# Overall Hospital Quality Star Ratings on *Hospital Compare*

# Methodology Report (v2.0)

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

**May 2016**

# 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 |
| **Angela Hsieh, PhD** | Lead Analyst (2014-2016) |
| **Li Qin, PhD** | Lead Analyst (2016) |
| **Haikun Bao, PhD** | Supporting Analyst |
| **Jaymie Simoes, MPH** | Project Coordinator II |
| **Mallory Perez, BSPH** | Research Assistant II |
| **Erica Norton, 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, Vinitha Meyyur, PhD, Kia Stanfield, Alicia Budd, MPH, Pierre Yong, MD, MPH, Kristie Baus, MS, RN, William Lehrman, MA, PhD, and Kate Goodrich, MD, MHS 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 Hector Cariello, BS, MPH 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, Kanchana Bhat, MPH, Benjamin Clopper, MPH, Rana Searfoss, BA, 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 Technical Expert Panel (TEP) and our patient and patient advocate working group.

# I. Executive Summary

This report presents the methodology (v2.0) for the Overall Hospital Quality Star Ratings, 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*.

## 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 Ratings 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 Ratings 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.

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 Ratings" refers to Overall Hospital Quality Star Ratings 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 Ratings 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 a multi-stakeholder Technical Expert Panel (TEP); holding two public comment periods, a National Stakeholder call, a hospital dry run; and convening a patient and patient advocate working group, to date.

CMS designed several aspects of the Star Ratings development process to include the patient and consumer perspective in key methodological and policy decisions. Both the TEP and working 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 five-step approach to calculating the Star Ratings. The measures are first 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. 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. In the fourth step, a weight is applied to each group score, and all available groups are averaged to calculate the hospital summary score. Finally, to assign a Star Rating, the hospital summary scores are organized into five ordered categories using a clustering algorithm.

In addition to the Star Ratings, CMS also organized hospitals into one of three group performance categories (above, same as, and below the national average) for each of the hospital's available groups, providing additional detail for patients and consumers.

## Conclusion

The overarching goal of the Overall Hospital Quality Star Ratings is to improve the usability and interpretability of *Hospital Compare* for patients and consumers. This report reflects a two-year effort to develop a scientifically robust Star Rating methodology that considers and incorporates feedback from a diverse set of stakeholders. As CMS continues to develop its Star Rating intiatives, this methodology may be updated and revised as needed.  CMS will continue to collaborate with stakeholders and users of *Hospital Compare*, aiming to support the appropriate use of the Overall Hospital Quality Star Ratings and be responsive to the dynamic realm of hospital quality measurement.

# 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 Ratings 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 Ratings 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 Star Rating 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 Ratings are designed to provide summary information for consumers about existing publicly-reported quality data. 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. Throughout the remainder of this report, "Star Ratings" refers to Overall Hospital Quality Star Ratings unless otherwise noted.

## 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 Ratings, 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 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 of measures and different 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.

The methodology (v2.0), presented in this report, reflects slight modifications made since the July 2015 hospital dry run (v1.0). 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 Ratings and conducting validity and reliability analyses (Appendix A).

The methodology (v2.0) calculates the Star Ratings through a five-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 Star Ratings
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 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. Finally, in Step 5, to assign Star Ratings, hospital summary scores were organized into five categories using a clustering algorithm.

Of note, CMS reports hospital performance at the group level, separately categorizing each of a hospital's available group scores into one of three group performance categories (above, same as, and below the national average). These performance categories provide additional detail for patients and consumers for comparing across the seven groups (Section V).

## 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 Ratings, the Star Ratings include a broad and diverse set of measures.

### *Criteria for Selecting Measures for Overall Hospital Quality Star Ratings*

CMS vetted measure selection criteria with stakeholders through the TEP and public comments to ensure that the Star Ratings 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 Ratings.

Because the Star Ratings are 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 Ratings for April 2016 was 113 measures.[1] 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 no longer required for Inpatient Quality Reporting (IQR) Program or Outpatient Quality Reporting (OQR) Program; and
6. Duplicative measures.

### *Standardization of Measure Scores*

Before combining measures into a score, each measure was first converted into a common scale of measurement. Hospital quality measure results include many different types of scoring information,

[1] Of note, Star Rating results calculated using April 2016 data were not publicly reported. These results were calculated for the purposes of this report and for the hospital preview period, which allowed hospitals to review their hospital-specific results in advance of public reporting.

ranging from time (e.g., median time in minutes from 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 AMI). Therefore, to enable the combination of information, CMS used standardization to ensure all measure scores were in a common scale with a common direction. This did 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 standardized 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; 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 Winsorized 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 bellow -3 are set to be -3. This was 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. Ultimately, CMS grouped measures into seven groups (Table 1). The use of groups in Star Ratings is consistent with other CMS Star Rating initiatives (Nursing Home Compare Star Ratings, Medicare Plan Finder Star Ratings, and Dialysis Facility Compare).

**Table 1. Overall Hospital Quality Star Rating Groups for April 2016**

| Overall Star Rating Groups |
|---|
| Mortality |
| Safety of Care |
| Readmission |
| Patient Experience |
| Effectiveness of Care |
| Timeliness of Care |
| Efficient Use of Medical Imaging |

The rationale for these seven 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 Ratings.
- The groups were vetted and supported by the TEP.

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 April 2015 *Hospital Compare* dataset supported CMS's decision to assign measures to groups. Out of the 75 measures in the Star Ratings for this reporting period, the average hospital reported 44.8 measures (Interquartile range: 19 to 69). The distribution of hospital quality reporting by group for the April 2015 reporting period is described in Appendix D.

The group names were finalized with input from a patient and patient advocate working 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 April 2016, please refer to Section V.

## 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 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 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 initially understand, CMS aimed to overcome this challenge by embedding multiple channels for stakeholder education throughout development and preparation for implementation of the Star Rating methodology.

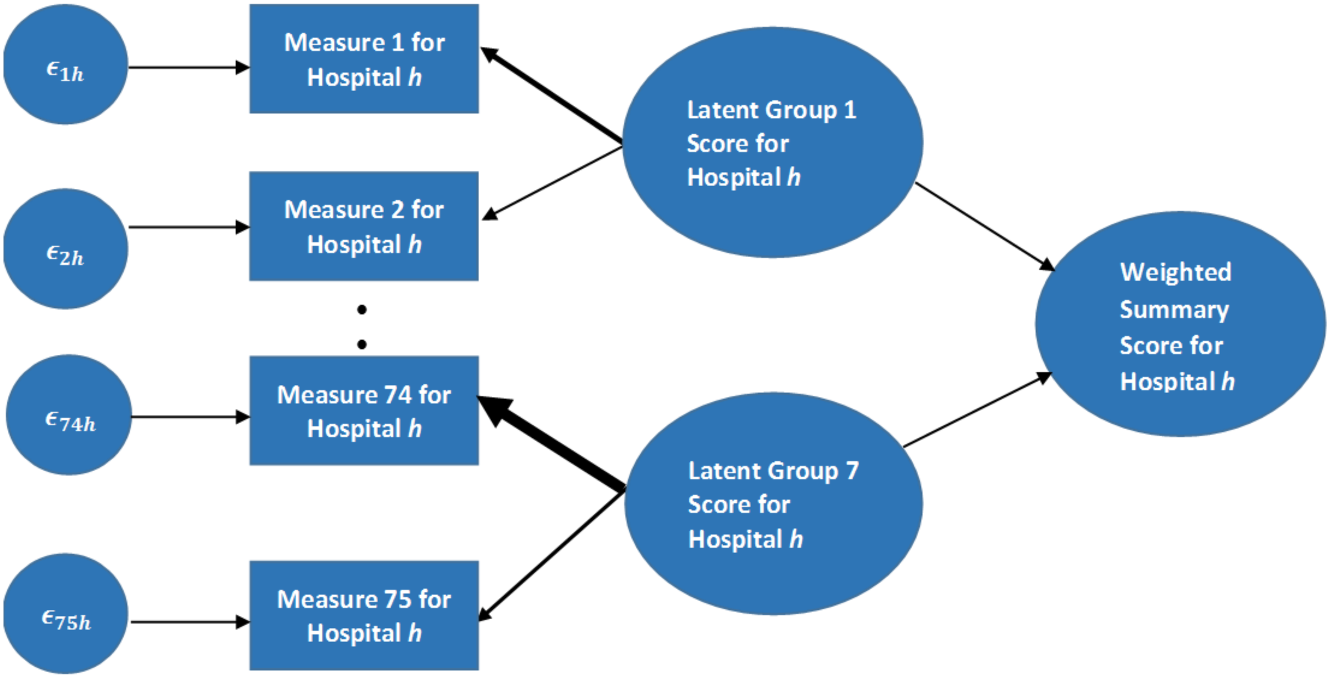
## *Detailed Description of LVM*

In this section, CMS presents a sample path diagram for the LVM of each group in the Star Ratings as well as the statistical equations used to calculate group scores. CMS constructed the LVM using standard software, SAS Proc NLMIXED. All parameters in the models were estimated by maximum likelihood method, and the group scores were obtained as empirical Bayes estimates.

### *Path Diagram*

In the sample path diagram presented in Figure 1, the ovals represent the group scores and hospital summary score. 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 Ratings. 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**


$\epsilon_{1h}$
Measure 1 for Hospital *h*
$\epsilon_{2h}$
Measure 2 for Hospital *h*
$\epsilon_{74h}$
Measure 74 for Hospital *h*
$\epsilon_{75h}$
Measure 75 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 statistical equation for LVM to calculate a hospital's group score (Equation 1) is as follows:

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

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

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

Let $Y_{khd}$ denote the standardized score for hospital *h* and measure *k* in group *d*. $\alpha_{hd}$ is the hospital-specific group-level latent trait (random effect) for hospital *h* and group *d*. $\gamma_{kd}$ 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*. $\alpha_{hd}$ follows a Normal distribution with mean 0 and variance 1. The assumption of unit variance here is an innocuous choice of units required to identify the parameter $\mu_{kd}$ and $\gamma_{kd}$.

