall performance on selected quality measures. As a secondary objective, hospitals are also classified according to performance on groups or domains of measures.

CMS intends for the Overall Hospital Quality Star Ratings to complement existing efforts, such as the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) star rating (implemented in April 2015), and will continue to report individual quality measures for stakeholders seeking more detailed information.

In what follows, "Star Rating" refers to Overall Hospital Quality Star Rating, unless otherwise noted. The Star Rating methodology was developed over two years and included substantial stakeholder input. This development work began with defining guiding principles.

Guiding Principles for Developing Star Ratings

Based on a systematic review of the literature,[1] lessons from prior star rating efforts, and the CMS quality strategy, CMS defined the following principles to guide the Star Rating work:

- Alignment with *Hospital Compare* and other CMS programs;
- Transparency of methodological decisions; and
- Responsiveness to stakeholder input.

CMS has sought to meet the third principle by assembling two multi-stakeholder Technical Expert Panels (TEP); holding three public input periods, two National Stakeholder calls, a hospital dry run; and convening a Patient and Patient Advocate Work Group, as well as a Provider Leadership Work Group (PLWG).

CMS designed several aspects of the Star Rating development process to include the patient and consumer perspective in key methodological and policy decisions. Both the TEP and Patient & Patient Advocate Work Group included diverse patient and patient advocate representation (Appendix B). These individuals were supportive of CMS's decision to develop a hospital quality star ratings system, expressing its potential value and importance to patients and consumers.

Overview of Methodology

The methodology takes a six-step approach to calculating the Star Rating. In the first step, the measures are selected based on their relevance and importance as determined through stakeholder and expert feedback, and the included measures are standardized to be consistent in terms of direction and magnitude. In the second step, these standardized measures are then organized into seven groups according to measure type. Third, for each group a latent variable model is used to estimate a group score for each hospital reporting measures in that group; as a secondary task objective, these groups scores are also used to classify hospitals into performance categories (above, same as, or below the national average) for each quality domain. In the fourth step, a weight is applied to each group score and all available groups are averaged to calculate the hospital summary score. In the fifth step, the public reporting threshold is applied, with hospitals reporting too few measures or groups excluded. Finally, to assign star ratings, hospital summary scores are organized into five ordered categories using a clustering algorithm.

Conclusion

The overarching goal of the Overall Hospital Quality Star Rating is to improve the usability and interpretability of *Hospital Compare* for patients and consumers. This report reflects the results of a year-long effort to re-evalaute the original methodology and engage stakeholders in a broad discussion about potential enhancements. The original methodology (V2.0), which was released at the inception of public reporting in July 2016, was similiarly a two-year effort that included multiples forms of stakeholder feedback. In the future, CMS will continue to regularly re-evaluate the Star Rating intiative, and this methodology may continue to be updated and revised as necessary.

II. Introduction

Project Objective

CMS contracted with CORE to work in collaboration with other contractors to develop the methodology for the Overall Hospital Quality Star Rating on *Hospital Compare*. *Hospital Compare* includes information on over 100 quality measures and more than 4,000 hospitals. The primary objective of the Overall Hospital Quality Star Rating project is to summarize information from the existing measures on *Hospital Compare* in a way that is useful and easy to interpret for patients and consumers through the development of a statistically sound methodology. Consistent with other CMS star rating programs, this methodology assigns each hospital between one and five stars, reflecting the hospital's overall performance on selected quality measures.

The Overall Hospital Quality Star Rating is designed to provide summary information for consumers about existing publicly reported quality data. CMS intends for the Overall Hospital Quality Star Rating to complement existing efforts, such as the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) star rating (implemented in April 2015), and will continue to report individual quality measures for stakeholders seeking more detailed information. Throughout the remainder of this report, "Star Rating" refers to the Overall Hospital Quality Star Rating, unless otherwise noted.

This report reflects changes since the v2.0 reported information including:

- Methodology enhancements and updated language;
- Updated validity and reliability calculations; and
- Updated hospital distributions and summary score ranges.

Why Develop Hospital Quality Star Ratings?

In 2014, CMS conducted a review of the literature and prior star rating efforts which supported the notion that patients care about quality information. The results also suggested that patients' use of this information is limited by low understanding of quality information and some inconsistency in the facets of quality that interest them most. Consumers need help understanding hospital quality information, and prefer information be presented in a more condensed and annotated manner to convey the many facets of quality. These key findings are consistent with consumers' priorities of bringing a wide variety of measures together into a single overall star rating, and also point to the need for extensive engagement and education of stakeholders throughout development and implementation.

In addition to patients' and consumers' informational needs, CMS developed the Overall Hospital Quality Star Rating methodology to complement the methodologies and goals of other CMS programs and star rating initiatives, including: Dialysis Facility Compare Star Ratings, Home Health Compare Quality of Patient Care Star Ratings, HCAHPS, Nursing Home Compare Star Ratings, Medicare Plan Finder Star Ratings, and Qualified Health Plans (QHPs) Quality Rating System (QRS).[2,3]

III. Overall Hospital Quality Star Rating Methodology

CMS considered various approaches for calculating the Overall Hospital Quality Star Rating, including simple or weighted averages of all the measures and more complex statistical approaches utilizing factor analysis and latent variable models. CMS evaluated each approach in the context of the project goals, stakeholder input, and timeline.

CMS sought to identify an approach that would:

- Generate a single, aggregate measure of available hospital quality information;
- Account for the heterogeneity of measures available (process, outcome, etc.);
- Account for the fact that different hospitals are reporting different numbers and types of measures;
- Accommodate changes in the included measures (for example, retirement of measures); and
- Utilize an evidence-based approach reflecting modern statistical methods that previously have been applied to health care.

To assist readers as they review this report, CMS has provided a glossary of statistical terms used when describing CMS's approach to calculating the Star Rating and conducting validity and reliability analyses (Appendix A).

The methodology calculates the Star Rating through a six-step process (Appendix C). These steps are listed below and are described in greater detail in subsequent sections.

Step 1: Selection and standardization of measures for inclusion in the Star Rating
Step 2: Assignment of measures to groups
Step 3: Calculation of latent variable model group scores
Step 4: Calculation of hospital summary scores as a weighted average of group scores
Step 5: Application of minimum thresholds for receiving a star rating
Step 6: Application of clustering algorithm to categorize summary scores into star ratings

The measures were first selected based on their relevance and importance as determined through stakeholder and expert feedback. The selected measures were standardized to be consistent in terms of direction and magnitude, with outlying values trimmed (Step 1). In Step 2, the measures were organized into seven groups by measure type. In Step 3, the standardized measures for each group were used to construct a latent variable statistical model that reflected the dimension of quality represented by the measures within the given group. Each of the seven statistical models generated a hospital-specific group score, which is obtained as a prediction of the latent variable. In Step 4, a weight was applied to each group score, and all available groups were averaged to calculate a hospital summary score. In Step 5, the public reporting threshold is applied, requiring hospitals to have a minimum of three measure groups (one of which must be an outcome group) with at least three measure in each of the three groups to be included in the final clustering step. Finally, in Step 6, to assign star ratings, hospital summary scores were organized into five categories using a clustering algorithm.

Of note, CMS also reports hospital performance at the measure group level using the results of Step 3, separately categorizing each of a hospital's available group scores into one of three group performance categories (above, same as, or below the national average). These performance categories provide additional details for patients and consumers comparing hospitals across the seven groups (Section IV).

Step 1: Measure Selection for Inclusion and Standardization

Introduction to Hospital Compare Measures

Hospital Compare includes measures that reflect a range of different dimensions of quality, from clinical care processes to measures focused on care transitions to measures of patients' experiences. The measures on *Hospital Compare* represent a variety of measure types, and cover a broad set of clinical conditions and care processes. Though not all measures reported on *Hospital Compare* were selected for inclusion in the Star Rating, the Star Rating includes a broad and diverse set of measures.

Criteria for Selecting Measures for the Overall Hospital Quality Star Rating

CMS vetted measure selection criteria with stakeholders through the original TEP and public input periods to ensure that the Star Rating captured the diverse aspects of quality represented by the measures on *Hospital Compare*.

All measures for acute care hospitals reported on *Hospital Compare*, as determined using the data reported in the CMS *Hospital Compare* downloadable data file, were included in the Star Rating.

Because the Star Rating is intended for acute care hospitals, CMS first omitted all measures on *Hospital Compare* that were specific to specialty hospitals (such as a cancer hospital or inpatient psychiatric facility) or ambulatory surgical centers prior to applying any measure selection criteria. With these measures omitted, the total number of measures eligible for inclusion in the Star Rating for December 2017 was 124 measures. The Star Rating measure selection criteria are presented in the subsequent text and in Figure 3.

Measure Selection Criteria

CMS used the following criteria to exclude measures from the Star Rating calculation:

1. Measures suspended, retired, or delayed from public reporting on *Hospital Compare*;
2. Measures with no more than 100 hospitals reporting performance publicly;
3. Structural measures;
4. Measures for which it is unclear whether a higher or lower score is better (non-directional);
5. Measures not required for Inpatient Quality Reporting (IQR) Program or Outpatient Quality Reporting (OQR) Program; and
6. Overlapping measures (for example, measures that are identical to another measure, or measures with substantial overlap in cohort and/or outcome).

Standardization of Measure Scores

Before combining measures into a score, each measure is first converted into a common scale. Hospital quality measure results include many different types of scoring information, ranging from time (e.g., median time in minutes from emergency department [ED] arrival to ED departure for admitted ED patients) to percentages (e.g., percentage of patients given antibiotics prior to surgery); quality measures also have two directions, with either “lower is better” (readmissions, mortality) or “higher is better” (use of aspirin for acute myocardial infarction [AMI]). Therefore, to enable the combination of information, CMS uses standardization to ensure all measure scores are in a common scale with a common direction. This does not change the measure information, just the scale for scoring in order to make it possible to combine the measures in the Star Rating calculation. Specifically, CMS standardizes a hospital’s score on each measure by calculating “Z–scores” for each measure, reversing the direction if necessary so that higher values were always ‘better’; the measure “Z–score” is the difference between an individual hospital’s score and the overall mean score for hospitals divided by the standard deviation across hospitals.

For example, in April 2015, OP-21 (Median Time to Pain Management for Fractures) had a national average performance of 55.6 minutes with a standard deviation of 17.75 minutes. In contrast, VTE-6 (Incidence of Potentially Preventable Blood Clots) had a national average of 7.23% with a standard deviation of 9.10%. After standardization and redirection, both measures had a mean score of 0 and standard deviation of 1 and both were reversed so that a higher standardized score indicates better quality. For an individual hospital with an OP-21 score of 65 minutes, the standardized score was -0.53, while the standardized score for a hospital with a score of 45 minutes was 0.602. Henceforth in this report, a measure score refers to the standardized measure score or “Z- score”.

CMS further Winsorizes the standardized measure score at the 0.125th percentile (Z= -3) and the 99.875 percentile (Z=3) of a Standard Normal distribution; thus, all standardized scores above 3 were set to be 3, and all standardized scores below -3 are set to be -3. This is done to avoid extreme outlier performance for which it is unclear if the reported measure score represented an extreme performance or potentially inaccurate reporting, as well as to avoid values that would make estimation technically challenging.

Step 2: Assignment of Measures to Groups

Approach to Grouping Measures

CMS evaluated several options for organizing quality measures into mutually exclusive conceptual groups with the TEP. Ultimately, CMS grouped measures into seven groups. The use of groups in the Star Rating is consistent with other CMS star rating initiatives (Nursing Home Compare Star Ratings, Medicare Plan Finder Star Ratings, and Dialysis Facility Compare). The Overall Hospital Quality Star Rating measure groups for December 2017 are:

- Mortality;
- Safety of Care;
- Readmission;
- Pateint Experience;
- Effectveness of Care;
- Timeliness of Care; and
- Efficient Use of Medical Imaging.

The rationale for these seven measure groups is as follows:

- The seven groups are aligned with the CMS Hospital Value-Based Purchasing (HVBP) program, the current categories on the *Hospital Compare* website, and other national quality initiatives.[4]
- The groups are clinically reasonable in that they capture common components of quality for which hospital quality is likely linked across measures.
- The groups allow for future measures to be added or removed from the Star Rating as they are updated on *Hospital Comapre*.

CMS conducted descriptive analyses to better understand the variability in hospital-level reporting of quality information. The average number of measures reported by hospitals using the December 2017 *Hospital Compare* dataset supported CMS's decision to assign measures to groups. Out of the 57 measures in the Star Rating for this reporting period, the average hospital reported 36 measures (Interquartile range: 21 to 50). The distribution of hospital quality reporting by group for the December 2017 reporting period is described in Appendix D.

The group names were finalized with input from the Patient & Patient Advocate Work Group, convened in collaboration with the National Partnership for Women & Families (NPWF), and previous CMS consumer testing.

For the number of measures in each group for December 2017, please refer to Section IV.

