Hospital Quality Star Ratings on *Hospital Compare*

July 2015 Dry Run

Methodology of Overall Hospital Quality Star Ratings

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

July 2015

Table of Contents

List of Tables

List of Figures

Section 1: Introduction

Within this report, the Centers for Medicare & Medicaid Services (CMS) is presenting the methodology under development for the Overall Hospital Quality Star Ratings to stakeholders in order to seek input. Exploration of the Star Ratings methodology is also a result of the requirement to summarize data on performance measures that was mandated in Sec. 399JJ of the Patient Protection and Affordable Care Act. This report describes the methodology used to calculate hospitals' scores during the dry run of Overall Hospital Quality Star Rating. Concurrent with this dry run, CMS will also hold a public comment period to seek public input on the methodology under development for the Overall Hospital Quality Star Ratings.

Goal of Project

Under contract with CMS, the Yale New Haven Health Services Corporation – Center for Outcomes Research & Evaluation (CORE), in collaboration with the Lantana Consulting Group, is developing the Overall Hospital Quality Star Ratings methodology. The star ratings would provide an overall quality rating for each hospital that currently has a sufficient amount of reported hospital quality information available on *Hospital Compare*. The *Hospital Compare* website would continue to provide the results of individual quality measures to inform consumers about the quality of care for particular conditions, procedures, and aspects of care such as patient experience.

The development team aims to develop an approach that provides patients and consumers with scientifically valid information in order to inform them about multiple dimensions of quality in a single measure. The star ratings on *Hospital Compare* would use a five-star rating system for consistency and alignment with existing CMS star ratings efforts for other providers. Key principles guiding the development effort are as follows:

- **Stakeholder Engagement:** Star Ratings development should seek stakeholder input from multiple channels from project start to finish.
- **Inclusivity:** Star Ratings should be as inclusive as possible of measures reported on *Hospital Compare.*
- **Consistency:** Star Ratings should be aligned with existing CMS programs and allow for consistent measure selection over time.
- **Simplicity and Accessibility:** Measures included in the Star Ratings should convey directional, evidence-based information.

The star ratings for the dry run summarize hospital quality based on existing measures, which capture some, though not all, aspects of quality through 75 measures across 7 distinct quality dimensions. As CMS develops more measures, the methodology for the star ratings would be able to capture hospital quality more comprehensively.

Stakeholder Engagement

The development team strove to engage stakeholders early on in the development of the methodology for the star ratings. The development team convened a 15-member technical expert panel (TEP) of patients, hospitals, purchasers, and star ratings developers. Following the first TEP meeting, CMS held a public comment period. A summary of the discussion from the three TEP meetings can be found under *the Hospital Quality Star Ratings on Hospital Compare* download at http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/CallforPublicComment.html.

The first public comment period was dedicated to soliciting input regarding the project objectives and the criteria to select measures for inclusion in the Overall Hospital Quality Star Ratings. The results of this public comment are available on the CMS Call for Public Comment webpage (http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/CallforPublicComment.html).

This dry run and the second public comment period seek input from a wide variety of stakeholders regarding several key decisions made during the development of the methodology including the analytic approach for summarizing individual measures, the use of weights to combine several aspects of quality into a single measure, and the approach to categorizing hospitals into star categories.

Purpose of Document

This report aims to highlight important technical and policy considerations for the public and hospitals. This document presents the methodology that will be used during the June 2015 Hospital Dry Run. Specifically, the document:

1. Describes the process for providing feedback during the Dry Run (Section 2)
2. Reviews the phases of Star Ratings development (Section 3)
3. Lists the measure exclusion criteria (Section 4)
4. Presents the approach for calculating hospital summary scores (Section 5)
5. Presents the approach for translating summary scores to star ratings (Section 6)
6. Discusses threshold for star ratings calculations (Section 7)

In the appendix sections, we provide a glossary of technical terms used throughout this report (Appendix A); the list of measures that would be included in the star ratings as of April 2015 (Appendix B); box-plots to describe hospital-level distribution of measure group scores in star categories (Appendix C); the loadings for each measure used in the statistical model (Appendix D); and hospital distributions by reported measures (Appendix E).