## *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 applying 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 Ratings are 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 gave 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.

## 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 calculated 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 comment 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 public comment, the hospital dry run, and the patient and patient advocate working group.

Given the TEP's feedback and criteria set during development, CMS finalized a policy-based weighting scheme modified from that used for the HVBP program (Table 2). The statistical equation that uses these weights to calculate hospital summary scores is presented in Equation 3.

**Table 2. 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**

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

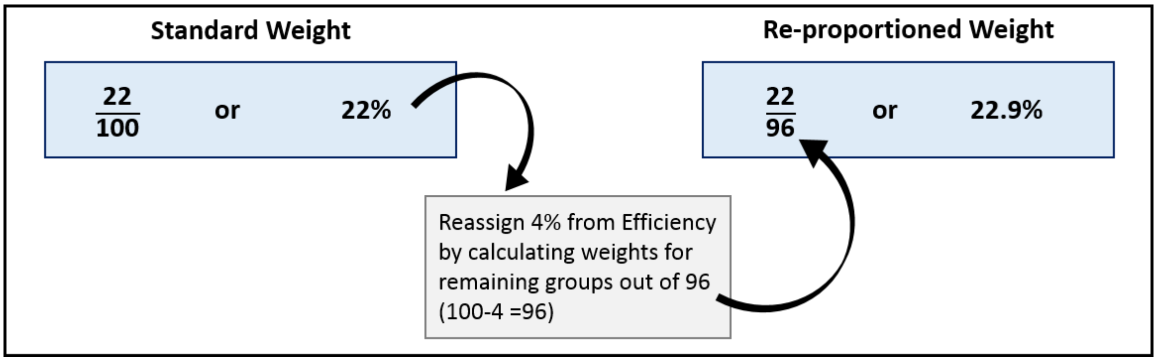
## *Method for Re-weighting When Missing Group(s)*

If a hospital reports no measures for a given group, CMS considered that group to be "missing." When a hospital is missing one or more groups, CMS applied 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 3 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.

**Table 3. 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% | --- |

**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)

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

## *Winsorization of Summary Scores*

Next, CMS analyzed the distribution of hospital summary scores and performed Winsorization. Winsorization is a common strategy used to set extreme outliers to a specified percentile of the data; in this case, any extreme outlier values lower than the 0.5th percentile and higher than the 99.5th percentile is set to have the 0.5th percentile or 99.5th percentile value, respectively. The decision to Winsorize hospital summary scores, a modification from the Star Rating dry run, was based on comments received during the second public comment period and patients' and consumers' preference for a broader distribution of Star Ratings.

## Step 5: 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 margin of a star category (some hospitals will border a higher/lower star category).
- Similar to other CMS Star Ratings efforts, a three-star rating will be considered “average.”
- The objective of this project is to develop whole-star ratings (not half-stars).
- Star ratings do 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.
- Star ratings are 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. 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 *k*-means clustering analysis generates five categories based on hospital summary scores in a way that minimizes the 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 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, the *k*-means clustering approach produced 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 VI.

# IV. Minimum Thresholds for Reporting a Star Rating

CMS aimed 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.

## Minimum Threshold of Measures per Group

CMS set 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 4). In Table 4, the "Required N" refers to the number of measures needed to meet the desired level of reliability (R).

**Table 4. 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 set 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 ensured that the Star Ratings were inclusive of publicly reported measures and was vetted with the public through the second Star Rating public comment.

## 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 (Table 5). Setting increasingly higher thresholds for both measures and groups would exclude more hospitals from the Star Ratings.

The minimum measure and group thresholds were applied solely for reporting purposes and had no effect on the calculation of hospital summary scores or the star categorization.

**Table 5. Hospitals with Star Rating based on minimum thresholds using April 2015 Data**

| | Minimum Group Threshold | | | | | |
|---|---|---|---|---|---|---|
| Minimum Measure Threshold | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 4,617 (97%) | 4,330 (91%) | 3,958 (83%) | 3,713 (78%) | 3,353 (71%) | 3,009 (63%) |
| 2 | 4,329 (91%) | 4,020 (85%) | 3,639 (77%) | 3,319 (70%) | 3,061 (64%) | 2,789 (59%) |
| 3 | 3,988 (84%) | 3,709 (78%) | 3,307 (70%) | 3,044 (64%) | 2,845 (60%) | 2,411 (51%) |
| 4 | 3,499 (74%) | 3,277 (69%) | 3,036 (64%) | 2,801 (59%) | 2,481 (52%) | 1,831 (39%) |

Note: The fixed number of minimum groups shown in Table 5 must include at least one Outcome group.

# V. Development Results of the Star Rating Methodology

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

## *Distribution of Star Ratings (April 2016)*

CMS calculated the Star Ratings for April 2016 using the April 2016 *Hospital Compare* dataset.[2] The frequency of hospitals by each Star Rating category is shown in Table 6.

**Table 6. Distribution of Star Ratings for April 2016**

| Star Rating | Number of Hospitals (% Total) |
|---|---|
| ★★★★★ | 87 (2.39%) |
| ★★★★ | 821 (22.51%) |
| ★★★ | 1,881 (51.58%) |
| ★★ | 716 (19.63%) |
| ★ | 142 (3.98%) |

## *Measure Selection*

Figure 3 depicts the process of measure selection for April 2016. Out of a possible 113 eligible for inclusion in the Star Ratings, 51 measures were excluded based on our selection criteria.

## *Assignment to Groups*

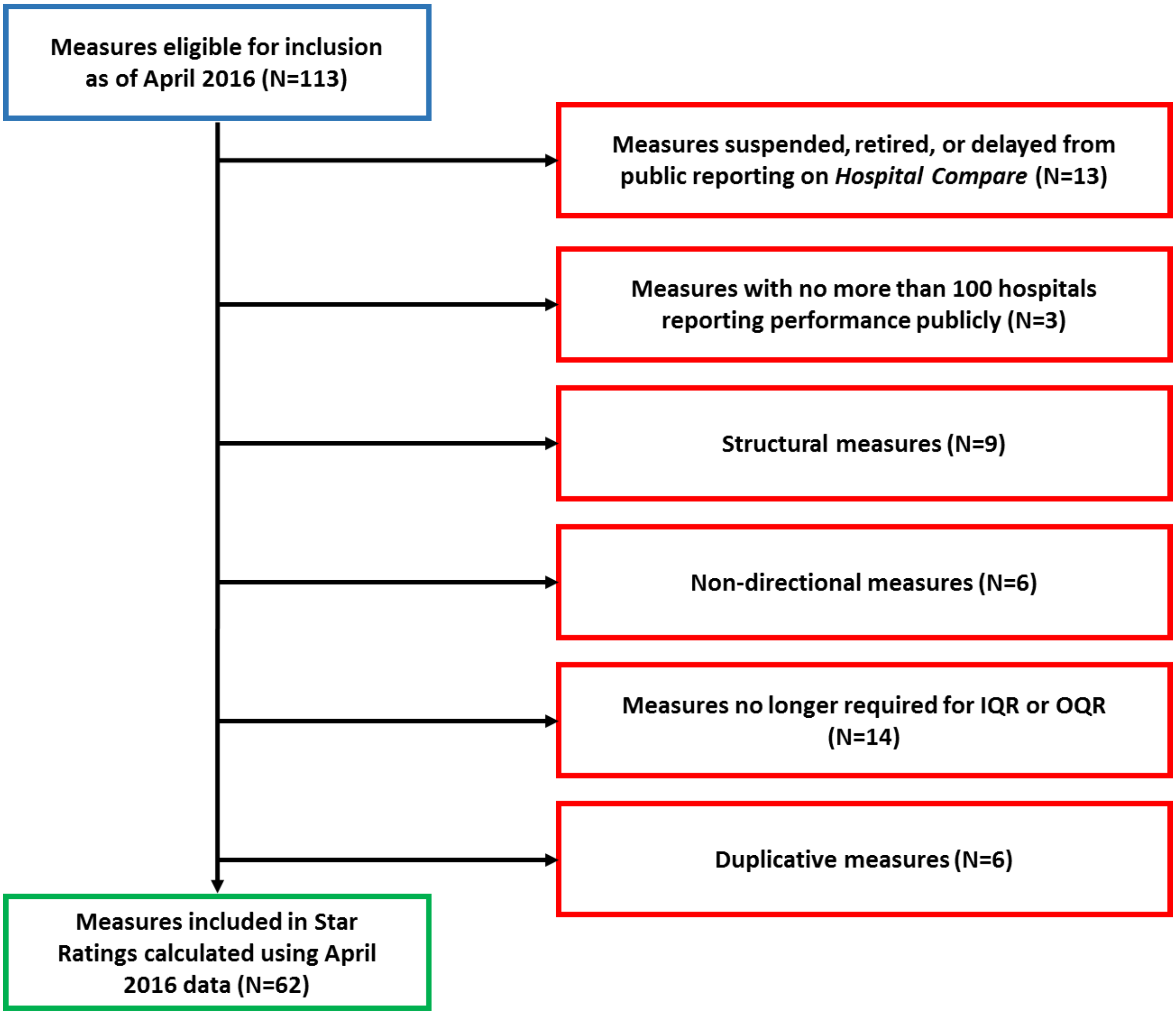
Table 7 displays the number of measures assigned to each group, out of the total 62 measures included in the Star Ratings for April 2016. The complete list of measures by name per group can be found in the Overall Hospital Quality Star Rating Updates and Specifications Report: April 2016, available on *QualityNet* at www.qualitynet.org > Hospitals – Inpatient > Hospital Star Ratings.