Step 3: Calculation of Group Scores using Latent Variable Models

Overview of Latent Variable Model (LVM)

CMS employed latent variable modeling (LVM) to estimate a group score for the dimension of quality represented by the measures in each group. CMS constructed a separate LVM for each group so that a total of seven latent variable models are used.

LVM is a statistical modeling approach which assumes each measure reflects information about an underlying, unobserved dimension of quality. LVM accounts for the relationship, or correlation, between measures for a single hospital. Measures that are more consistent with each other, as well as measures with larger denominators, have a greater influence on the derived latent variable. The model estimates for each hospital and each measure group the value of a single latent variable representing an underlying dimension of quality; this estimate is the hospital's group score.

Rationale for Using LVM

CMS considered the following assumptions and advantages of LVM prior to selecting this approach for calculating group scores.

Assumptions of Using LVM

- Each LVM assumes that each group reflects a single distinct underlying aspect of quality.
- Each measure contributes to exactly one group score even if it may potentially reflect more than one aspect of quality.
- Each included measure is a valid indicator of quality.

Advantages of LVM

- The LVM method is used for composite measures in healthcare quality literature.[5-7]
- LVM accounts for consistency of performance by giving more importance to measures that are correlated within a group.
- LVM accounts for missing measures by using all available information to generate a group score; hospitals with varying amounts of information can be accommodated in the model.
- The model can account for sampling variance, reflecting the differences in precision for each hospital's measure score as a result of differences in hospitals' volumes used to calculate each measure.

Although LVM is an accepted technique for summarizing individual indicators, CMS realized that this approach may be challenging to understand or replicate. LVMs can be difficult to estimate and may often require assumptions regarding model parameters such as the error structure. Nonetheless, CMS ultimately determined that the advantages of LVM outweighed the challenges: CMS determined that the use of LVM with minimal, reasonable assumptions could overcome any technical challenges presented during methodology development and testing. Furthermore, while the modeling technique may be difficult for patients and consumers to understand initially, CMS aimed to overcome this challenge by embedding multiple channels for stakeholder education throughout development and preparation for implementation of the Star Rating methodology.

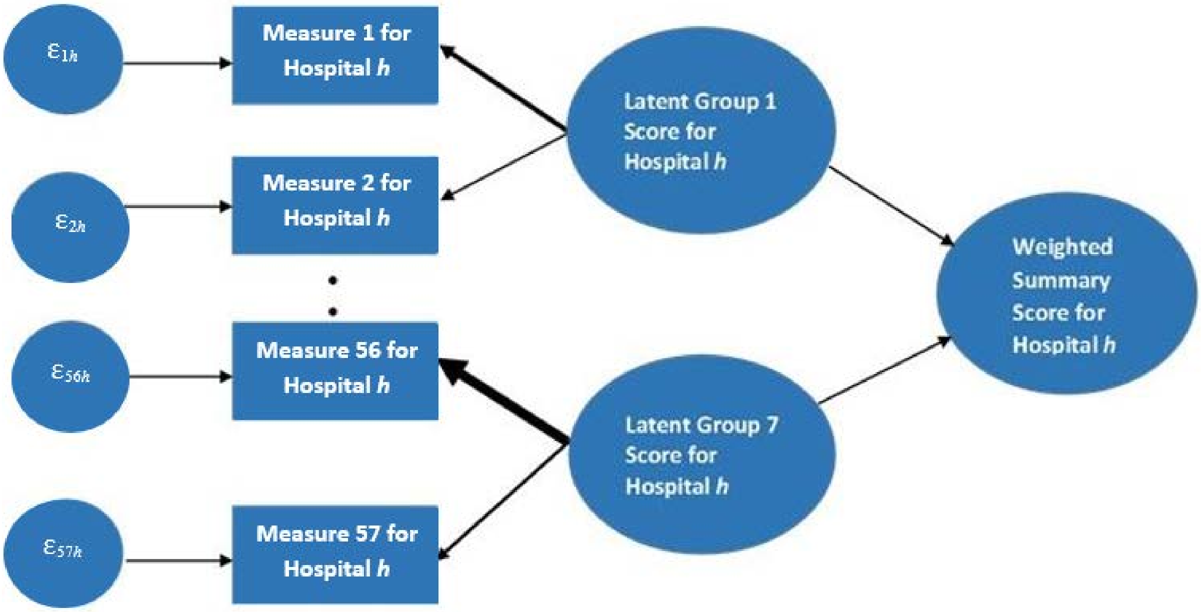
Detailed Description of LVM

In this section, CMS presents a sample path diagram for the LVM of each group in the Star Rating as well as the statistical equations used to calculate group scores. CMS constructed the LVM using the standard procedure, Proc NLMIXED in SAS software. All parameters in the models were obtained by maximum likelihood estimation (MLE), and the group scores were obtained as empirical Bayes estimates. The MLE process involves the numerical approximation of an integral for random effect, in which a quadrature technique is applied. This is done in a two-step process, where first non-adaptive Gaussian Quadrature is used for obtaining initial values of parameters, and then adaptive quadrature is used to strategically place quadrature points over the LVM integral range so fewer quadrature points are needed to achieve highly accurate estimations of parameters including group scores, resulting in stable star ratings.

Path Diagram

In the sample path diagram presented in Figure 1, the ovals represent the group scores and hospital summary scores. The group score is not directly observed but estimated from the models using the individual measures. The arrows between the group scores and each individual measure represent the relationship of that measure to the aspect of quality reflected by each measure with respect to the other measures in that group; each arrow has a different degree of association, also known as a "loading" or coefficient. The small circles on the left represent the residual error within each hospital for each of the measures included in the Star Rating. The residual error (ε) is the variation which could not be explained by the group score (random effect).

Figure 1. Sample Path Diagram of Group-Specific LVM


ε_{1h}
Measure 1 for Hospital *h*
ε_{2h}
Measure 2 for Hospital *h*
ε_{56h}
Measure 56 for Hospital *h*
ε_{57h}
Measure 57 for Hospital *h*
Latent Group 1 Score for Hospital *h*
Latent Group 7 Score for Hospital *h*
Weighted Summary Score for Hospital *h*

Statistical Equation for LVM

The LVM used to derive a hospital's group score (Equation 1) is as follows:

Equation 1. Latent Variable Model within Each Group, *d*

$$Y_{khd} = \mu_{kd} + \gamma_{kd}\alpha_{hd} + \varepsilon_{khd},\ k\text{=1,…,}N_d$$

$$\alpha_{hd} \sim N(0,1) \text{ and } \varepsilon_{khd} \sim N(0, \sigma_{kd}^2)$$

Let Y_{khd} denote the standardized score for hospital *h* and measure *k* in group *d*. α_{hd} is the hospital-specific group-level latent trait (random effect) for hospital *h* and group *d* and follows a Normal distribution with mean 0 and variance 1. The estimated value of α_{hd} will be used as a group score. γ_{kd} is the loading (regression coefficient of the latent variable) for measure *k*, which shows the relationship with the group score of group *d*. N_d is the total number of measures in group *d*. The assumption of unit variance here is an innocuous choice of units required to identify the parameter μ_{kd} and γ_{kd}.

Loadings of Measures within Each Group

As noted in the advantages of LVM (page 13), measures that are more consistent, or more correlated, with other measures within the group have a greater influence on the hospital's group score. The influence of an individual measure on the group score is represented by the measure's "loading."

A loading is produced for each measure in a group when estimating the LVM; these statistically estimated measure loadings are regression coefficients based on maximum likelihood methods using observed data and are not subjectively assigned. A loading reflects the degree of the measure's influence on the group score relative to the other measures included in the same group. Key considerations for measure loadings include:

- A measure's loading is specific to the measure, considering national performance on the measure and the measure's relationship to other measures in the group and the group's latent variable. It is the same for all hospitals reporting that measure.
- Measures with higher loadings are more strongly associated with the group score. These more "consistent" measures, in terms of hospital performance, give us more signal or information about a hospital's quality profile than measures with "random" performance. Loadings are estimated using maximum likelihood. If several measures all point consistently in one direction, but one points in the opposite direction, the outlier receives less loading.
- Large measure loadings do not directly imply that only a few measures "matter" towards the group score. However, measures with higher loadings do have a greater association (or 'impact') on the group score than measures with much lower loadings. There could be multiple measures with large loadings in one group. Measures that are reported by more hospitals with consistent performance will tend to have higher loadings, as they reflect a stronger "signal" of hospital quality.
- Given that CMS will re-estimate the loadings each time the Star Rating is updated, the loadings for an individual measure can dynamically change as the distribution of hospitals' performance on the measure and its correlation with other measures evolve over time.

Accounting for Measure Sampling Variation

Hospitals' reported measures may include different numbers of patients, depending on the measure. For each measure, some hospitals may report a score based on data from fewer cases while other hospitals report scores based on more cases, resulting in differing precision for each hospital's individual measure score. This variability in precision is usually known as "sampling variation."

CMS gives more weight to measure scores that are more precise by using a weighted likelihood method. This method (Equation 2) uses the hospital's measure denominator (hospital case count or sample size) to weight the observed value (hospital's individual measure score). A weighted likelihood ensures that a hospital with a larger denominator, or a more precise measure score, contributes more in calculating the measure loadings.

Equation 2. Weighted Likelihood for accounting for sampling variation within Each Group, d

$$L = \prod_{k=1}^{K} \prod_{h=1}^{H} (L(Y_{khd}))^{w_{khd}} \qquad w_{khd} = \frac{n_{khd}}{\sum_{h=1}^{N_{kd}} n_{khd}} \times N_{kd}$$

L is the likelihood function. N_{kd} is the total number of hospitals for measure *k* in group *d* and n_{khd} is the denominator for hospital *h* and measure *k* in group *d*. A hospital with larger denominator will be weighted more in the LVM. The specified weighted likelihood is maximized with respect to all the parameters in Equation 1.

Group Performance Categories

In addition to a hospital's star rating, CMS decided to report categorical group performance for each of a hospital's available (i.e., meeting the minimum threshold) groups. To assign each group score to a group performance category, CMS compares a hospital's group score to the national average group score. The LVM for each group produces a point estimate and standard error for each hospital's group score that CMS uses to construct a 95% confidence interval for each hospital's group score. CMS compares this 95% confidence interval to the national mean group score. CMS defines the group performance categories as follows:

- "Above the national average," defined as a group score with a confidence interval that fell entirely *above* the national average;
- "Same as the national average," defined as a group score with a confidence interval that included the national average; and
- "Below the national average," defined as a group score with a confidence interval that fell entirely *below* the national average.

Step 4: Weighted Average of Groups to Calculate Summary Scores

Approach to Developing the Weighting Scheme

After estimating the group score for each hospital and each group, CMS calculates a weighted average to combine the seven group scores into a single hospital summary score. CMS evaluated potential weighting options considering the following three criteria:

- Group Importance
  - The weight of outcome groups (Mortality, Safety of Care, and Readmission) should be greater than that of process groups (Effectiveness of Care & Timeliness of Care).
  - The weight of the Efficient Use of Medical Imaging group should take into account the limited population captured by these measures.
- Consistency with Existing CMS Policies and Priorities
  - The weights should align with the existing weighting schemes of other CMS programs to ensure consistent incentives.
  - The weights should reflect CMS's priorities as reflected in the CMS Quality Strategy.
- Stakeholder Input
  - The weights should reflect the prioritization of the groups by the TEP as well as feedback received during the public input periods, the Star Ratings dry run, and additional sources of patient and consumer feedback.

Final Weighting Scheme

To obtain stakeholder input, CMS surveyed the TEP asking them to rank the groups for the purposes of weighting. The final weighting scheme set by CMS incorporated the TEP's feedback and was vetted with other stakeholders through a public input period, the hospital dry run, and the Patient & Patient Advocate Work Group. The weights were brought back to the second TEP, the PLWG, and public input in 2017, where the majority of feedback warrented no change to the original weighting scheme.

Given the feedback and criteria set during development, CMS finalized a policy-based weighting scheme modified from that used for the HVBP program (Table 1). The statistical equation that uses these weights to calculate hospital summary scores is presented in Equation 3.