Section 2: Providing Feedback

CMS requests that hospitals and stakeholders submit comments on the methodology under development for the Overall Hospital Quality Star Ratings. CMS asks that stakeholders provide comments regarding the approaches to calculating hospital summary scores and translating summary scores to star ratings.

Instructions for Providing Comments

By email:

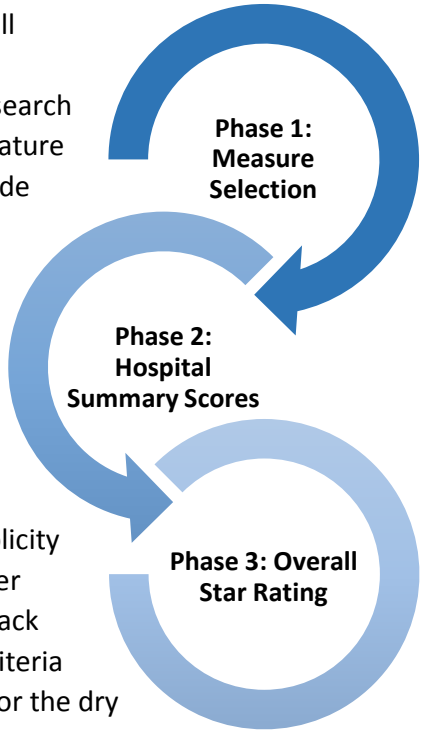
- If you are providing comments on behalf of an organization, include the organization's name and contact information.
- If you are commenting as an individual, submit identifying or contact information.
- Emailed comments are due by close of business August 17, 2015.
- **Please do not include personal health information in your comments**.
- Send your comments to cmsstarratings@lantanagroup.com.

Section 3: Phases of Development

In this section, we describe the key steps we took to establish the methodology for calculating an overall star rating.

Literature Review and Environment Scan

The development team began constructing the methodology for Overall Hospital Quality Star Ratings by both reviewing previous star ratings development experiences and conducting a comprehensive literature search on methodologies, best practices, and consumer preferences. The literature review and environmental scan supported the concept of bringing a wide variety of measures together into a single overall star rating for consumers. In addition, this work pointed to the need for extensive engagement and education of stakeholders to ensure that the methodology will facilitate stakeholder's understanding of the star ratings.



Phase 1: Measure Selection

To determine the appropriate set of measures for inclusion in the star ratings, the development team adhered to five guiding principles: simplicity and accessibility, inclusivity, scientific rigor, incorporation of stakeholder feedback, and consistency across CMS programs. After receiving feedback from the TEP and the first public comment period, CMS finalized the criteria for excluding measures from the Overall Hospital Quality Star Ratings for the dry run (Section 4).

Phase 2: Calculation of Hospital Summary Scores

During our second and third TEP meetings, we evaluated the options for summarizing the measure information available for each hospital into a single summary score. The methodology under development utilizes a two-stage approach (Section 5). First, measures are grouped by type (e.g., mortality, process, etc.), and a statistical model is applied to each group to generate a measure group score. Hospital quality is reflected by many dimensions, represented by the measure groups. The star ratings seek to combine these groups into a single measure of quality. In the second stage, the group scores are combined as a weighted average to produce the hospital summary score.

Phase 3: Translation of Summary Scores to Star Ratings

During the third TEP meeting, the development team also evaluated options for translating hospital summary scores into one of five star categories (Section 6). CMS continues to seek stakeholder input and feedback on the approach considered for categorizing hospital summary scores into one of five star ratings.

Section 4: Measure Exclusion Criteria

This section describes the criteria to determine which measures from *Hospital Compare* are included in the calculation of the Overall Hospital Quality Star Ratings. These criteria were vetted through the TEP and first public comment period and are being used for the dry run.

The Overall Hospital Quality Star Ratings project strives to be as inclusive as possible of existing hospital quality measures in order to fully represent the information available to consumers on *Hospital Compare* in a star rating. However, some measures that are currently included on *Hospital Compare* would be excluded from the calculation of the Overall Hospital Quality Star Ratings. Certain measures would be excluded in order to generate a star rating based on measures that are actively collected and reported, widely available, suitable for combination, and interpretable by patients and consumers.