**Table 7. Total Number of Measures by Group for April 2016**

| Group | Number of Measures (N=62) |
|---|---|
| Mortality | 7 |
| Safety of Care | 8 |
| Readmission | 8 |
| Patient Experience | 11 |
| Effectiveness of Care | 16 |
| Timeliness of Care | 7 |
| Efficient Use of Medical Imaging | 5 |

[2] Of note, Star Rating results calculated using April 2016 data were not publicly reported. These results were calculated for the purposes of this report and for the hospital preview period, which allowed hospitals to review their hospital-specific results in advance of public reporting.

Figure 3. Measure Selection Flowchart (April 2016 Data)


Measures eligible for inclusion as of April 2016 (N=113)
Measures suspended, retired, or delayed from public reporting on *Hospital Compare* (N=13)
Measures with no more than 100 hospitals reporting performance publicly (N=3)
Structural measures (N=9)
Non-directional measures (N=6)
Measures no longer required for IQR or OQR (N=14)
Duplicative measures (N=6)
Measures included in Star Ratings calculated using April 2016 data (N=62)

## Group Scores

### 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 compared a hospital's group score to the national average group score. The LVM for each group produced a point estimate and standard error for each hospital's group score that CMS used to construct a 95% confidence interval for each hospital's group score. CMS compared this 95% confidence interval to the national mean group score. CMS defined 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.

### Group Peformance Category Results for April 2016

Table 8 displays the frequency of hospitals in each group performance category for April 2016.

**Table 8. Frequency of Hospitals by Group Performance Category**

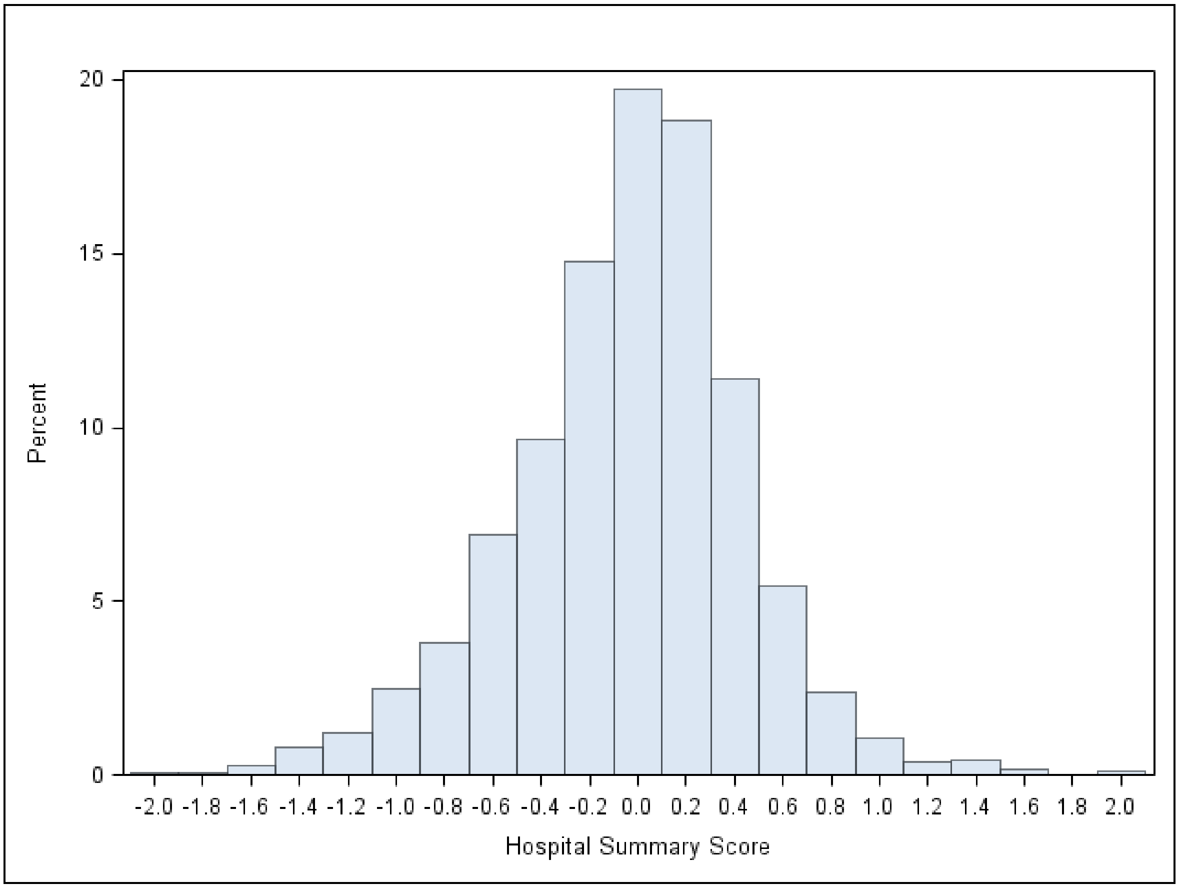
| | Frequency (Number of Hospitals) by Group Performance Category | | |
|---|---|---|---|
| **Group** | **Above the National Average** | **No Different than the National Average** | **Below the National Average** |
| Mortality (N=3,524) | 406 (11.52%) | 2791 (79.20%) | 327 (9.28%) |
| Safety of Care (N=2,817) | 795 (28.20%) | 1321 (46.86%) | 703 (24.94%) |
| Readmission (N=2,126) | 853 (22.04%) | 2126 (54.92%) | 892 (23.04%) |
| Patient Experience (N=3,549) | 1032 (29.08%) | 1206 (33.98%) | 1311 (36.94%) |
| Effectiveness of Care (N=3,642) | 1185 (32.54%) | 1954 (53.65%) | 503 (13.81%) |
| Timeliness of Care (N=3,425) | 1238 (36.15%) | 1295 (37.81%) | 892 (26.04%) |
| Efficient Use of Medical Imaging (N=2,787) | 354 (12.70%) | 2087 (74.88%) | 346 (12.41%) |

Note: The total number of hospitals in the *Hospital Compare* dataset as of April 2016 was 4,604 hospitals. Results shown are for all hospitals with ≥ 3 measures by group.

### Hospital Summary Scores

Figure 4 presents the distribution of hospital summary scores (N=4,604) for April 2016. 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 April 2016**



### Star Classification

Table 9 shows the frequency (number of hospitals) in each of the five star categories for April 2016. 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 9. Frequency of Hospitals by Star Category using k-Means for April 2016**

| Rating | Frequency of Hospitals | Summary Score Range in Category | Mean (sd) |
|---|---|---|---|
| 1 Star | 142 (3.89%) | (-2.01, -0.97) | -1.23 (0.22) |
| 2 Star | 716 (19.63%) | (-0.96, -0.33) | -0.58 (0.17) |
| 3 Star | 1881 (51.58%) | (-0.33, 0.25) | -0.01 (0.16) |
| 4 Star | 821 (22.51%) | (0.25, 0.86) | 0.46 (0.15) |
| 5 Star | 87 (2.39%) | (0.86, 1.96) | 1.17 (0.28) |

Note: The total number of hospitals in the *Hospital Compare* dataset for April 2016 was 4,604 hospitals. The table shows the results for the 3,647 hospitals that met the reporting criteria.

# VI. Testing the 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 finalize the Star Rating methodology and 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 Ratings, 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 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 first iteration of the Star Ratings 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, CMS conducted an analysis that describes the distribution of group scores for hospitals in each Star Rating category. 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 (Table E.1).

CMS found statistically significant differences in group scores between each Star Rating category for every group except Mortality, Effectiveness of Care, Timeliness of Care, and Efficient Use of Medical Imaging. For these groups, CMS found statistically significant differences in group scores for the majority of comparisons between Star Ratings, except for the Efficient Use of Medical Imaging group.

### *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 ($p = 0.20$); the higher the group score, the higher the hospital's Star Rating. The results of this analysis, using April 2015 data, are depicted as box plots by group in Appendix E.

Despite the absence of meaningful differences in hospital performance across Star Rating categories in the Efficient Use of Medical Imaging group, the linear-trend is positive with higher-star hospitals performing better and lower-star hospitals performing worse at the group score level. Balancing these technical limitations with the original development principle of inclusivity and the TEP's support for including efficiency measures in the Star Ratings, CMS kept this group in the Star Ratings.

## Reliability Testing

To confirm the reliability of the Star Rating methodology, CMS tested the stability of the Star Ratings, both over time and across simulations of data within the same time period. First, CMS compared the Star Ratings 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 organizaing 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 Analysis)*

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]).[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 is 0.45 (comparing to 1 if complete agreement), which indicates moderate agreement of Star Ratings between these two quarters. The ICC score is 0.91, which indicates substantial agreement of summary score between these two quarters.

### *Re-classification Analysis of Star Ratings*

To evaluate the relaibilty of *k*-means clustering, CMS used simulation to calculate 5,000 summary scores for each hospital. In each simulation, each hospital's summary score was randomly selected based on the distribution of the hospital's summary score and the summary score standard error using April 2015 data. Within each simulation, CMS reclassified the hospitals into 5 star categories, fixing the range of summary scores included in each category based on the results of *k*-means clustering for the 2015 hospital dry run (April 2015 data). CMS examined the proportion of the simulations that reclassified hospitals into the same star category as they were assigned during the 2015 dry run (Table 10). The higher the proportion, the higher the reliability of the classification step of the methodology .