Table 1. Star Ratings Weighting by Group

Group	Star Ratings Weight
Mortality	22%
Safety of Care	22%
Readmission	22%
Patient Experience	22%
Effectiveness of Care	4%
Timeliness of Care	4%
Efficient Use of Medical Imaging	4%

Equation 3. Calculation of Hospital Summary Score from Group Scores

$$Summary\ Score_h = \frac{\sum_{d=1}^{7} W_d \alpha_{hd}}{\sum_{d=1}^{7} W_d}$$

Method for Re-weighting When Missing One or more Groups

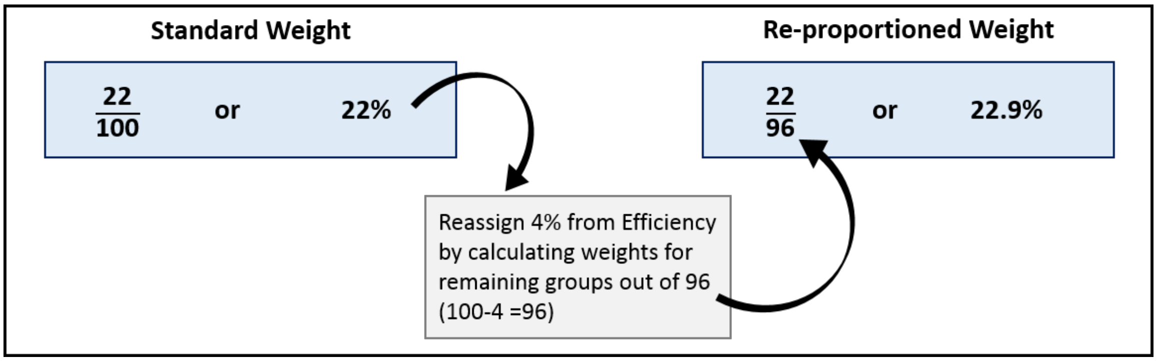
If a hospital reports no measures for a given group, CMS considers that group to be “missing.” When a hospital is missing one or more groups, CMS applies the HVBP program’s approach of re-proportioning the weight of the missing group(s) across the groups for which the hospital does report measures. Table 2 and Figure 2 provide examples of how the weighting scheme is adjusted for a hospital that is missing the Efficient Use of Medical Imaging group.

The final summary score for each hospital is the weighted average of that hospital’s group scores.

Table 2. Example Re-weighting Scheme for Hospital Missing Efficient Use of Medical Imaging Group

Group	Standard Weight	Re-proportioned Weight
Mortality	22%	22.9%
Safety of Care	22%	22.9%
Readmission	22%	22.9%
Patient Experience	22%	22.9%
Effectiveness of Care	4%	4.2%
Timeliness of Care	4%	4.2%
Efficient Use of Medical Imaging **(N=0)**	4%	0

Figure 2. Example Calculation for Re-proportioning Group Weights



Reassign 4% from Efficiency by calculating weights for remaining groups out of 96 (100-4 =96)

Step 5: Minimum Thresholds for Reporting a Star Rating

CMS aims to assign star ratings on the basis of adequate information regarding hospitals' quality. Thus, CMS evaluated and developed standards regarding the minimum number of measures and groups a hospital must report in order to receive a publicly reported star rating on *Hospital Compare*. CMS set these thresholds to allow for as many hospitals as possible to receive a star rating without sacrificing the validity and reliability of the Star Rating methodology. This step is applied prior to clustering hospitals into star categories; K-means clustering is inherently a comparative analytic procedure, and therefore conceptually only the subset of hospitals for which a star rating is ultimately reported should be clustered together.

Minimum Threshold of Measures per Group

CMS sets the minimum measure threshold at three measures per group. Setting a minimum measure threshold of three measures for each group exceeded a desired reliability level of 0.75 for all groups (Table 3). In Table 3, the "Required N" refers to the number of measures needed to meet the desired level of reliability (R).

Table 3. Minimum Measure Thresholds using Reliability Calculation and April 2015 Data

Group	Measures	Required N (for R =0.6)	Required N (for R =0.7)	Required N (for R=0.75)	Required N (for R=0.8)
Patient Experience	11	0.73	1.14	1.46	1.95
Readmission	7	1.21	1.89	2.43	3.23
Mortality	6	1.28	1.99	2.56	3.41
Safety of Care	8	1.14	1.78	2.28	3.05
Efficient Use of Medical Imaging	5	0.98	1.52	1.96	2.61
Effectiveness of Care	30	0.90	1.40	1.80	2.41
Timeliness of Care	8	0.80	1.24	1.60	2.13

Minimum Threshold of Groups in Summary Score

In addition to setting a minimum number of measures per group, CMS sets the final minimum group threshold for a hospital to receive a star rating at three groups, with at least one outcome group (that is, Mortality, Safety of Care, or Readmission). This minimum group threshold, requiring at least one outcome group, is similar to the eligibility requirements for hospitals to participate in HVBP.

After a hospital satisfies the minimum measure threshold for three groups (of which one must be an outcome group), any additional measures are included in the hospital's star rating, even if the total number of measures in the additional groups is fewer than three. This decision ensures that the Star Rating is inclusive of publicly reported measures and was vetted with the public through the second public input period, and again in the third public input period.

Results after Applying Minimum Thresholds

Together, both the minimum measure and group thresholds resulted in 78% of hospitals (N=4,746) receiving a star rating using the April 2015 dry run dataset. Since the initial public reporting in July 2016, between 78.9 and 80.3% of hospitals have received a star rating each reporting period. Setting increasingly higher thresholds for both measures and groups would exclude more hospitals from the Star Rating.

In the original methodology, the minimum measure and group thresholds were applied solely for reporting purposes and had no effect on the calculation of hospital summary scores. In the current methodology, application of the reporting threshold is conducted prior to k-means clustering to ensure that only hosptials meeting the threshold will be included in calculations and ultimately receive a star rating.

Step 6: Application of Clustering Algorithm to Obtain Star Ratings

Assumptions for Translating Summary Scores to Stars

Prior to selecting an approach for translating hospital summary scores to stars, CMS considered several important assumptions.

- Any approach selected will always result in some hospitals having summary scores at the boundary of two star categories (some hospitals will border a higher/lower star category).
- Similar to other CMS star rating efforts, a three-star rating will be considered "average."
- The objective of this project is to develop whole-star ratings (not half-stars).
- The Star Rating does not reflect an "apples to apples" comparison between hospitals (in other words, just because two hospitals may have the same star rating does not mean they have identical hospital quality). Rather, the star ratings reflect the weighted average of the summarized, group-level quality information available for a given hospital.
    - For example, there are many ways a hospital can be three stars. One hospital may do exceedingly well on the Process and Efficiency groups but perform poorly on Patient Experience. Another hospital with the same rating may do average across all available groups.
    - Because each hospital may have a different set of measures contributing to its star rating, patients and consumers should evaluate individual measure scores in addition to the star rating. Individual measure performance can be found for a given hospital on *Hospital Compare* at: hospitalcompare.hhs.gov.
- The Star Ratings is not intended to guide specific hospital quality improvement efforts, but rather to make summary information available to the public.

Overview of k-Means Clustering

CMS considered several approaches for translating summary scores to star ratings, including categorizing hospitals by percentile, setting statistical significance cutoffs, and using a clustering algorithm.

The Star Rating methodology utilizes *k*-means clustering to complete convergence. The *k*-means clustering analysis is a standard method for creating categories (or clusters) so that the observations (or scores) in each category are closer to their category mean than to any other category mean. The number of categories is pre-specified; CMS specified five categories, so that the *k*-means clustering analysis generates five categories based on hospital summary scores in a way that minimizes the squared distance between summary scores and their assigned category mean. Stated in another way, hospitals were organized into one of five categories such that a hospital's summary score is "more like" that of the other hospitals in the same category and "less like" the summary scores of hospitals in the other categories. The methodology runs clustering until complete convergence, in which the procedure iteratively examines solutions until it can find no better solution. The final star rating categories were structured such that the lowest group is one star and the highest group is five stars.

The rationale for the decision to use *k*-means clustering is as follows:

- *k*-means optimally designates five "means" for five star categories within the distribution of hospital summary scores. This minimizes the within-category and maximizes the between-category differences in summary scores.
- Hospitals in a cluster will have similar summary scores.
- In comparison to alternative approaches, such as quintiles, the *k*-means clustering approach produces a slightly broader distribution of star ratings.
- This approach is aligned with the similar clustering approach used to calculate the HCAHPS Star Ratings, also reported on *Hospital Compare*.

Results of CMS's analyses of the validity and the reliability of this approach are shown in Section V.

IV. Development Results of the Star Rating Methodology

This section describes results obtained throughout the development and reevaluation of the Star Rating methodology. CMS presents the national distribution of the Star Rating and results for each step of the methodology (measure selection, assignment to groups, group scores, summary scores, and star classification).

Distribution of Star Ratings (December 2017)

CMS calculated the Star Rating for December 2017 using the December 2017 *Hospital Compare* dataset. The frequency of hospitals by each Star Rating category is shown in Table 4.

Table 4. Distribution of Star Ratings for December 2017

Star Rating	Number of Hospitals (% Total)
★★★★★	337 (9.1%)
★★★★	1,155 (31.3%)
★★★	1,187 (32.2%)
★★	753 (20.4%)
★	260 (7.0%)

Measure Selection

Figure 3 depicts the process of measure selection for December 2017. Out of a possible 124 eligible for inclusion in the Star Rating, 67 measures were excluded based on our selection criteria.

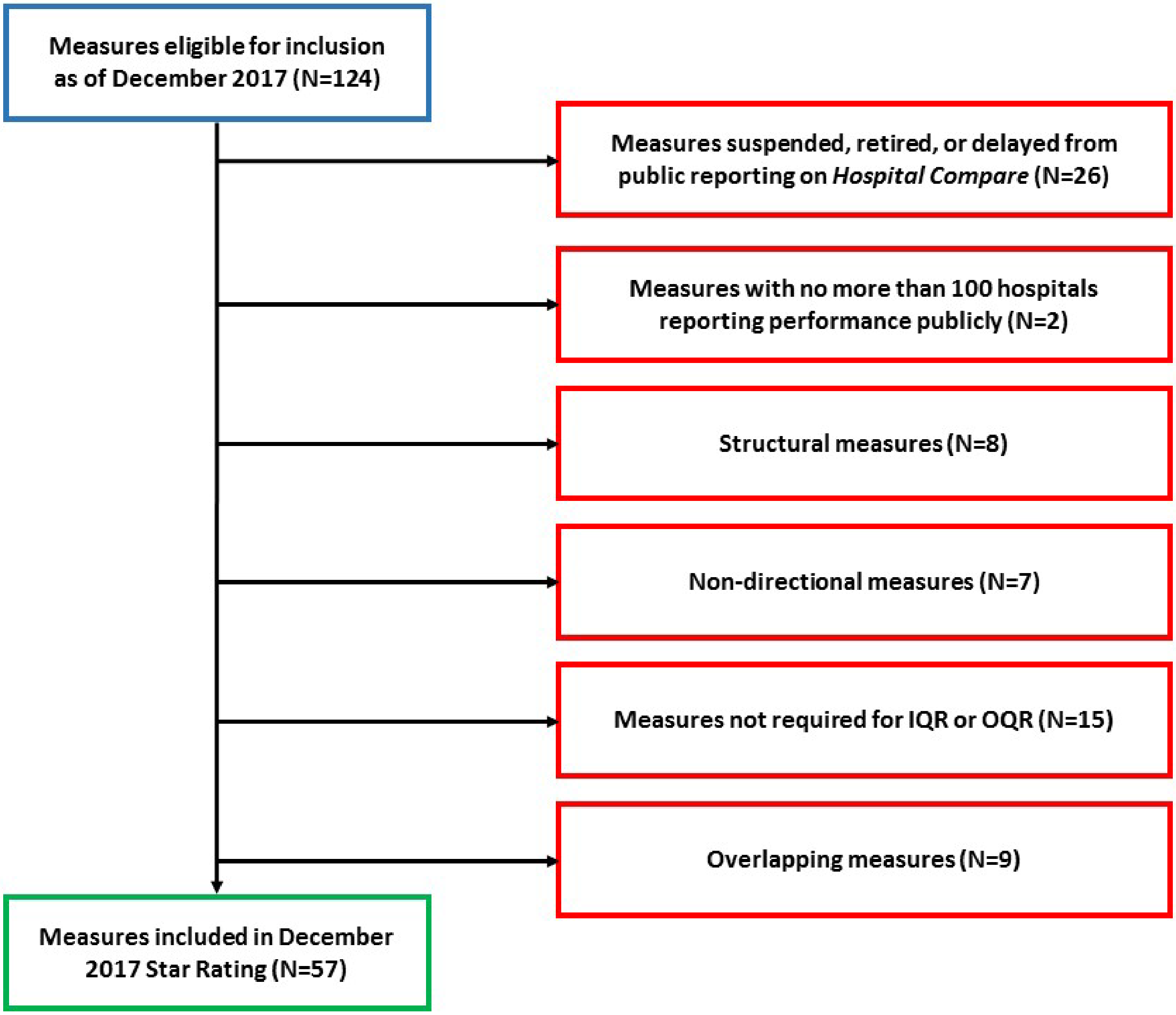
Assignment to Groups

Table 5 displays the number of measures assigned to each group, out of the total 57 measures included in the Star Rating for December 2017. The complete list of measures by name per group can be found in the Overall Hospital Quality Star Rating Quarterly Updates and Specifications Report: December 2017, available on *QualityNet* at www.qualitynet.org > Hospitals – Inpatient > Hospital Star Ratings.