As of April 2015, there are 106 hospital quality measures potentially available for public reporting on *Hospital Compare*. These measures represent a variety of measure types and cover a broad set of clinical conditions and care processes. In order to understand the variety of the quality measures currently on *Hospital Compare*, the development team abstracted important measure details and guidance from FY 2015 rulemaking and sub-regulatory reports, Measure Applications Partnership Reports, the *Hospital Compare* website, 2014 Inpatient Quality Reporting (IQR) Measure Comparison Table, and other publicly published resources.[1-10] We used this information to exclude measures according to the final exclusion criteria determined by CMS following the first TEP meeting and public comment period. Please note that the dry run star ratings methodology includes only the composite measure PSI-90 and PSI-4. The methodology does not double count as both individual measures and a composite.

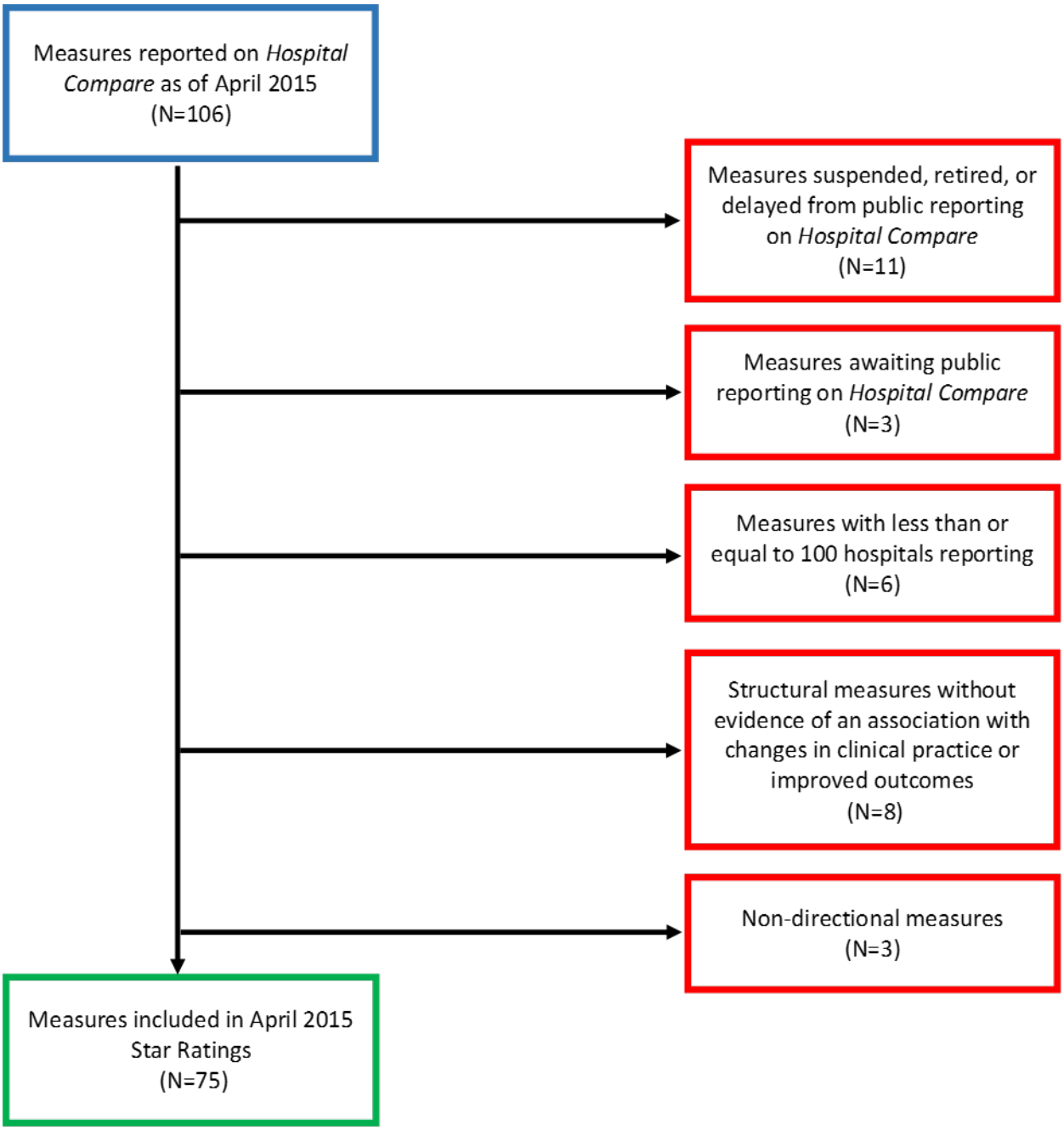
Seventy-five (Table B.1) of the 106 potentially reportable measures on *Hospital Compare* as of April 2015 are recommended for inclusion. A flowchart that diagrams the measure selection process is included in Figure 1.