In Table 10, The percentage in each cell represents the proportion of hospitals with the assigned Star Rating during 2015 April dry run (the first column) that were re-classified into the corresponding Star categories using 5,000 simulations. For example, in 5,000 simulations, 78.91% of the time, the 573 hospitals classified as two-star during the 2015 April dry run were re-classified as two-star hospitals.

**Table 10. Classification Analysis of k-Means Clustering using April 2015 Data**

| April 2015 Rating | Rating for Simulated Samples | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 71.51% | 28.08% | 0.37% | 0.03% | 0.00% |
| 2 | 1.28% | 78.91% | 19.53% | 0.27% | 0.01% |
| 3 | 0.28% | 13.81% | 69.09% | 16.17% | 0.65% |
| 4 | 0.00% | 0.40% | 24.98% | 71.75% | 2.85% |
| 5 | 0.00% | 0.24% | 1.76% | 19.54% | 78.46% |

## *Re-classification Analysis of Group Performance Categories*

In addition to checking the reliability of the Star Ratings 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 3 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 2015 April dry run (Table 11). The higher the proportion, the higher the reliability.

**Table 11. Re-classification Analysis of Group Performance Categories using April 2015 Data**

| | Reliability in Each Group Performance Category | | |
|---|---|---|---|
| Group | Above the National Average | Same as the National Average | Below the National Average |
| Mortality | 83.57% | 87.68% | 80.36% |
| Safety of Care | 89.06% | 86.06% | 91.15% |
| Readmission | 89.65% | 86.90% | 91.87% |
| Patient Experience | 90.42% | 80.43% | 92.20% |
| Effectiveness of Care | 81.62% | 87.00% | 88.17% |
| Timeliness of Care | 82.35% | 82.26% | 91.76% |
| Efficient Use of Medical Imaging | 68.44% | 83.84% | 85.80% |

## Summary of Testing

The analyses conducted by CMS supported several key assumptions as well as confirmed the level of the validity and reliability sought by CMS for public reporting.

The underlying assumption that each of the Star Rating groups represents 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 almost 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 when comparing results from April 2015 to July 2015, a time period that included the addition and removal of several measures. Furthermore, within the same performance period, the reclassification rate (hospitals being re-classified into their original Star Rating category and group performance category) for nearly all Star Rating and group performance categories demonstrated high reliability (R>0.7).

As the distribution of hospital performance evolves and/or updates to the Star Rating methodology are made, CMS will continue to test the validity and reliability of 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. |

# Appendix B: Stakeholder Roster

**Table B.1. Techinical Expert Panel (TEP) Roster**

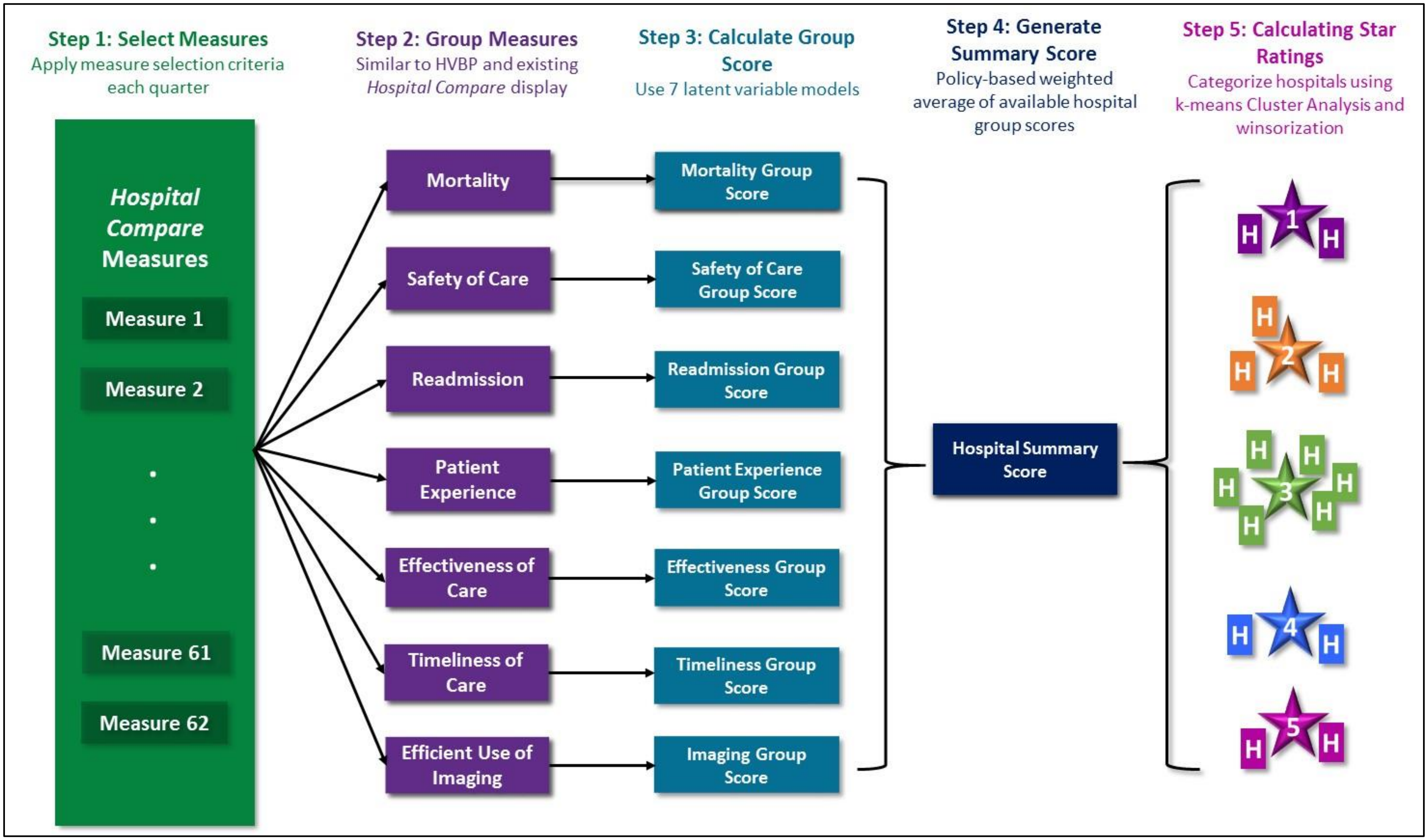
| TEP Member | Title |
|---|---|
| Matt Austin, PhD | Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University *(Assistant Professor)* |
| Vinita Bahl, DMD, MPP | Performance Assessment & Clinical Effectiveness, University of Michigan Health System *(Director)* |
| John Bott, 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. Star Ratings Working Group Roster**

| Working 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 (*Patient/Family Advisor*) |

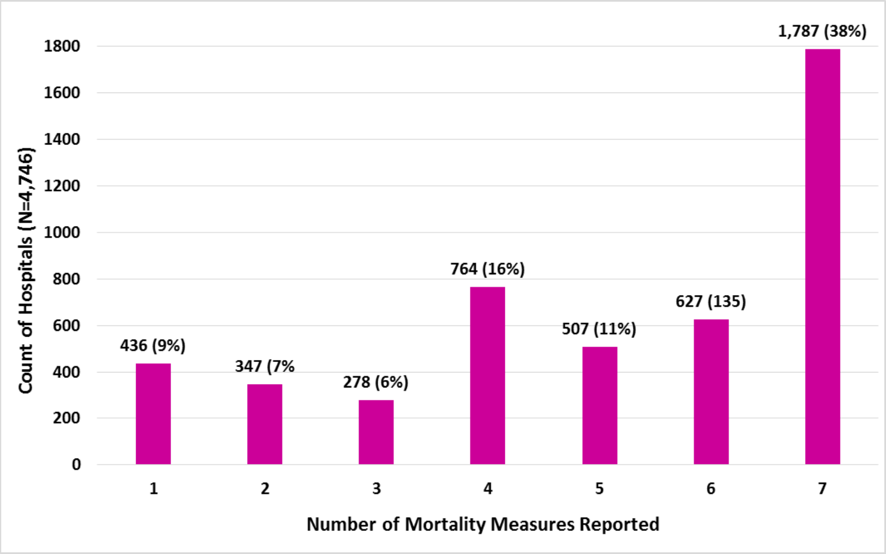
# Appendix C: Flowchart of Five-Step Overall Star Rating Methodology

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


Step 1: Select Measures
Apply measure selection criteria each quarter
Step 2: Group Measures
Similar to HVBP and existing *Hospital Compare* display
Step 3: Calculate Group Score
Use 7 latent variable models
Step 4: Generate Summary Score
Policy-based weighted average of available hospital group scores
Step 5: Calculating Star Ratings
Categorize hospitals using k-means Cluster Analysis and winsorization
*Hospital Compare* Measures
Measure 1
Measure 2
•
•
•
Measure 61
Measure 62
Mortality
Safety of Care
Readmission
Patient Experience
Effectiveness of Care
Timeliness of Care
Efficient Use of Imaging
Mortality Group Score
Safety of Care Group Score
Readmission Group Score
Patient Experience Group Score
Effectiveness Group Score
Timeliness Group Score
Imaging Group Score
Hospital Summary Score
H
1
H
H
H
2
H
H
H
H
H
3
H
H
H
H
4
H
H
5
H