Table 5. Total Number of Measures by Group for December 2017

Group	Number of Measures (N=57)
Mortality	7
Safety of Care	8
Readmission	9
Patient Experience	11
Effectiveness of Care	10
Timeliness of Care	7
Efficient Use of Medical Imaging	5

Figure 3. Measure Selection Flowchart (December 2017 Data)


Measures eligible for inclusion as of December 2017 (N=124)
Measures suspended, retired, or delayed from public reporting on *Hospital Compare* (N=26)
Measures with no more than 100 hospitals reporting performance publicly (N=2)
Structural measures (N=8)
Non-directional measures (N=7)
Measures not required for IQR or OQR (N=15)
Overlapping measures (N=9)
Measures included in December 2017 Star Rating (N=57)

Group Scores

Group Peformance Category Results for December 2017

Table 6 displays the frequency of hospitals in each group performance category for December 2017.

Table 6. Frequency (Number of Hospitals) by Group Performance Category

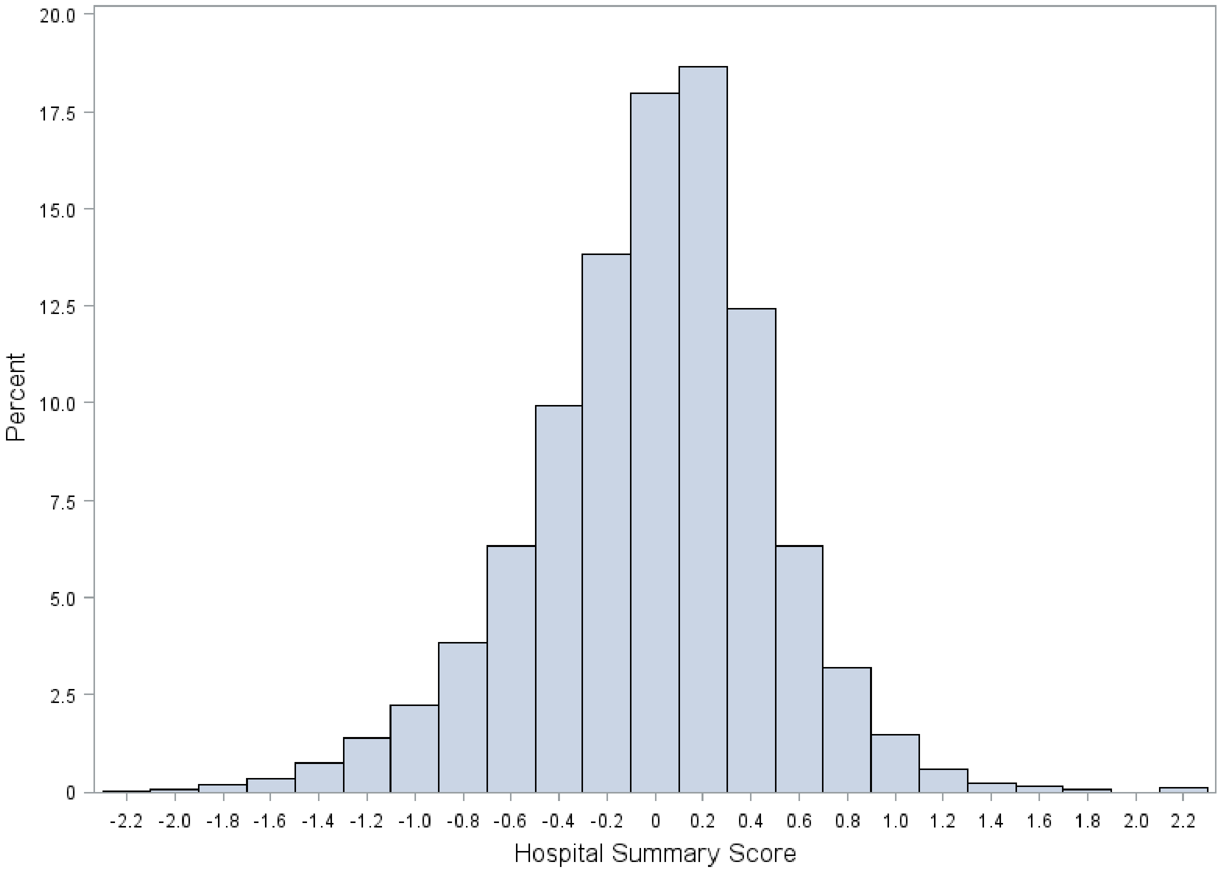
Group	Above the National Average	No Different than the National Average	Below the National Average
Mortality (N=3,444)	389 (11.3%)	2,703 (78.5%)	352 (10.2%)
Safety of Care (N=2,644)	1,148 (43.4%)	584 (22.1%)	912 (34.5%)
Readmission (N=3,892)	1,537 (39.5%)	977 (25.1%)	1,378 (35.4%)
Patient Experience (N=3,485)	1,213 (34.8%)	1,152 (33.1%)	1,120 (32.1%)
Effectiveness of Care (N=3,485)	70 (1.9%)	3,421 (91.5%)	250 (6.7%)
Timeliness of Care (N=3,688)	1,159 (31.4%)	1,533 (41.6%)	996 (27.0%)
Efficient Use of Medical Imaging (N=2,925)	502 (17.2%)	1,988 (68.0%)	435 (14.9%)

Note: The total number of hospitals in the *Hospital Compare* dataset as of December 2017 was 4,579 hospitals. The table shows the results for the 3,692 hospitals that met the reporting criteria.

Hospital Summary Scores

Figure 4 presents the distribution of hospital summary scores (N=3,692) for December 2017. The bars represent the percentage of hopsitals (y-axis) with a given hospital summary score (x-axis).

Figure 4. Distribution of Hospital Summary Scores for December 2017



Star Classification

Table 7 shows the frequency (number of hospitals) in each of the five star categories for December 2017. In addition, CMS displays the range of summary scores captured by each star category and the mean and standard deviation of hospital summary scores for each category.

Table 7. Frequency of Hospitals by Star Category using k-Means for December 2017

Rating	Frequency of Hospitals	Summary Score Range in Category	Mean (sd)
1 Star	260 (7.0%)	-2.12,-0.78	-1.09 (0.27)
2 Star	753 (20.4%)	-0.77,-0.26	-0.47 (0.14)
3 Star	1,187 (32.2%)	-0.26,0.13	-0.05 (0.11)
4 Star	1,155 (31.3%)	0.13,0.56	0.31 (0.12)
5 Star	337 (9.1%)	0.56,2.21	0.83 (0.26)

Note: The total number of hospitals in the *Hospital Compare* dataset for December 2017 was 4,579 hospitals. The table shows the results for the 3,692 hospitals that met the reporting criteria.

V. Validity & Reliability of the Star Rating Methodology

In this section, CMS presents their approach and results of testing the validity and reliability of the Star Rating methodology.

In order to confirm the validity of CMS's approach, CMS tested a number of key assumptions. Described in detail in the subsequent text, CMS tested the underlying assumption of LVM that each group represents a single latent variable. CMS also tested for meaningful differences in performance across the star rating categories and the relationship between these categories' and both groups and summary scores.

Validity Testing

Assumption of Single Latent Variable per Group

For the Star Rating, CMS assumed that each group conveys information about a single latent quality trait that corresponds with the type of measures included in the group. In other words, measures in one group convey information about one aspect of hospital quality. CMS conducted a confirmatory factor analysis during initial development using the April 2015 data to confirm that there is a single latent trait (a single dimension of quality) in each group. The clinical assumption that the measures in one group represent a single dimension of quality (one factor) held true for all groups except the Efficient Use of Medical Imaging group, which appeared to include more than one latent trait. Appendix E illustrates the factor analysis results graphically using scree plots.

Despite the empirical evidence that the efficiency measures might reflect more than one aspect of latent quality, CMS maintained the group for the Star Rating as a result of stakeholder support for the inclusion of these measures and the face validity of our clinically defined groups. Stakeholders felt that the efficiency measures collectively conveyed important information about hospital's utilization of resources and focus on patient safety.

Pairwise Correlation between Star Ratings

To test the validity of the star ratings generated using *k*-means clustering to complete convergence, CMS conducted an analysis that describes the distribution of group scores for hospitals in each star rating category using December 2017 data. In particular, CMS tested, using Tukey's method for multiple comparisons, the association between the mean group score for one star rating category and the mean group score for each of the other star rating categories for each group. This validation analysis demonstrated statistically different group scores between each star rating category in many groups, supporting the ability of *k*-means clustering to distinguish hospital performance across the five clusters (See Pairwise Comparison of Mean Group Scores between Star Classes).

CMS found statistically significant differences in most comparisons of group scores between each star rating category. Group scores were statistically different between every star rating category for the Safety of Care, Readmission, and Patient Experient groups and between the majority of star rating comparisons for the Mortality (9 out of 10), Effectiveness of Care (9 out of 10), Timeliness of Care (8 out of 10) and Efficient Use of Medical Imaging (5 out of 10) groups.

Test for the Linear Trend of Group Scores

To further confirm the validity of the Star Rating methodology, CMS examined the relationship across the mean groups scores for each star rating category, comparing the slope across the mean group scores for each category to zero. CMS found a statistically significant (p<0.0001) linear trend for each group, except the Efficient Use of Medical Imaging group (p = 0.20) during initital development with the April 2015 data. Due to changes in star rating distributions with the methodology enhancements, CMS re-assessed this using December 2017 data and found a statistically significant (p<0.0001) linear trend for each group including the Efficient Use of Medical Imaging group; the higher the group score, the higher the hospital's star rating. The results of this analysis, using December 2017 data, are depicted as box plots by group in Appendix E.

Reliability Testing

To confirm the reliability of the Star Rating methodology, CMS tested the stability of the Star Rating, both over time and across simulations of data within the same time period. First, CMS compared the Star Rating across time intervals and summary scores across time intervals (comparing across distinct quarters of reporting). Next, CMS conducted re-classificaiton analyses that tested: 1) the reliability of *k*-means for organizing summary scores into star ratings; and 2) the reliability of LVMs for calculating group scores, used to organize hospitals into group performance categories.

Reliability of Star Ratings over Time (Quarter-to-Quarter Stability)

To assess the reliability of the star ratings and the summary scores across different time intervals, CMS calculated kappa coefficients [8] and Intraclass Correlation Coefficients (ICC [2, 1]) in prior quarters.[9] Cohen's kappa coefficient was used to measure the agreement of star ratings between different quarters. The ICC score was used to determine the extent to which assessments of a hospital using data during different time periods produces similar summary scores.

CMS calculated summary scores and star ratings using April 2015 and July 2015 *Hospital Compare* data. The kappa coefficient was 0.45 (comparing to 1 if complete agreement), which indicates moderate agreement of Star Ratings between these two quarters. The Intraclass Correlation Coefficient was 0.91, which indicates substantial correlation of summary score between these two quarters.

As an additional test of stability, CMS calculated the concordance correlation for October to December 2016, which was found to be 0.99, indicating substantial correlation of summary scores between these two quarters.

Future Quarterly Update Specification Reports will update these analyses using the December 2017 data for comparison, as the analyses require comparing across quarters that utilize the same methodology.

Re-classification Analysis of Star Ratings

To evaluate the relaibilty of *k*-means clustering, CMS used simulations to calculate 5,000 summary scores for each hospital. In each simulation, each hospital's summary score was randomly selected from a normal distribution with a mean equal to the hospital's summary score and the summary score standard error using December 2017 data. Within each simulation, CMS reclassified the hospitals into five star categories, fixing the range of summary scores included in each category based on the results of *k*-means clustering for the December 2017 reporting period. CMS examined the proportion of the simulations that reclassified hospitals into the same star category as they were assigned during the December 2017 reporting. The higher the proportion, the higher the reliability of the classification step of the methodology.

In Table 8, The percentage in each cell represents the proportion of hospitals with the assigned star rating during December 2017 (the first column) that were re-classified into the corresponding star categories using 5,000 simulations. For example, in 5,000 simulations, 79.51% of the time, the hospitals classified as two-star during the December 2017 reporting period maintained two stars throughout the 5,000 simulations.

Table 8. Classification Analysis of k-Means Clustering using December 2017 Data

December 2017 Rating	Re-classify as 1-Star	Re-classify as 2-Star	Re-classify as 3-Star	Re-classify as 4-Star	Re-classify as 5-Star
1	78.04	21.93	0.03	0	0
2	2.82	79.51	17.6	0.07	0
3	0.09	9.81	76.6	13.29	0.22
4	0	0.3	23.57	73.51	2.61
5	0	0	0.06	32.92	67.02

Re-classification Analysis of Group Performance Categories

In addition to checking the reliability of the Star Rating through a re-classification analysis, CMS sought to similarly evaluate the reliability of the classficiation method for the group performance categories. For each group, CMS simulated group scores 5,000 times for each hospital based on the hospitals' group score and their standard error estimated from the LVM. Then, CMS reclassified the hospitals into three group performance categories (Above the National Average, Same as the National Average, and Below the National Average) by comparing the simulated 95% confidence interval to the simulated mean group score (correlation between simulated value and mean of simulated data is considered). CMS examined the proportion of the simulations that reclassified hospitals into the same group performance category as they were assigned during the December 2017 reporting period (Table 9). The higher the proportion, the higher the reliability.