Criteria for Measure Exclusion

The following criteria will be used for the hospital dry run to exclude measures from star ratings:

1. Measures suspended, retired, or delayed from public reporting on *Hospital Compare*;
2. Measures awaiting public reporting on *Hospital Compare;*
3. Measures with less than or equal to 100 hospitals reporting performance publicly;
4. Structural measures without evidence of an association with changes in clinical practice or improved outcomes (e.g., volume and registry participation); and
5. Measures for which it is not clear whether a higher or lower scores is better (i.e., non-directional measures).

Figure 1. Measure Selection Flowchart (April 2015 data)



Section 5: Approach under Development for Calculating Hospital Summary Scores

The methodology under development for Overall Hospital Quality Star Ratings uses a two-stage approach to calculating hospital summary scores. We are seeking stakeholder feedback on this approach in general and with respect to several key elements.

The development team considered various approaches, including simple or weighted averages of all the measures and more complex statistical approaches utilizing factor analysis and latent variable models. The development team evaluated each approach in the context of the project goals and timeline.

The development team sought to identify an approach that would:

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

Ultimately, the development team consulted the TEP and expert stakeholders within CMS to reach a two-stage approach which employs latent variable modeling (LVM) and weighted averaging to meet these principles. Within this section we describe each of the specific steps following measure selection necessary to establish a hospital's star rating:

1. Standardization of measure scores into a common format
2. Potential measure groups to summarize aspects of quality
3. Two-stage approach to calculating hospital summary score

Standardization

Before combining measures into a score, each measure is first converted into a common scale of measurement. Hospital quality measure results include many different types of scoring information, 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, we used standardization to ensure all measure scores were in a common scale and in a common direction. This does not change the measure information – just the scale for scoring in order to make it possible to bring them together in the hospital star rating calculation. Specifically, we standardized

a hospital's score on each measure by calculating "Z" scores for each measure, reversing if necessary so that larger 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, OP-21 (Median Time to Pain Management for Fractures) has 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) has a national average of 7.23% with a standard deviation of 9.10%. After standardization and redirection, both measures have a mean score of 0 and standard deviation of 1; a higher standardized score indicates better quality. For an individual hospital with an OP-21 score of 65 minutes, the standardized score would be -0.53, while the standardized score for a hospital with a score of 45 minutes would be 0.602. Henceforth in this report, a measure score refers to the standardized measure score or "Z" score.

We further winsorize the standardized measure score at the 0.125th percentile (Z=-3) and the 99.875 percentile (Z=3) of a Standard Normal distribution to avoid extreme outlier performance for which it is unclear if the reported measure score represents extreme performance or potentially inaccurate reporting. All standardized scores above 3 are set to be 3 and all standardized scores bellow -3 are set to be -3.