# Appendix D: National Distribution of Measures per Group for April 2015

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



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

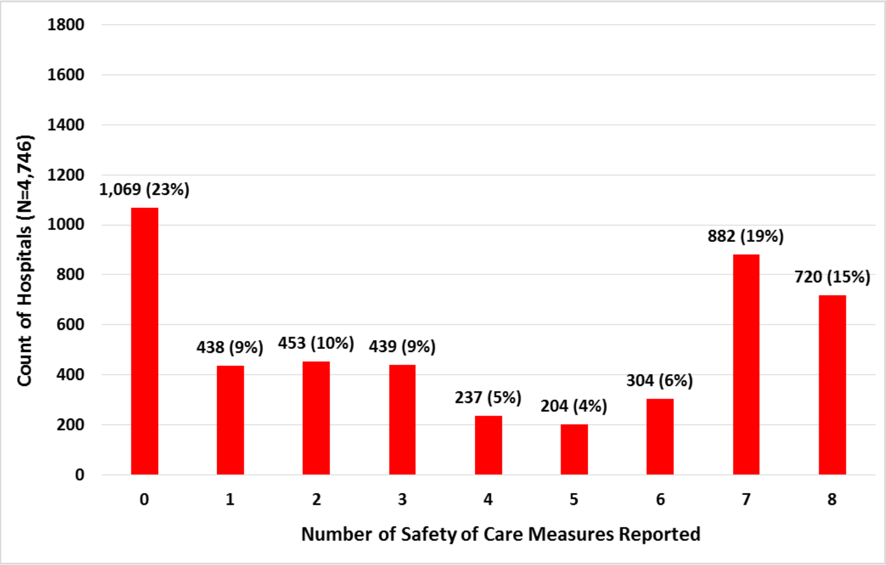


**Figure D.3. Count of Hospitals by Number of Readmission Measures**

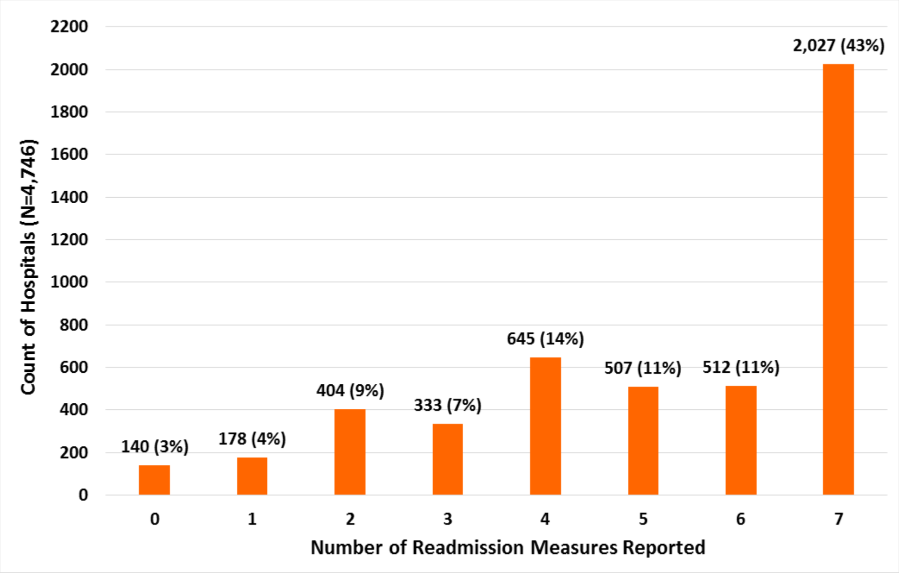


**Figure D.4. Count of Hospitals By Number of Patient Experience Measures**

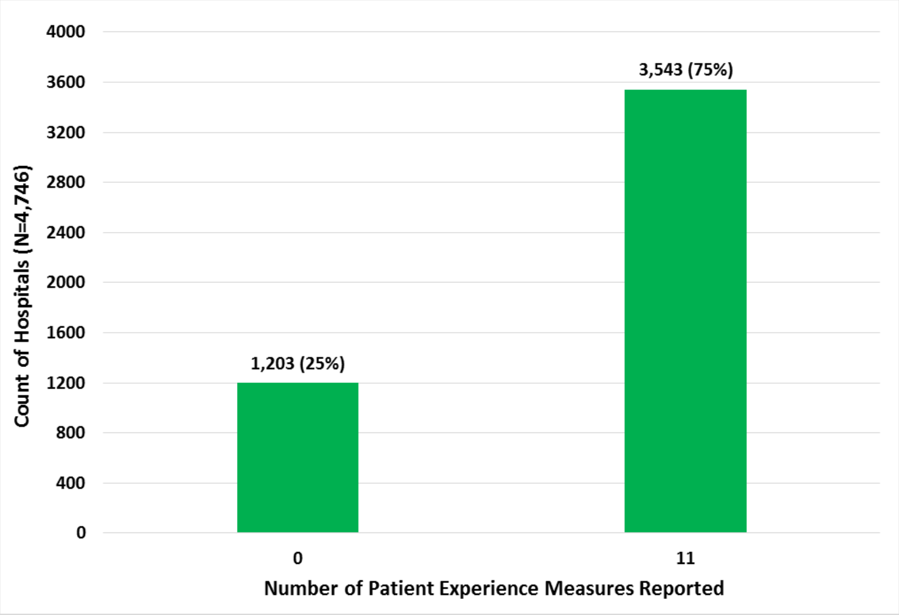


**Figure D.5. Count of Hospitals By Number of Effectiveness of Care Measures**

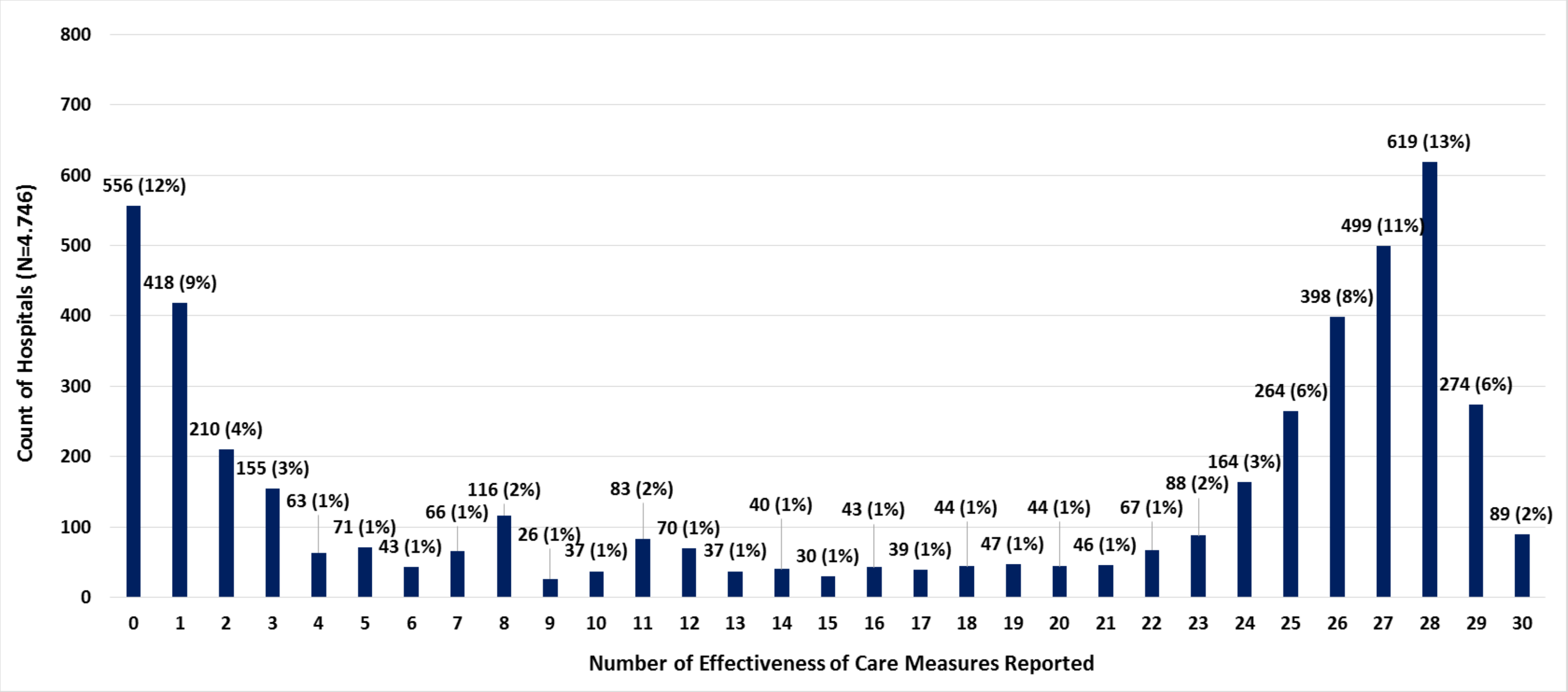