Table 9. Reclassification Analysis of Group Performance Categories using December 2017 Data

Group	Above the National Average	Same as the National Average	Below the National Average
Mortality	82.76%	86.63%	79.33%
Safety of Care	94.91%	79.94%	96.22%
Readmission	94.19%	81.13%	95.84%
Patient Experience	90.57%	80.23%	92.71%
Effectiveness of Care	65.5%	85.74%	78.5%
Timeliness of Care	83.84%	79.37%	93.19%
Efficient Use of Medical Imaging	72.65%	81.68%	88.98%

Summary of Testing

The analyses conducted by CMS supported several key assumptions as well as confirmed that methodology enhancements in v3.0 maintain or improve several measures of validity and reliability sought by CMS for public reporting.

The underlying assumption that each of the Star Rating measure groups represent a single latent quality trait was supported for all but one group. In addition, statistically significant differences exist between most group scores when compared between star rating categories. Moreover, a statistically significant linear trend exists at the group score-level across star rating categories for all groups, indicating that star ratings increase as hospital group scores increase.

Both hospitals' star ratings and summary scores proved reliable over time. Furthermore, within the same performance period, the reclassification rate (hospitals being re-classified into their original star rating category and group performance category) for all star rating and group performance categories demonstrated strong reliability (R>0.65).

As the distribution of hospital performance evolves and/or updates to the Star Rating methodology are continued to be made, CMS will continue to test the methodology and potential improvements to the Overall Hospital Quality Star Ratings.

References

1. Venkatesh AV, Hsieh A, Potteiger J, et al. Ad Hoc Analysis Report 3: Star Ratings Hospital Quality Star Ratings on Hospital Compare Methodology Report: Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluate (YNNHHSC/CORE); 2014.
2. Dialysis Facility Compare (DFC) star ratings and data release. 2015. at https://www.cms.gov/Newsroom/MediaReleaseDatabase/Fact-sheets/2015-Fact-sheets-items/2015-01-22.html.)
3. (CMS) CfMMS. Quality of Patient Care Star Ratings Methodology. 2010.
4. (CMS) CfMMS. Hospital-Value Based Purchasing. 2014.
5. Landrum M, Bronskill S, Normand S-L. Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers. Health Services and Outcomes Research Methodology 2000;1:23-47.
6. Henderson CR. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 1975;31:423-47.
7. Shwartz M, Ren J, Pekoz EA, Wang X, Cohen AB, Restuccia JD. Estimating a composite measure of hospital quality from the Hospital Compare database: differences when using a Bayesian hierarchical latent variable model versus denominator-based weights. Med Care 2008;46:778-85.
8. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960;20:37-46.
9. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.

Appendix A: Introduction to Statistical Terminology

In this Appendix, CMS defines the statistical terms relevant to this report. CMS intends for this section to help streamline communication and develop a common, foundational understanding of the approaches and analyses discussed.

Table A.1. Glossary of Key Terms

Term	Definition/Explanation
Standardization	The process of converting an individual score into a dimensionless quantity. The standardized score is the number of standard deviations an individual score is above or below the average score. This process may also be referred to as normalizing.
Winsorization	A typical strategy used to set all outliers to a specified percentile of the data; for example, a 99% Winsorization would set all data below the 0.5th percentile to the 0.5th percentile, and data above the 99.5th percentile set to the 99.5th percentile.
Weighting	Weighting considers the influence or importance of a component relative to the whole. Unequal weighting implies that some quantities contribute more than others.
Loading	A loading in structural equation modeling (SEM) is the regression coefficient between an indicator (measure) and its factor (group score). It indicates the strength of the relationship between the latent variable and the indicator(s).
Group	A subset of measures believed to be conceptually or empirically similar.
Summary score (latent variable)	An assumed, but unobserved, quantity that reflects some latent trait.
Quadrature	A statistical approach that seeks to obtain the numerical estimate of an integral by placing optimal points at which to evaluate the integral.

Appendix B: Stakeholder Roster

Table B.1. Technical Expert Panel (TEP) #1 Roster

TEP Member	Title
Matt Austin, PhD	Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University *(Assistant Professor)*
Vinita Bahl, DMD, MPP	Performance Assessment & Clinical Effectiveness, University of Michigan Health System *(Director)*
John Bott, MBA, MS	Consumers Union/Consumer Reports (*Measurement Consultant*); State of Wisconsin Department of Employee Trust Funds (*Manager of Performance Measurement*)
Kathy Ciccone, RN, MBA	Healthcare Association of New York State Quality Institute (*Executive Director*)
Kelly Court, MBA	Wisconsin Hospital Association (*Chief Quality Officer*)
Rachel Grob, PhD	Center for Patient Partnerships, University of Wisconsin-Madison (*Director of National Initiatives / Associate Clinical Professor*)
Rodney Hayward, MD	University of Michigan (Professor of Public Health and Internal Medicine, Director of the Robert Wood Johnson Foundation Clinical Scholars Program)
Emma Kopleff, MPH	National Partnership for Women & Families (*Senior Policy Advisor*)
Doris Peter, PhD	Consumer Reports Health Ratings Center (*Director*)
Laura Petersen, MD, MPH	Michael E. DeBakery VA Medical Center (*Associate Chief of Staff for Research*)
Casey Schwarz, JD	Medicare Rights Center (*Policy & Client Services Counsel*)
David Shahian, MD	Center for Quality and Safety, Massachusetts General Hospital (*Vice President*)
Brett Stauffer, MD, MHS	Clinical Decision Support, Baylor Scott & White Health (*Director*)
Guofen Yan, PhD	University of Virginia School of Medicine (*Associate Professor*)
Ben Yandell, PhD	Clinical Information Analysis, Norton Healthcare (*Associate Vice President*)

Table B.2. Technical Expert Panel (TEP) #2 Roster

TEP Member	Title
Andrew Amster, MSPH	Kaiser Permanente, Department of Care and Service Quality *(Senior Director of Center for Healthcare Analytics)*
Karl Bilimoria, MD, MS	Feinberg School of Medicine, Northwestern University *(Director of Surgical Outcomes and Quality Improvement)*
Larry Boress, MPA	Midwest Business Group on Health *(Chief Executive Officer);* National Association of Worksite Health Centers *(Executive Director)*
John Bott, MSSW, MBA	Consumer Reports *(Manager of Healthcare Ratings)*
Angelo Bufalino, PhD	Ascension Healthcare, Care Excellence *(Senior Director of Analytics)*
Steven Donnelly, PhD	HealthInsight *(Healthcare Analyst)*

TEP Member	Title
Cynthia Dunlap, DNP, RN, NEA-BC, FACHE	Texas Hospital Association Foundation *(Vice President of Clinical Initiatives and Quality)*
Rachel Grob, MA, PhD	Center for Patient Partnership, University of Wisconsin-Madison Law School *(Director of National Initiatives)*
Rodney Hayward, MD	University of Michigan and Ann Arbor VA Healthcare System *(Professor of Medicine and Public Health)*
Jonathan Jennings, MBA, RN	Hospital Corporation of America Corporate Office *(Assistant Vice President of Cardiovascular and Neuroscience Services)*
Maria de Jesus Diaz-Perez, PhD	Center for Improving Value in Health Care *(Quality Measures Program Manager)*
Paul Kallaur, MS	Center for the Study of Services/Consumers' CHECKBOOK *(Vice President of Surveys and Research)*
Casey Schwarz, JD	Medicare Rights Center *(Policy & Client Services Counsel)*
David Shahian, MD	Massachusetts General Hospital *(Vice President of Center for Quality and Safety);* Harvard Medical School *(Professor of Surgery)*
Brock Slabach, MPH, FACHE	National Rural Health Association *(Senior Vice President)*
Julie Wall, RN, MBA, FACMPE	Benefits Health System *(System Vice President of Quality & Patient Safety)*
Guofen Yan, PhD	University of Virginia, Department of Public Health *(Associate Professor of Biostatistics)*

Table B.3. Provider Leadership Work Group Roster

PLWG Member	Title
Maureen Dailey, PhD, RN, CWOCN, FAAN	Dailey Solution *(Quality Consultant)*
Nancy Foster, PhD, APRN	American Hospital Association *(Vice President for Quality and Patient Safety Policy)*
Anne Clouatre, MHS, EMT-P, CPXP	Centura Health, Littleton Adventist Hospital *(Director of Patient Experience and Service Excellence)*
Brian Waterman, PhD(c), MPH	Missouri Hospital Association *(Vice President of Analytic Business Solutions)*
Gayle Olano Hurt, MPA, CPHQ, PMC	Sheppard Pratt Health System *(Director of Data Management, Outcomes, Measurement, and Research Administration)*
James Stuart Wolf, MD, FACS	Dell Medical School, Department of Surgery and Peri-Operative Care *(Associate Chair for Clinical Integration and Operations)*
Nidia Williams, PhD, MBB, CPHQ	Lifespan Corporation *(Administrative Director of Operational Excellence)*
Sandi Hyde, MSPS	Lifepoint Health *(Senior Manager of Quality Data and Learning)*
Rhonda Anderson, RN, DNSc, FAAN, FACHE	RMA Consulting *(Healthcare Consultant);* Global Healthcare Accreditation *(Accreditor)*
Troy Williams, RN, MSN	Zuckerberg San Francisco General Hospital *(Chief Quality Officer)*
Ugochukwu Uwaoma, MD, MBA, MPH, FAC	Multicare Health System *(Chief Medical Officer and Physician)*
Bhargavi Degapudi, MD	Atlantic Care *(Medical Director for Care Transitions)*

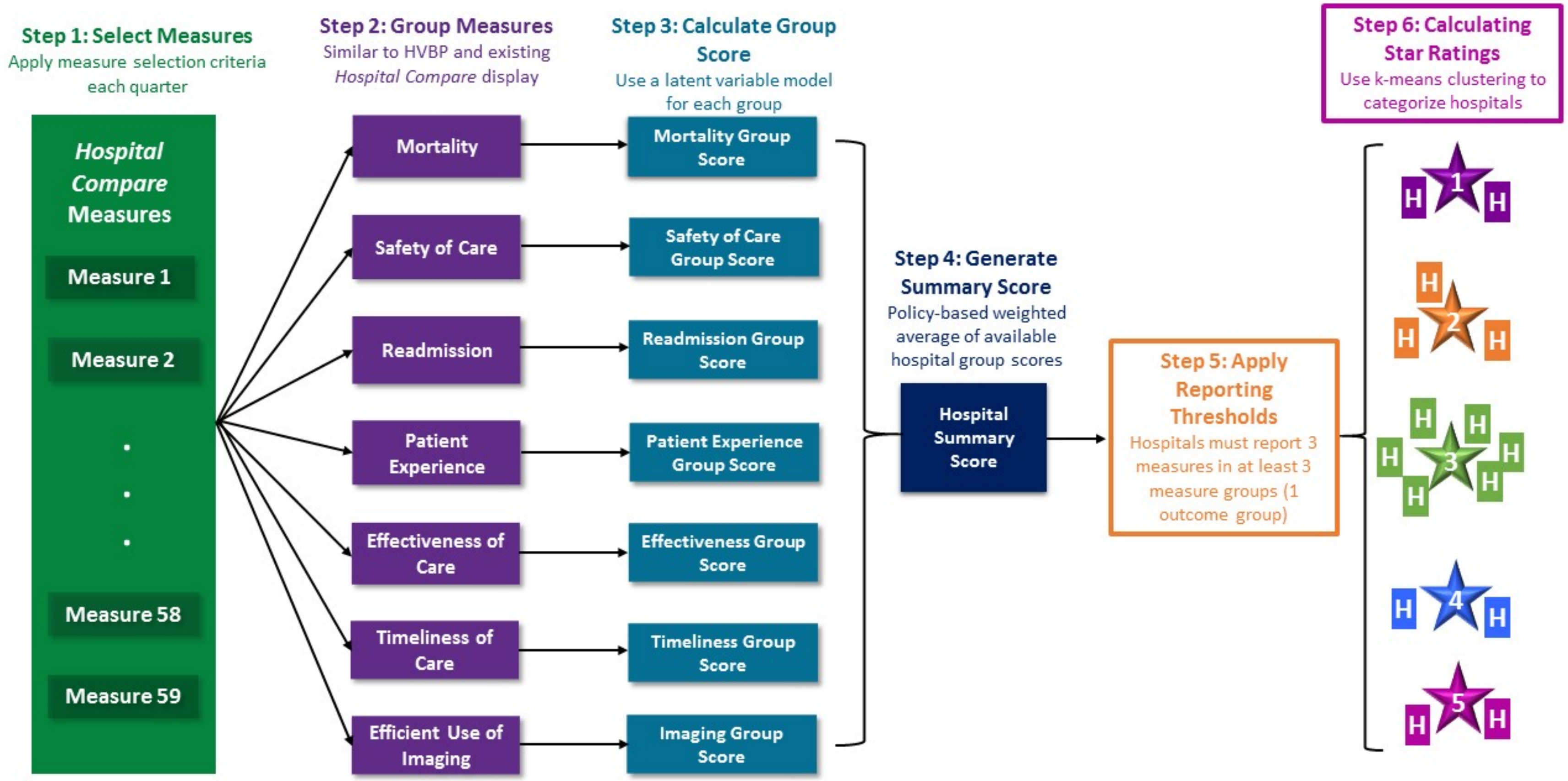
PLWG Member	Title
Elizabeth Mort, MD, MPH	Mass General Hospital, Mass General Physicians Organization *(Senior Vice President of Quality and Safety and Chief Quality Officer)*
Brian Wilmoth, CPA	UVA & VCA Health Systems *(Strategic Planning and Reimbursement Officer)*
Hazel Crews, PT, MHS, MHA, CPHQs	Indiana University Health *(Regional Director of Quality and Patient Safety Office and Director of Research)*
Kelly Gray-Eurom, MD, MMM, FACEP	University of Florida, Campus of Medicine *(Physician, Chief Quality Officer and Professor of Emergency Medicine)*
Lisa Shea, MD	Butler Hospital *(Medical Director of Quality and Regulation)*
Henry Pitt, MD	Temple University Health System *(Chief Quality Officer)*
Lloyd Guthrie, MBA, BSN	Center for Improving Value in Health Care *(Program Manager for Statewide Initiatives)*
Jens Eldrup-Jorgensen, MD	Vascular Center at Maine Medical Center and Society for Vascular Surgery Patient Safety Organization *(Medical Director);* Tufts University School of Medicine *(Professor of Surgery)*

Table B.4. Star Ratings Patient and Patient Advocate Work Group Roster

Work Group Member	Title
Anna Howard, JD	American Cancer Society Cancer Action Network (*Policy Principal*)
Gail Hunt	National Alliance for Caregiving (*President and CEO*)
Ann Monroe, MA	Health Foundation of Western and Central New York (*President*)
Claire Noel-Miller, MPA, PhD	American Association of Retired People (AARP) (*Senior Strategic Policy Advisor*)
Melissa Thomason	Vidant Health System (*Patient/Family Advisor*)
Linda Relyea	Primary Care Practice *(Patient Family Advisory Council)*
Mary Anne Lueken	King's Daughters Medical Center *(Patient Family Care Advisor)*

Appendix C: Flowchart of Six-Step Overall Star Rating Methodology

Figure C.1. The Six Steps of the Overall Star Rating Methodology


Step 1: Select Measures
Apply measure selection criteria each quarter
Hospital Compare Measures
Measure 1
Measure 2
.
.
.
Measure 58
Measure 59
Step 2: Group Measures
Similar to HVBP and existing Hospital Compare display
Mortality
Safety of Care
Readmission
Patient Experience
Effectiveness of Care
Timeliness of Care
Efficient Use of Imaging
Step 3: Calculate Group Score
Use a latent variable model for each group
Mortality Group Score
Safety of Care Group Score
Readmission Group Score
Patient Experience Group Score
Effectiveness Group Score
Timeliness Group Score
Imaging Group Score
Step 4: Generate Summary Score
Policy-based weighted average of available hospital group scores
Hospital Summary Score
Step 5: Apply Reporting Thresholds
Hospitals must report 3 measures in at least 3 measure groups (1 outcome group)
Step 6: Calculating Star Ratings
Use k-means clustering to categorize hospitals
1
2
3
4
5
H