Figure D.6. Count of Hospitals By Number of Timeliness of Care Measures

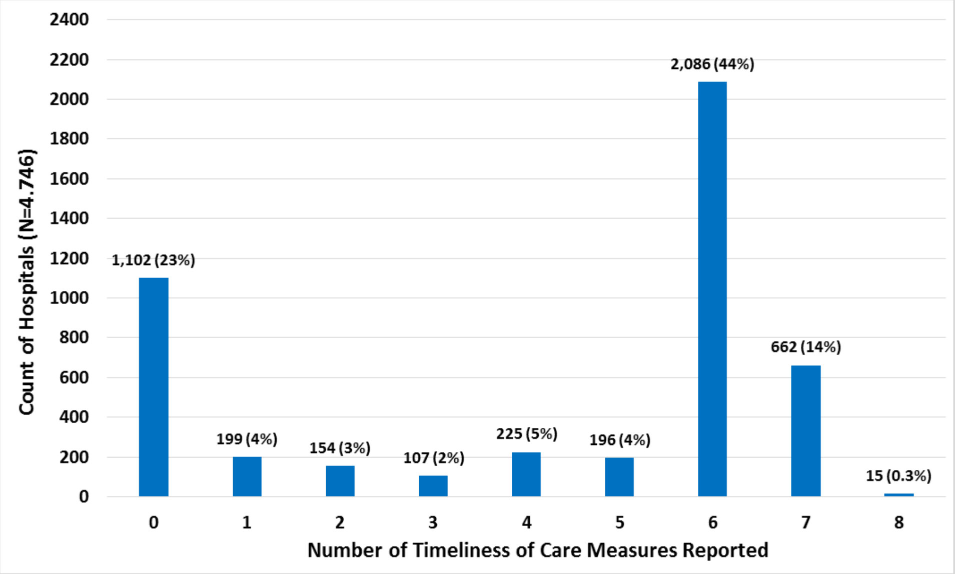
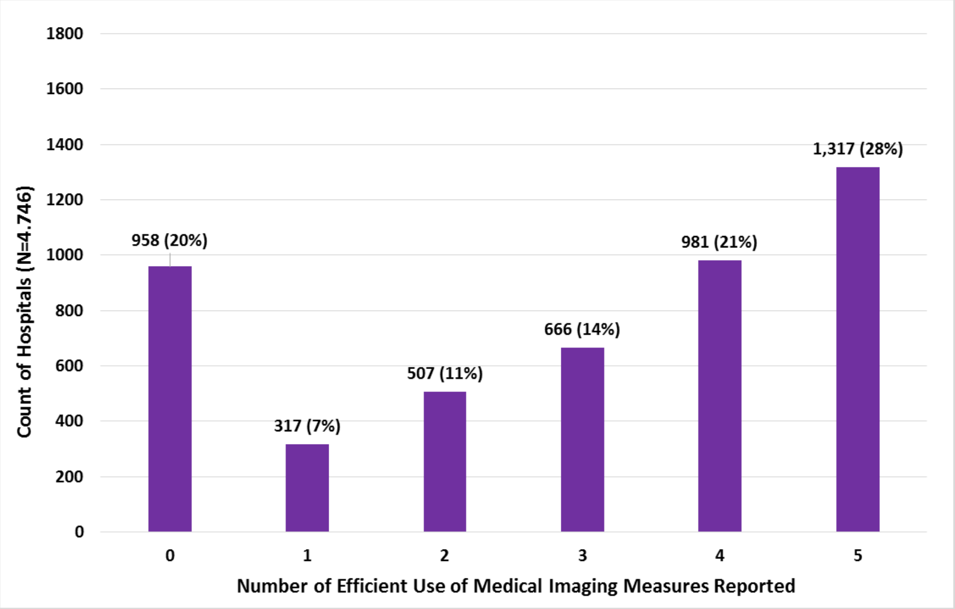


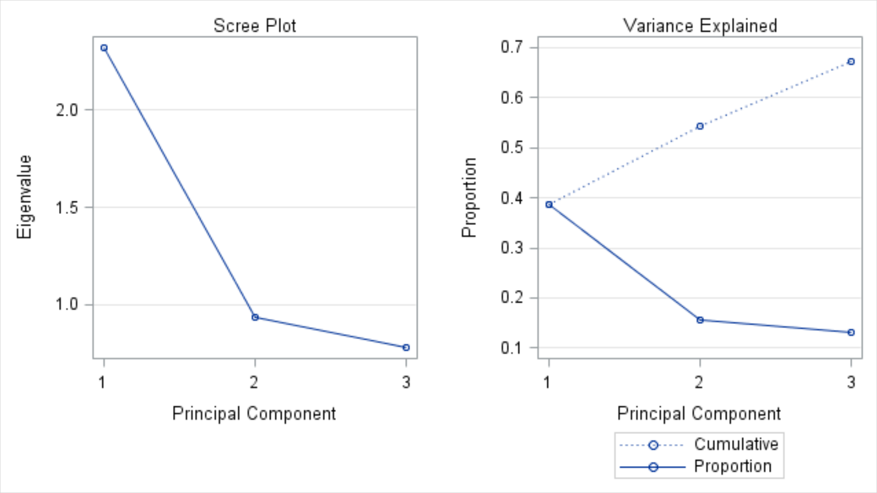
Figure D.7. Count of Hospitals By Number of Efficient Use of Medical Imaging Measures



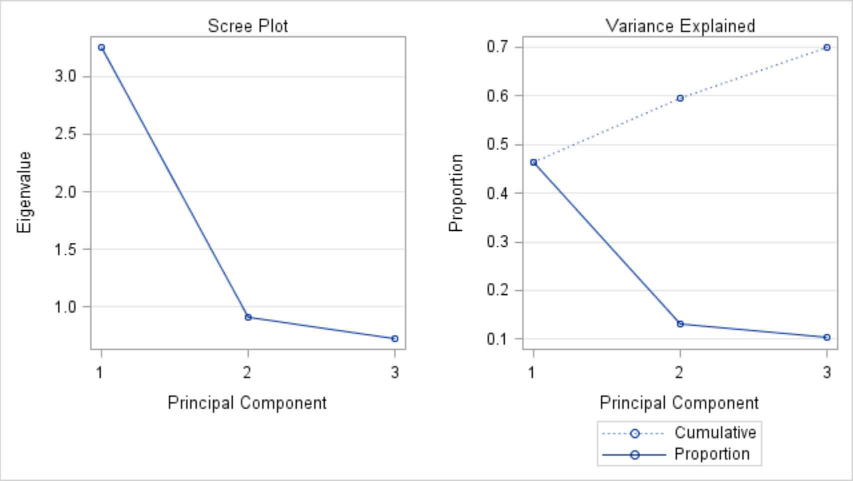
# Appendix E. Results of Validity Testing Using April 2015 Data

## *Testing Each Star Rating Group for a Single, Latent Trait*

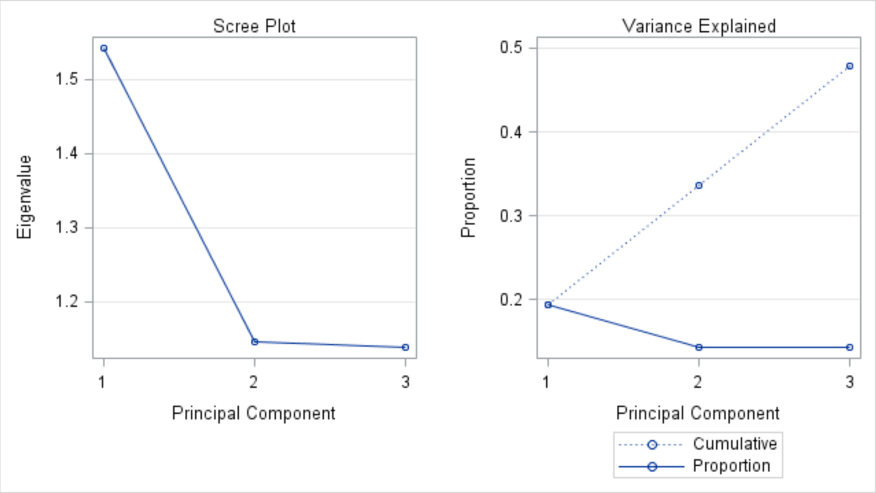
### Figure E.1. Scree Plot Results for Mortality Group