Appendix D: National Distribution of Measures per Group for December 2017

Figure D.1. Count of Hospitals by Number of Mortality Measures

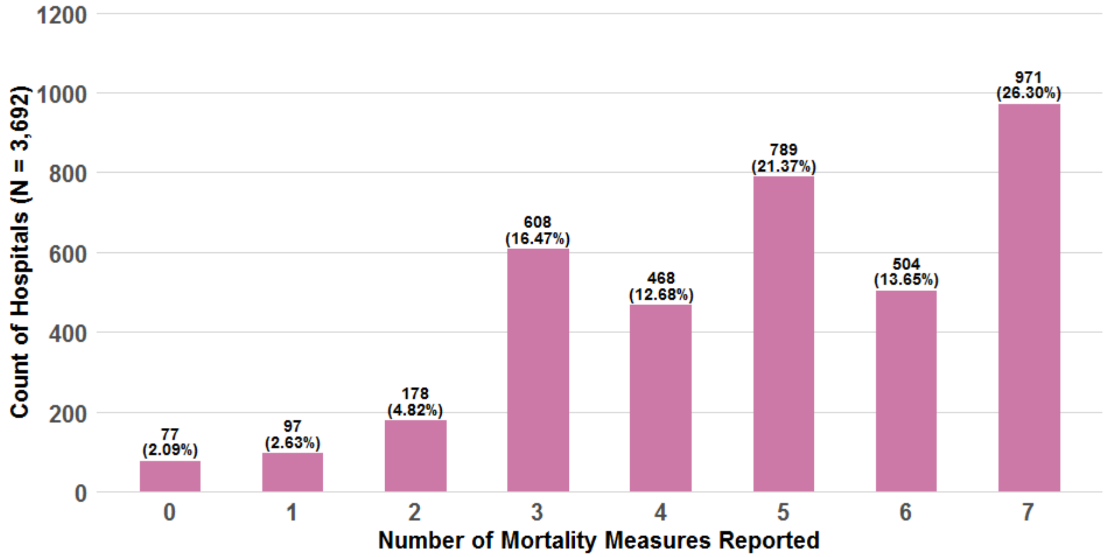


Figure D.2. Count of Hospitals by Number of Safety of Care Measures

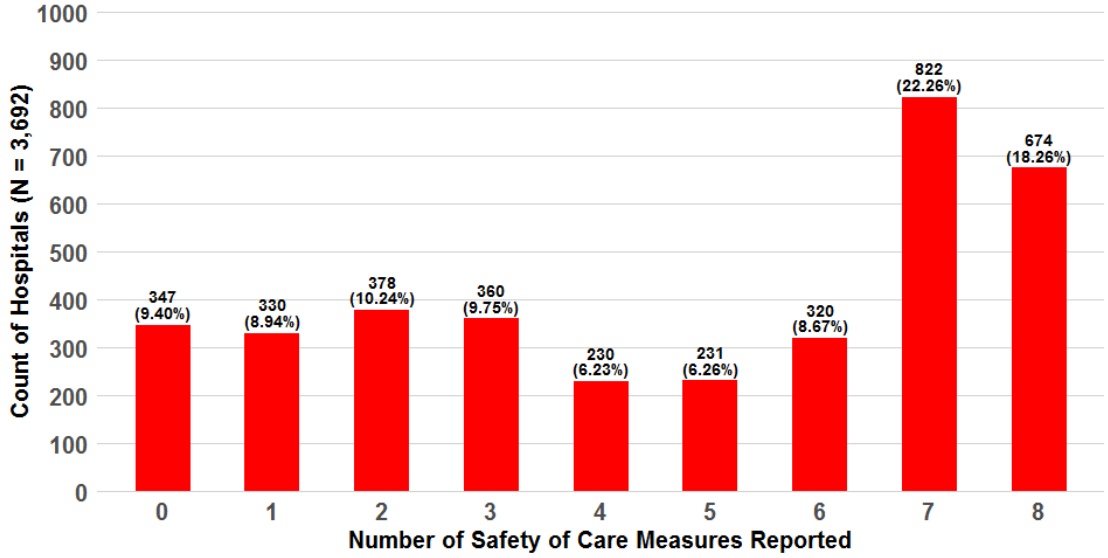


Figure D.3. Count of Hospitals by Number of Readmission Measures

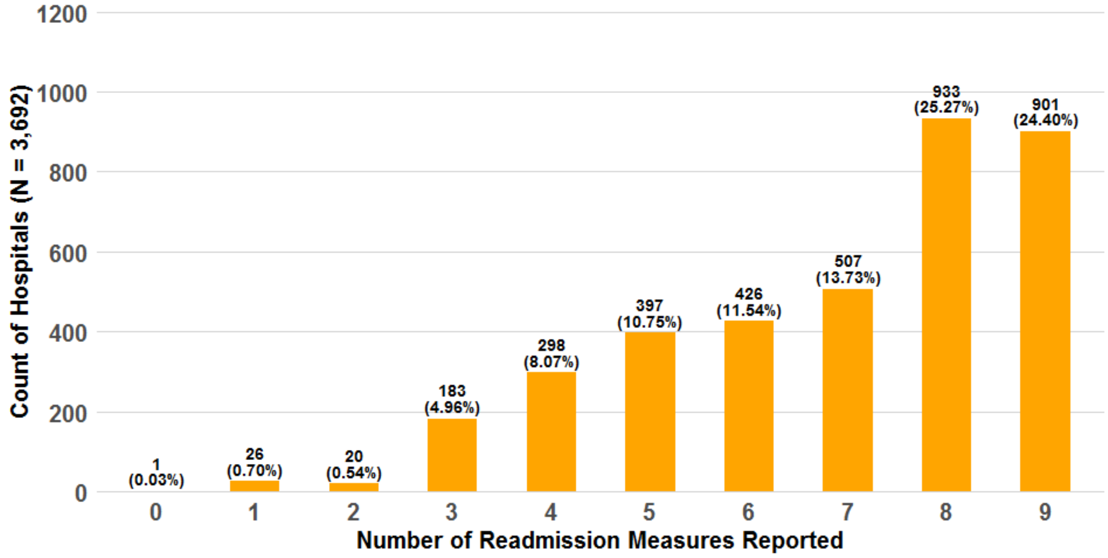


Figure D.4. Count of Hospitals By Number of Patient Experience Measures

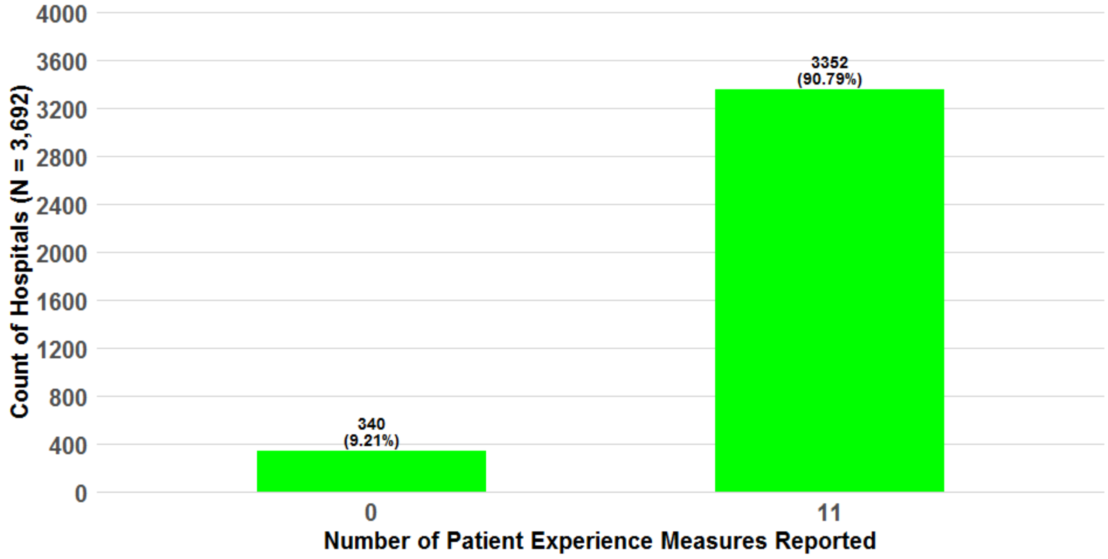


Figure D.5. Count of Hospitals By Number of Effectiveness of Care Measures

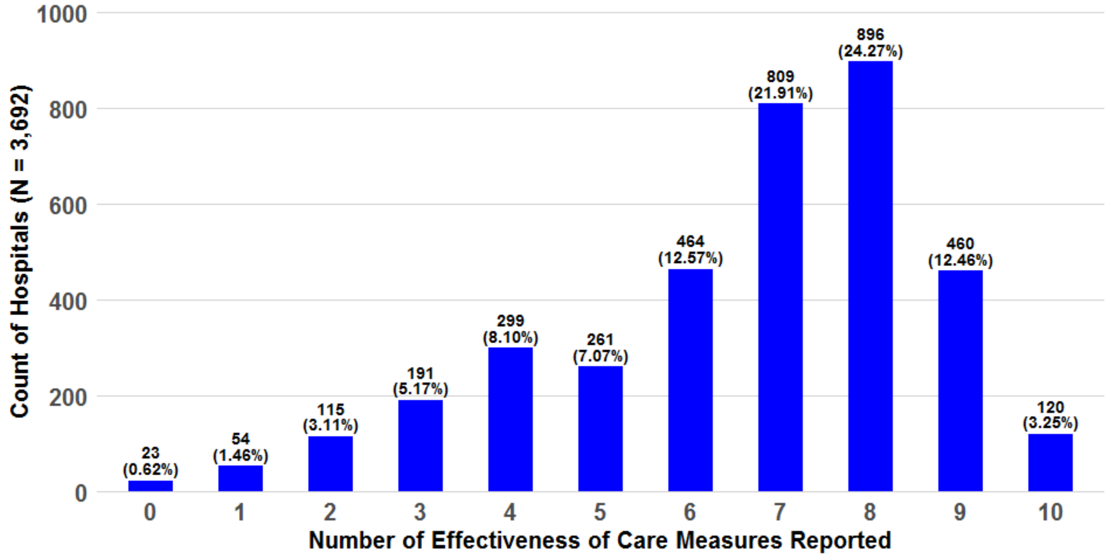


Figure D.6. Count of Hospitals By Number of Timeliness of Care Measures

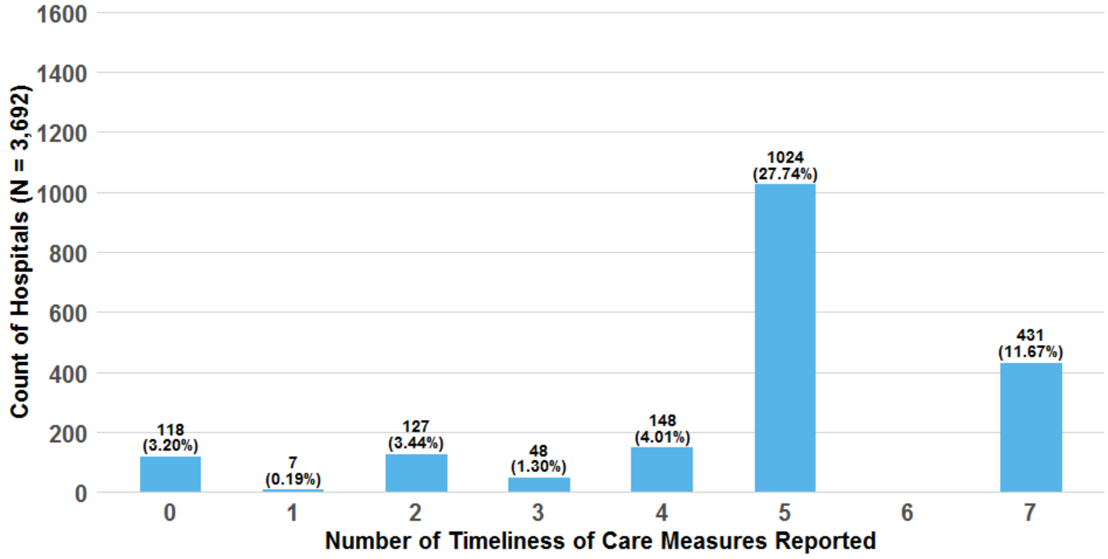
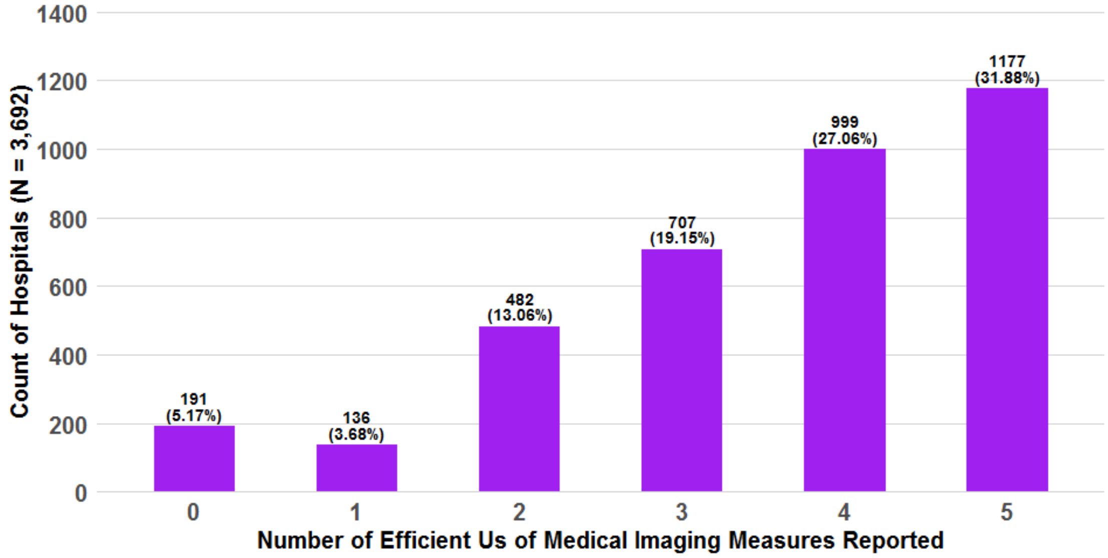


Figure D.7. Count of Hospitals By Number of Efficient Use of Medical Imaging Measures



Appendix E. Results of Validity Testing Using April 2015 Data

Testing Each Star Rating Group for a Single, Latent Trait

Figure E.1. Scree Plot Results for Mortality Group

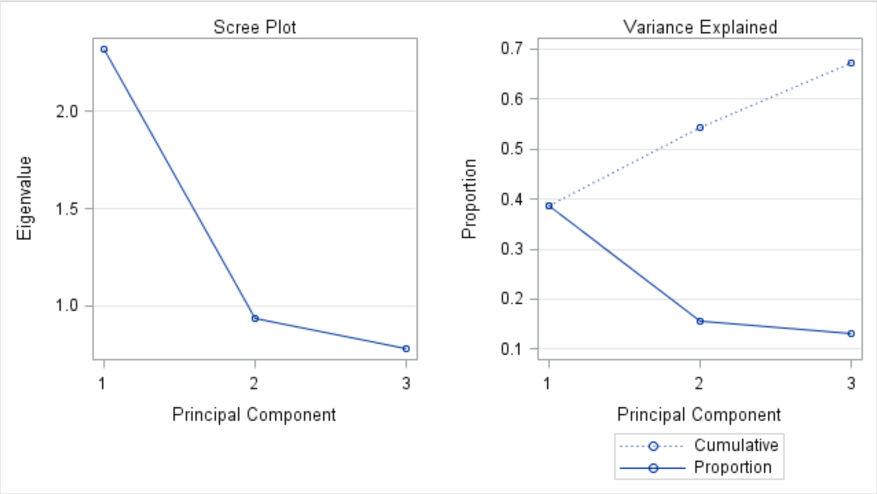


Figure E.2. Scree Plot Results for Safety of Care Group

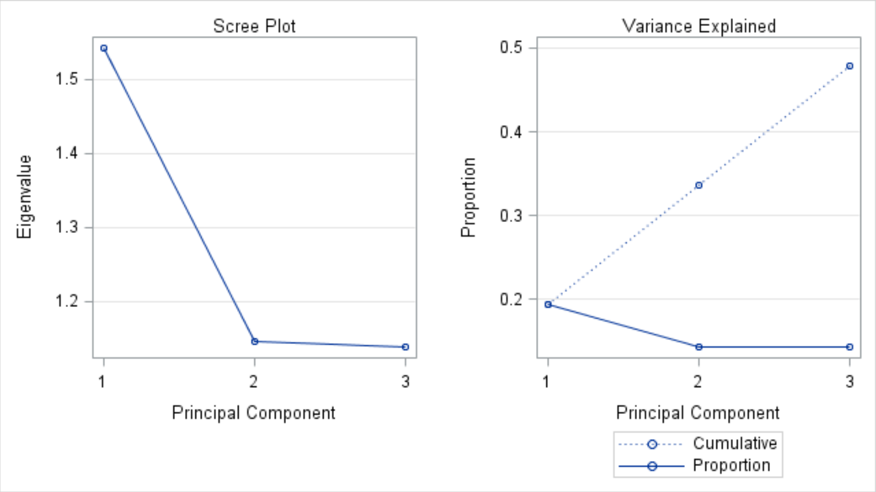


Figure E.3. Scree Plot Results for Readmission Group

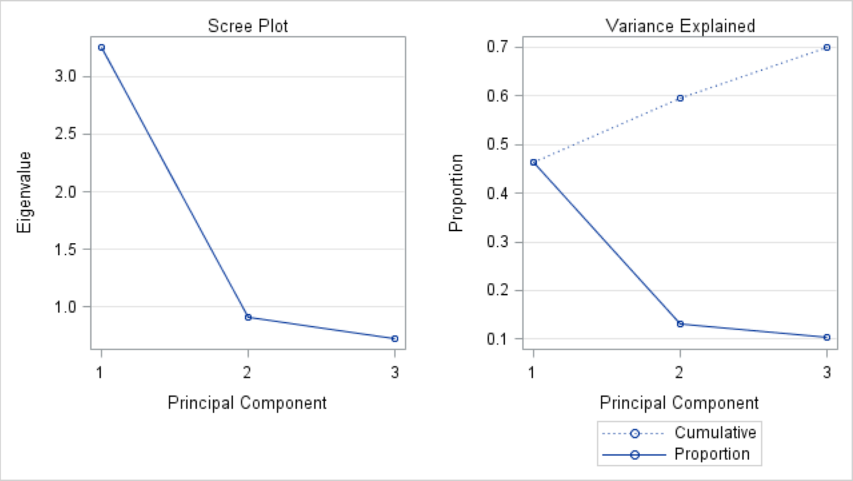


Figure E.4. Scree Plot Results for Patient Experience Group

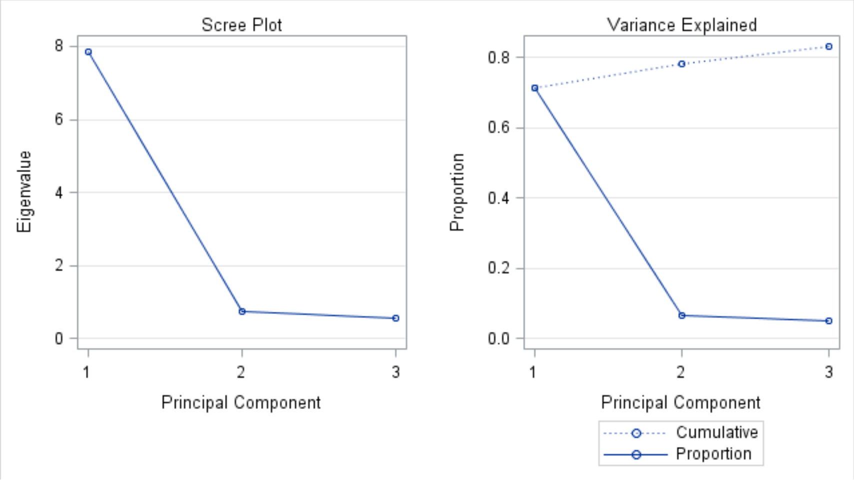


Figure E.5. Scree Plot Results for Effectiveness of Care Group

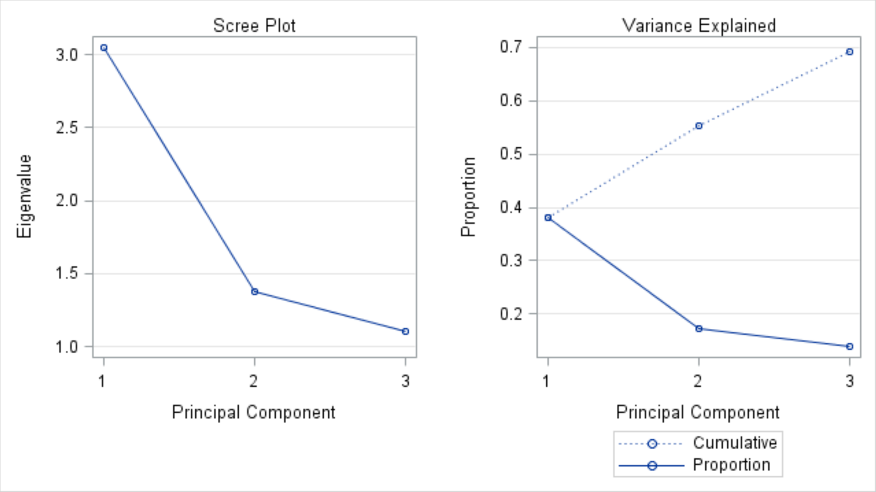


Figure E.6. Scree Plot Results for Timeliness of Care Group

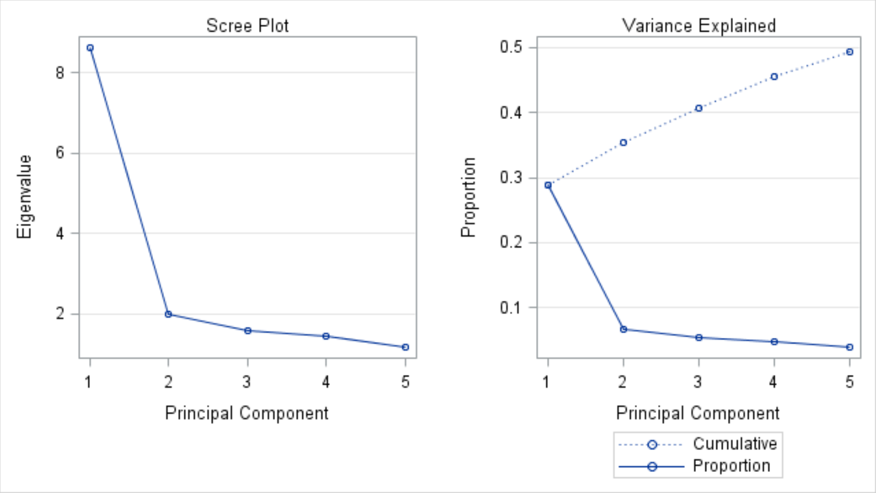
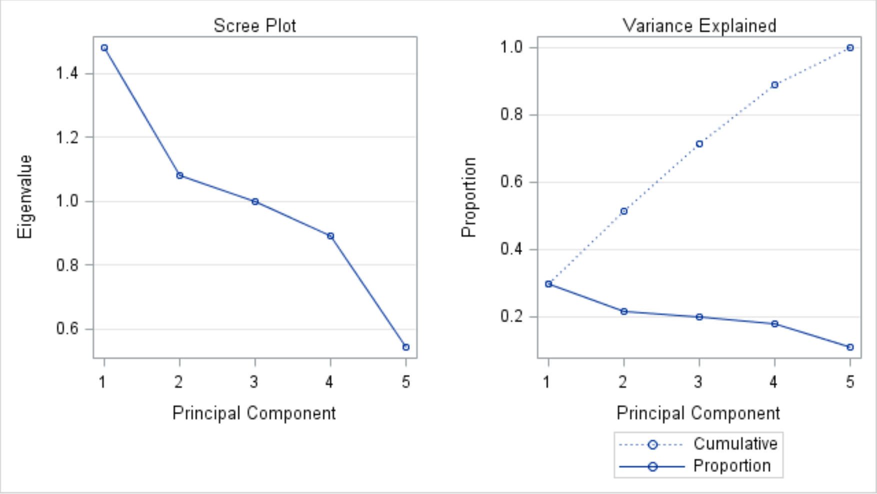


Figure E.7. Scree Plot Results for Efficient Use of Medical Imaging Group



Pairwise Comparison of Mean Group Scores between Star Classes

Tables E.1 – E.7 present the results of CMS's pairwise comparisons of mean group scores between the Star Rating classes using December 2017 data. The checkmarks (✓) presented in the table indicate statistically significant ($p<0.05$) differences. between the mean group scores of the two compared.

Table E.1. Pairwise Comparison of Star Classes using Mortality Mean Group Scores

Star Comparison	Difference between group scores	p<0.05
5-4	0.48	✓
5-3	0.69	✓
5-2	0.85	✓
5-1	0.99	✓
4-3	0.22	✓
4-2	0.37	✓
4-1	0.52	✓
3-2	0.16	✓
3-1	0.30	✓
2-1	0.15	

Table E.2. Pairwise Comparison of Star Classes using Safety of Care Mean Group Scores

Star Comparison	Difference between group scores	p<0.05
5-4	0.69	✓
5-3	0.98	✓
5-2	1.44	✓
5-1	2.36	✓
4-3	0.29	✓
4-2	0.75	✓
4-1	1.67	✓
3-2	0.47	✓
3-1	1.38	✓
2-1	0.91	✓

Table E.3. Pairwise Comparison of Star Classes using Readmission Mean Group Scores

Star Comparison	Difference between group scores	p<0.05
5-4	0.68	✓
5-3	1.23	✓
5-2	1.81	✓
5-1	2.64	✓
4-3	0.55	✓
4-2	1.12	✓
4-1	1.96	✓
3-2	0.57	✓
3-1	1.41	✓
2-1	0.84	✓

Table E.4. Pairwise Comparison of Star Classes using Patient Experience Mean Group Scores

Star Comparison	Difference between group scores	p<0.05
5-4	0.39	✓
5-3	0.93	✓
5-2	1.51	✓
5-1	2.25	✓
4-3	0.54	✓
4-2	1.11	✓
4-1	1.86	✓
3-2	0.58	✓
3-1	1.32	✓
2-1	0.74	✓

Table E.5. Pairwise Comparison of Star Classes using Effectiveness of Care Mean Group Scores

Star Comparison	Difference between group scores	p<0.05
5-4	0.13	✓
5-3	0.26	✓
5-2	0.35	✓
5-1	0.67	✓
4-3	0.13	✓
4-2	0.22	✓
4-1	0.55	✓
3-2	0.09	
3-1	0.41	✓
2-1	0.32	✓

Table E.6. Pairwise Comparison of Star Classes using Timeliness of Care Mean Group Scores

Star Comparison	Difference between group scores	p<0.05
5-4	-0.09	
5-3	0.08	
5-2	0.58	✓
5-1	1.27	✓
4-3	0.17	✓
4-2	0.68	✓
4-1	1.37	✓
3-2	0.51	✓
3-1	1.19	✓
2-1	0.69	✓

Table E.7. Pairwise Comparison of Star Classes using Efficient use of Medical Imaging Mean Group Scores

Star Comparison	Difference between group scores	p<0.05
5-4	0.11	
5-3	0.23	✓
5-2	0.29	✓
5-1	0.24	✓
4-3	0.12	✓
4-2	0.19	✓
4-1	0.13	
3-2	0.06	
3-1	0.01	
2-1	-0.05	

Linear Trend of Star Rating Group Scores across Star Categories Using December 2017 Data

Figure E.8. Mortality Group Scores across Star Categories

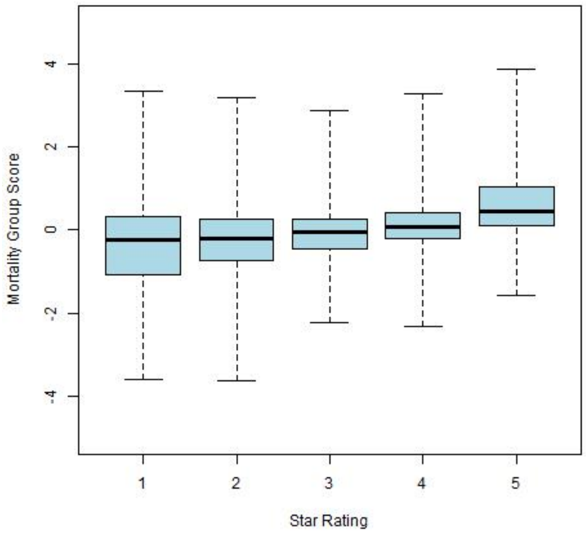


Figure E.9. Safety of Care Group Scores across Star Categories

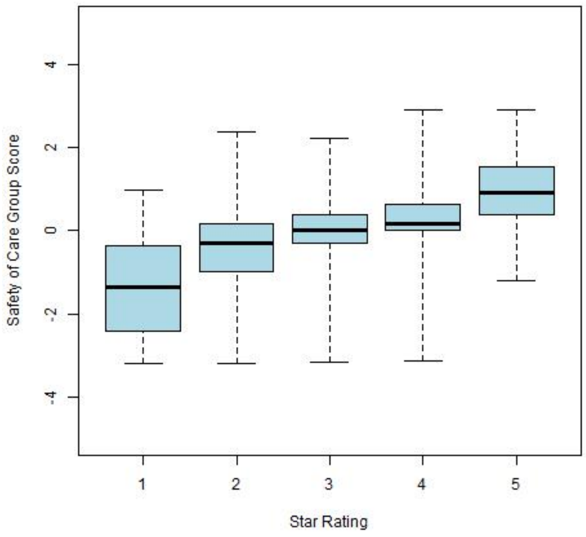


Figure E.10. Readmission Group Scores across Star Categories

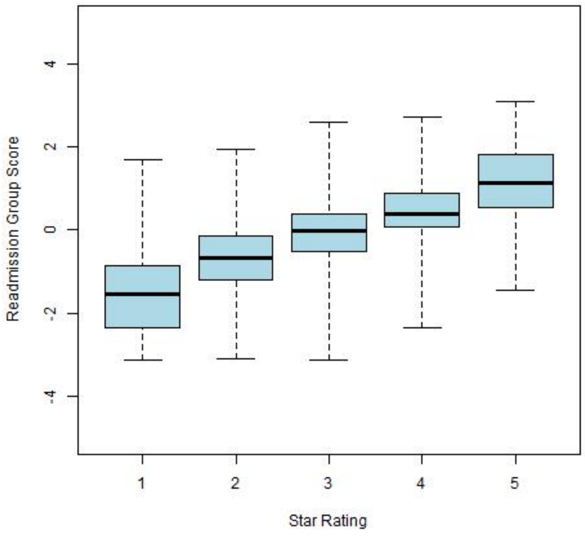


Figure E.11. Patient Experience Group Scores across Star Categories

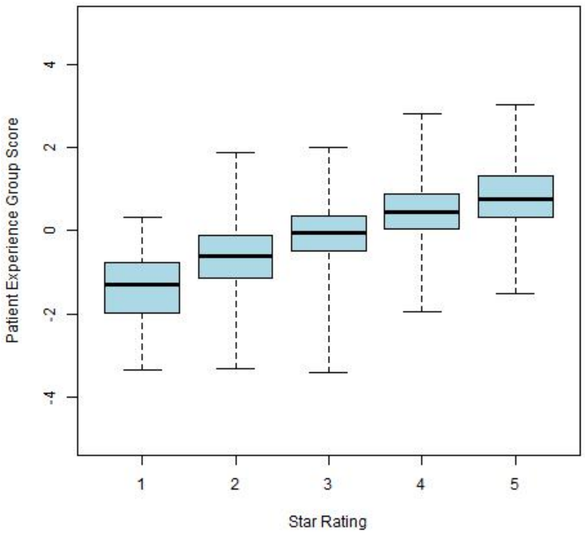


Figure E.12. Effectiveness of Care Group Scores across Star Categories

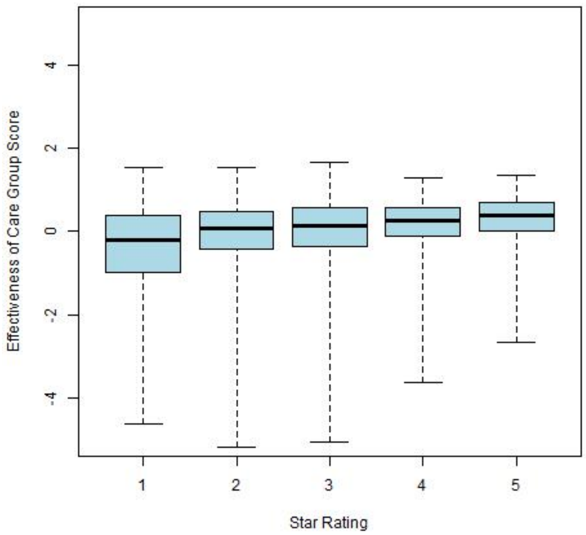


Figure E.14. Efficient Use of Medical Imaging Group Scores across Star Categories

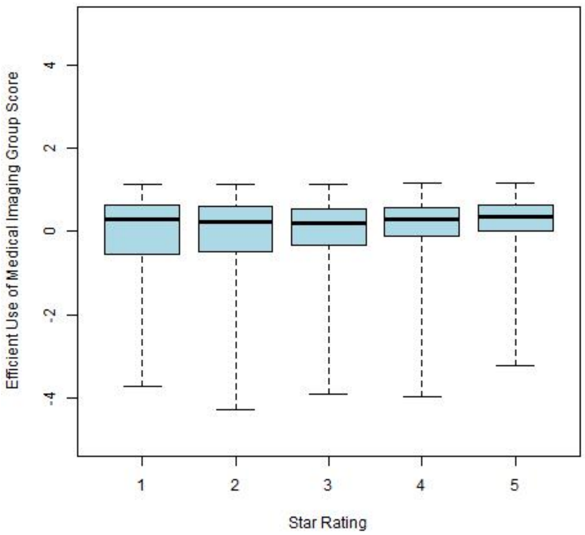
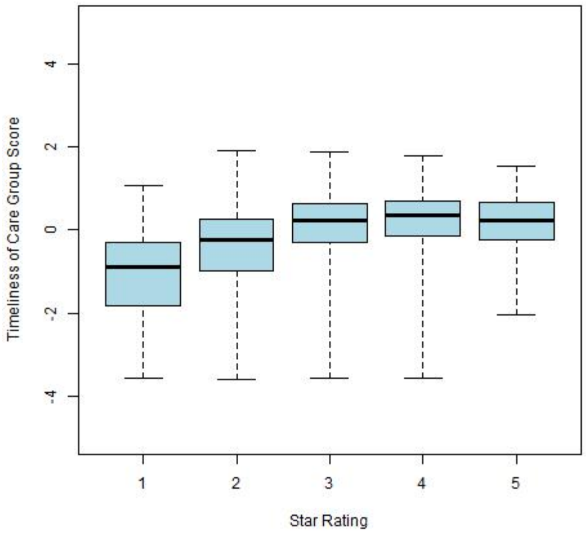


Figure E.13. Timeliness of Care Group Scores across Star Categories


Timeliness of Care Group Score
Star Rating