### Figure E.3. Scree Plot Results for Readmission Group



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



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

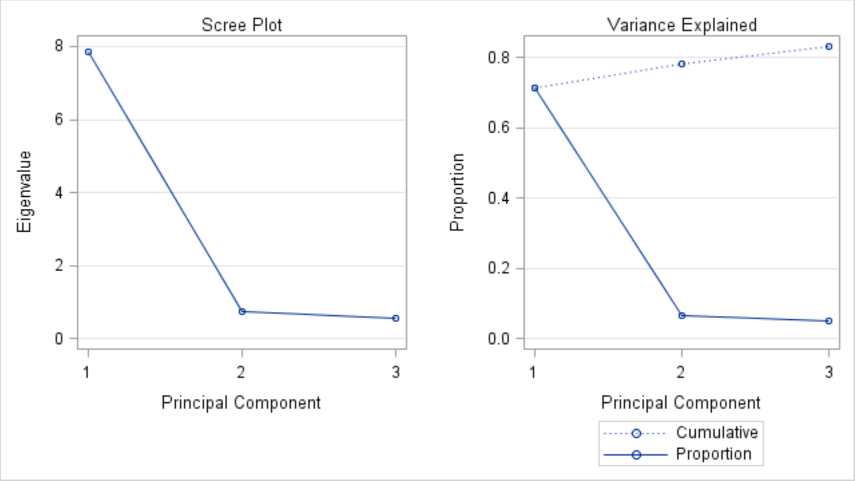


**Figure E.5. Scree Plot Results for Effectiveness of Care Group**

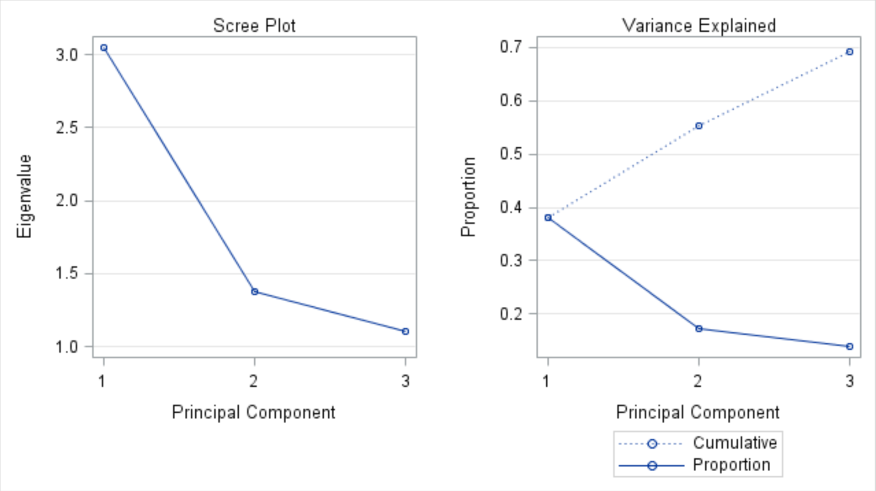


**Figure E.7. Scree Plot Results for Efficient Use of Medical Imaging Group**

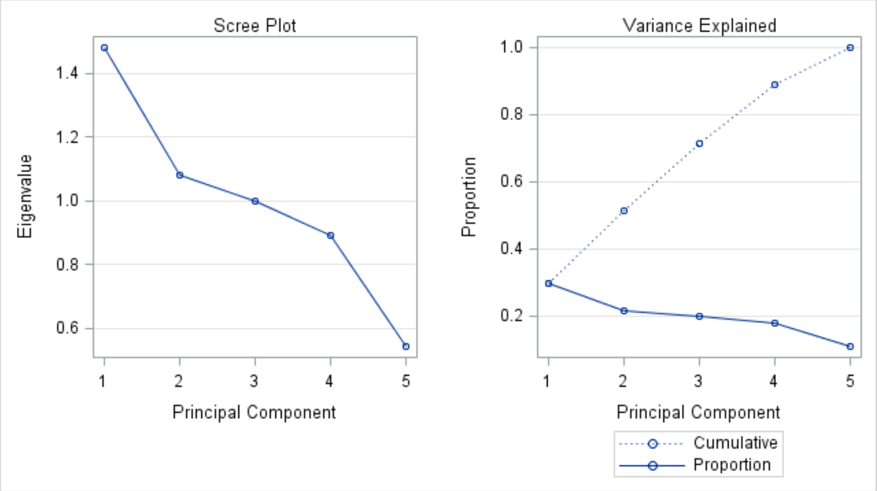
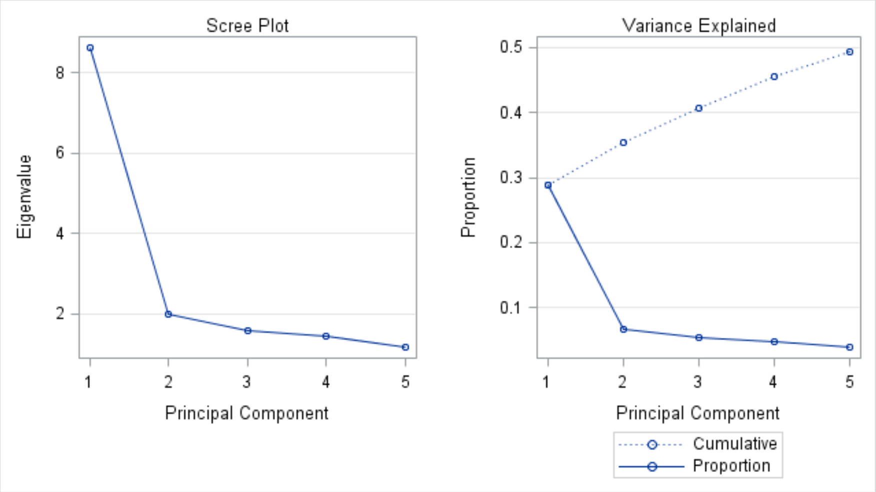


**Figure E.6. Scree Plot Results for Timeliness of Care Group**



## *Pairwise Comparison of Star Categories Testing Statistical Significance between Group Scores*

Table E.1. presents the results of CMS's pairwise comparison of Star Rating categories between group scores. The checkmarks (✓) presented in the table indicate that the difference between the mean group scores of the two compared Star Rating categories was statistically significant (p<0.05).

**Table E.1. Pairwise Comparison of Star Categories using Mean Group Scores by Group**

| Star Comparison | Mortality | | Safety of Care | | Readmission | | Patient Experience | | Effectiveness of Care | | Timeliness of Care | | Efficient Use of Medical Imaging | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Difference between group scores | p<0.05 | Difference between group scores | p<0.05 | Difference between group scores | p<0.05 | Difference between group scores | p<0.05 | Difference between group scores | p<0.05 | Difference between group scores | p<0.05 | Difference between group scores | p<0.05 |
| 5-4 | 0.34 | | 1.09 | ✓ | 1.56 | ✓ | 1.03 | ✓ | 0.61 | ✓ | 0.12 | | 0.03 | |
| 5-3 | 0.85 | ✓ | 1.49 | ✓ | 2.18 | ✓ | 2.05 | ✓ | 0.95 | ✓ | 0.31 | | 0.15 | |
| 5-2 | 1.19 | ✓ | 2.01 | ✓ | 3.16 | ✓ | 3.28 | ✓ | 1.08 | ✓ | 1.06 | ✓ | 0.23 | |
| 5-1 | 1.77 | ✓ | 3.30 | ✓ | 4.02 | ✓ | 5.00 | ✓ | 1.86 | ✓ | 2.74 | ✓ | 0.35 | |
| 4-5 | -0.34 | | -1.09 | ✓ | -1.56 | ✓ | -1.03 | ✓ | -0.61 | ✓ | -0.12 | | -0.03 | |
| 4-3 | 0.51 | ✓ | 0.40 | ✓ | 0.62 | ✓ | 1.02 | ✓ | 0.34 | ✓ | 0.19 | ✓ | 0.13 | ✓ |
| 4-2 | 0.84 | ✓ | 0.92 | ✓ | 1.60 | ✓ | 2.25 | ✓ | 0.47 | ✓ | 0.94 | ✓ | 0.21 | ✓ |
| 4-1 | 1.42 | ✓ | 2.21 | ✓ | 2.46 | ✓ | 3.97 | ✓ | 1.26 | ✓ | 2.62 | ✓ | 0.32 | |
| 3-5 | -0.85 | ✓ | -1.49 | ✓ | -2.18 | ✓ | -2.05 | ✓ | -0.95 | ✓ | -0.31 | | -0.15 | |
| 3-4 | -0.51 | ✓ | -0.40 | ✓ | -0.62 | ✓ | -1.02 | ✓ | -0.34 | ✓ | -0.19 | ✓ | -0.13 | ✓ |
| 3-2 | 0.34 | ✓ | 0.52 | ✓ | 0.98 | ✓ | 1.23 | ✓ | 0.13 | ✓ | 0.75 | ✓ | 0.08 | ✓ |
| 3-1 | 0.92 | ✓ | 1.80 | ✓ | 1.83 | ✓ | 2.95 | ✓ | 0.91 | | 2.43 | ✓ | 0.19 | |
| 2-5 | -1.19 | ✓ | -2.01 | ✓ | -3.16 | ✓ | -3.28 | ✓ | -1.08 | ✓ | -1.06 | ✓ | -0.23 | |
| 2-4 | -0.84 | ✓ | -0.92 | ✓ | -1.60 | ✓ | -2.25 | ✓ | -0.47 | ✓ | -0.94 | ✓ | -0.21 | ✓ |
| 2-3 | -0.34 | ✓ | -0.52 | ✓ | -0.98 | ✓ | -1.23 | ✓ | -0.13 | ✓ | -0.75 | ✓ | -0.08 | ✓ |
| 2-1 | 0.58 | | 1.29 | ✓ | 0.85 | ✓ | 1.72 | ✓ | 0.78 | | 1.68 | ✓ | 0.11 | |
| 1-5 | -1.77 | ✓ | -3.30 | ✓ | -4.02 | ✓ | -5.00 | ✓ | -1.86 | ✓ | -2.74 | ✓ | -0.35 | |
| 1-4 | -1.42 | ✓ | -2.21 | ✓ | -2.46 | ✓ | -3.97 | ✓ | -1.26 | ✓ | -2.62 | ✓ | -0.32 | |
| 1-3 | -0.92 | ✓ | -1.80 | ✓ | -1.83 | ✓ | -2.95 | ✓ | -0.91 | | -2.43 | ✓ | -0.19 | |
| 1-2 | -0.58 | | -1.29 | ✓ | -0.85 | ✓ | -1.72 | ✓ | -0.78 | | -1.68 | ✓ | -0.11 | |

## *Linear Trend of Star Rating Group Scores across Star Categories*

**Figure E.8. Mortality Group Scores across Star Categories**

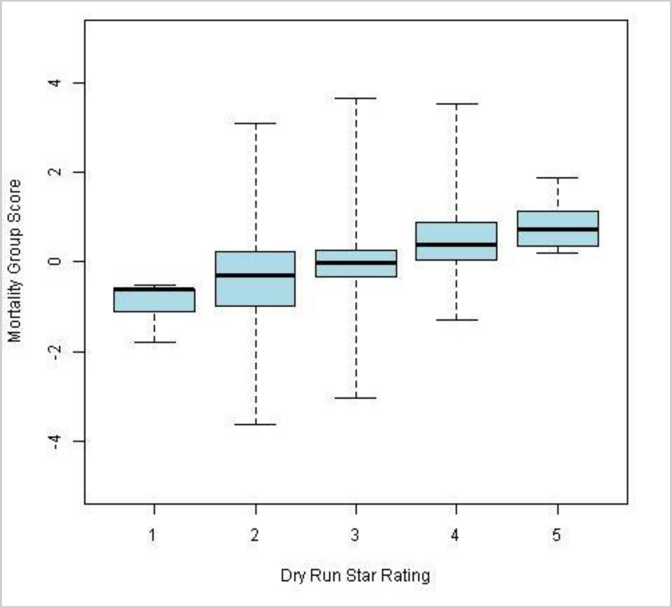


**Figure E.9. Safety of Care Group Scores across Star Categories**

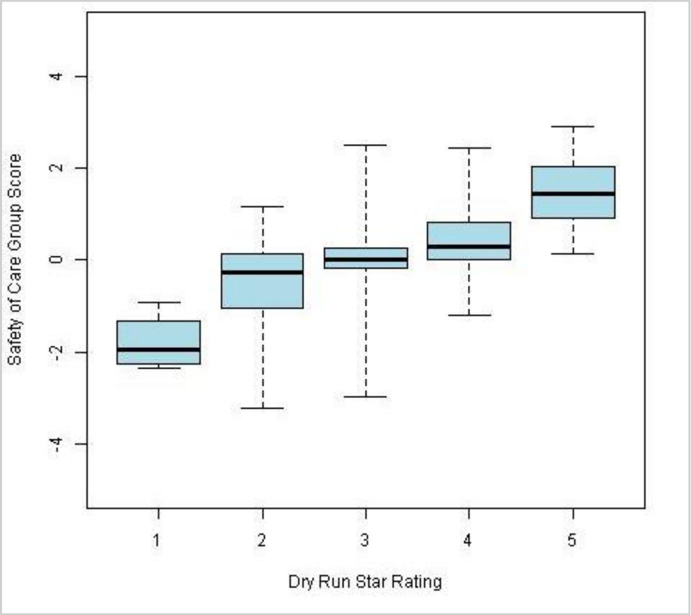


**Figure E.10. Readmission Group Scores across Star Categories**

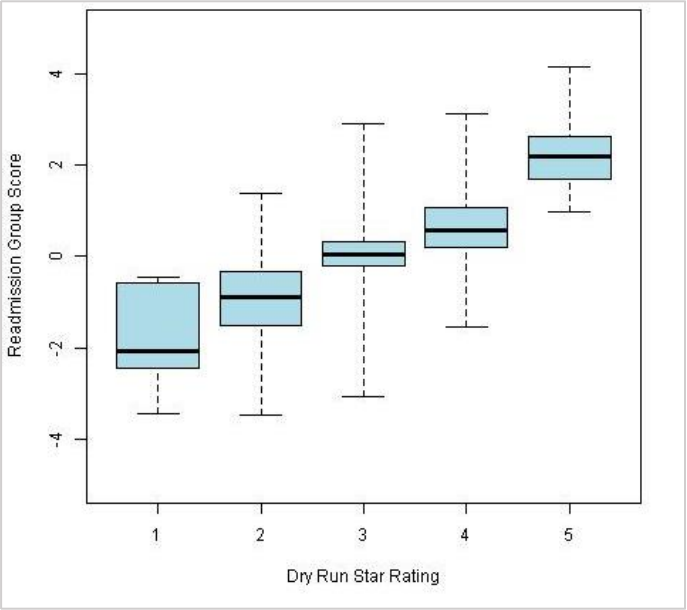


**Figure E.11. Patient Experience Group Scores across Star Categories**

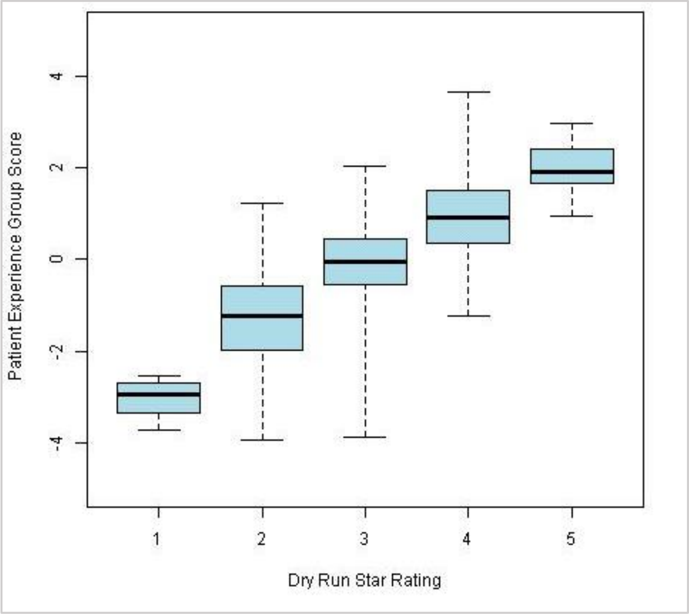


**Figure E.12. Effectiveness of Care Group Scores across Star Categories**

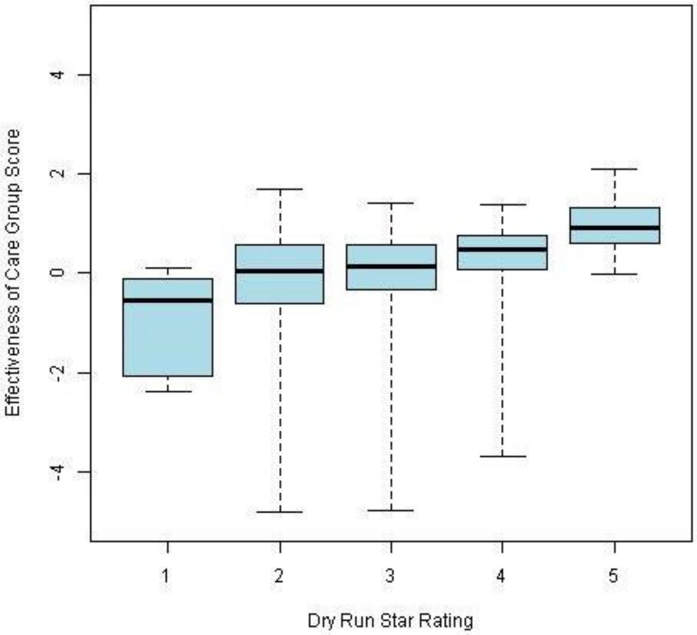


**Figure E.14. Efficient Use of Medical Imaging Group Scores across Star Categories**

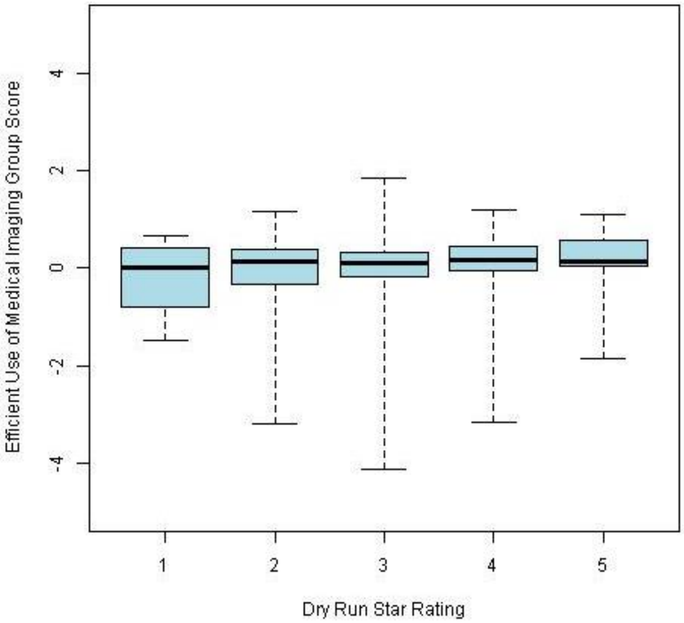
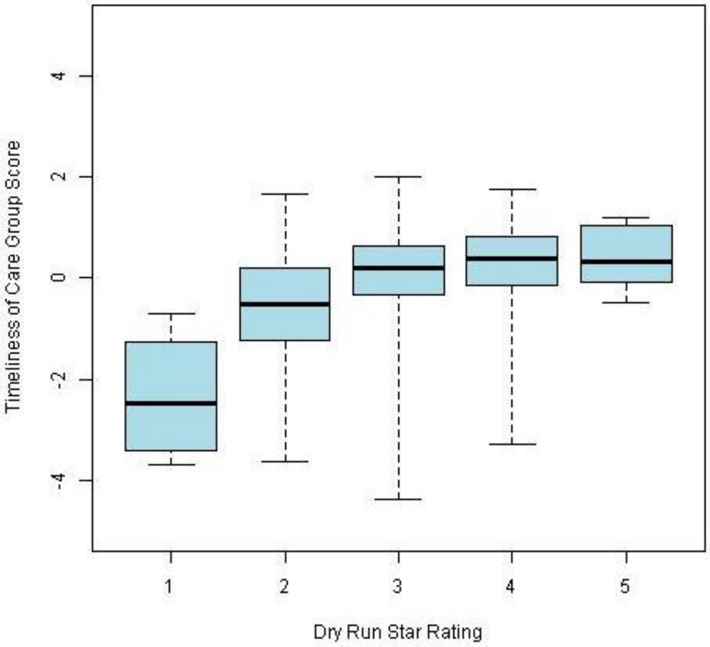


**Figure E.13. Timeliness of Care Group Scores across Star Categories**


Timeliness of Care Group Score
Dry Run Star Rating