Proposed Measure Groups

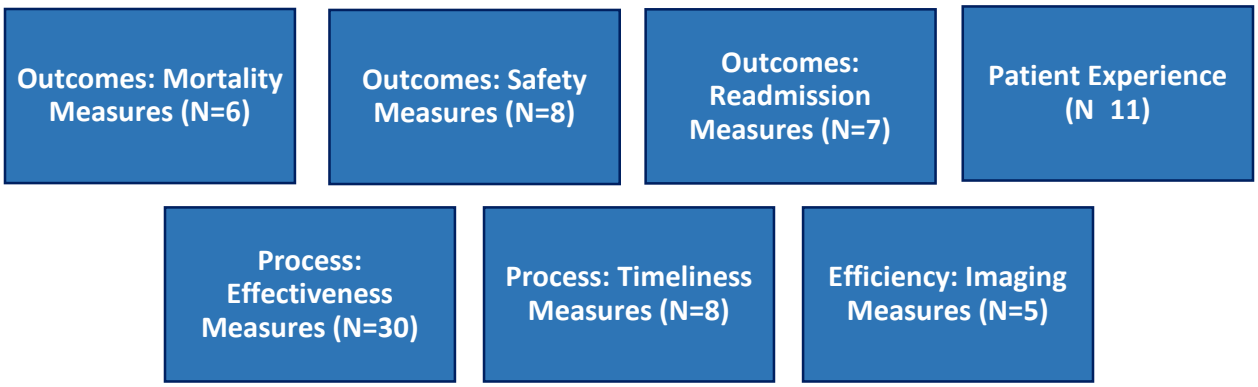
Hospital quality is represented by several dimensions, ranging from clinical care processes to readmission reduction initiatives focused on care transitions to patients' experiences. The development team evaluated several options for organizing quality measures into mutually exclusive conceptual groups. Each group accounts for measures that represent similar concepts of quality. For the two-stage approach, the development team has grouped measures into seven groups based on the type of measure and with support from the TEP (Outcomes – Mortality, Outcomes – Safety, Outcomes – Readmissions, Patient Experience, Process – Effectiveness, Process – Timeliness, and Efficiency – Imaging) (Appendix B). The rationale for these seven measure groups is as follows:

- The seven measure 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.
- Measure groups are clinically reasonable in that they capture common components of quality for which hospital quality is likely linked across measures. For example, the degree to which hospitals effectively manage care transitions or safely discharge patients is likely to be reflected in all readmission measures.

The proposed measure groups will allow for future measures to be added and removed from the star ratings. This option permits the mutually exclusive assignment of measures to each group,

whereas alternative approaches may be more subjective and require substantial deliberation and compromise to ensure consistency.

Figure 2. Proposed Measure Groups



Application of these measure groups to the April 2015 dataset, available for this dry run, demonstrates marked variability in hospital-level reporting of quality information. While 75 measures are available in total, the average hospital reported 44.8 measures in April 2015 (IQR: 19 to 69). We describe the distribution of hospital quality reporting overall and by measure group for the 75 measures included in this dry run methodology based on the April 2015 reporting period in Appendix E.

Overview of Approach to Hospital Summary Calculation

The current methodology calculates the final hospital summary score in two stages. In the first stage, the seven measure groups are used to calculate measure group scores, one for each group. This is derived for each hospital by fitting a latent variable model (LVM) on their standardized measure scores for each measure group. The LVM produces a hospital-specific group score that reflects the information about each aspects of quality that is conveyed by the available measures within a group, accounting for different measure availability across hospitals. In the second stage, the seven group scores are combined into one single overall hospital summary score using a policy-based weighting.

Below are the statistical equations describing the first stage using LVM (Equation 1) with weighted likelihood (Equation 2) and the second stage that calculates a weighted average (Equation 3).

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

$$Y_{khd} = \mu_{kd} + \gamma_{kd}\alpha_{hd} + \varepsilon_{khd},\ k=1,\ldots,N_d$$
$$\alpha_{hd} \sim N(0,1) \text{ and } \varepsilon_{khd} \sim N(0, \sigma_d^2)$$

Let Y_{khd} denote the standardized score for hospital *h* and measure *k* in measure group *d*. α_{hd} is the hospital-specific group-level latent trait (random effect) for hospital *h* and measure group *d*. γ_{kd} is the loading (coefficient) for measure *k*, which shows the relationship with the group score of measure group *d*. N_d is the total number of measures in measure group *d*. α_{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 μ_{kd} and γ_{kd}.

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

$$L = \prod_{k=1}^{K} \prod_{h=1}^{H} (L(y_{khd}))^{w_{khd}}$$

$$w_{khd} = \frac{n_{khd}}{\sum_{h=1}^{N_{kd}} n_{khd}} \times N_{kd}$$

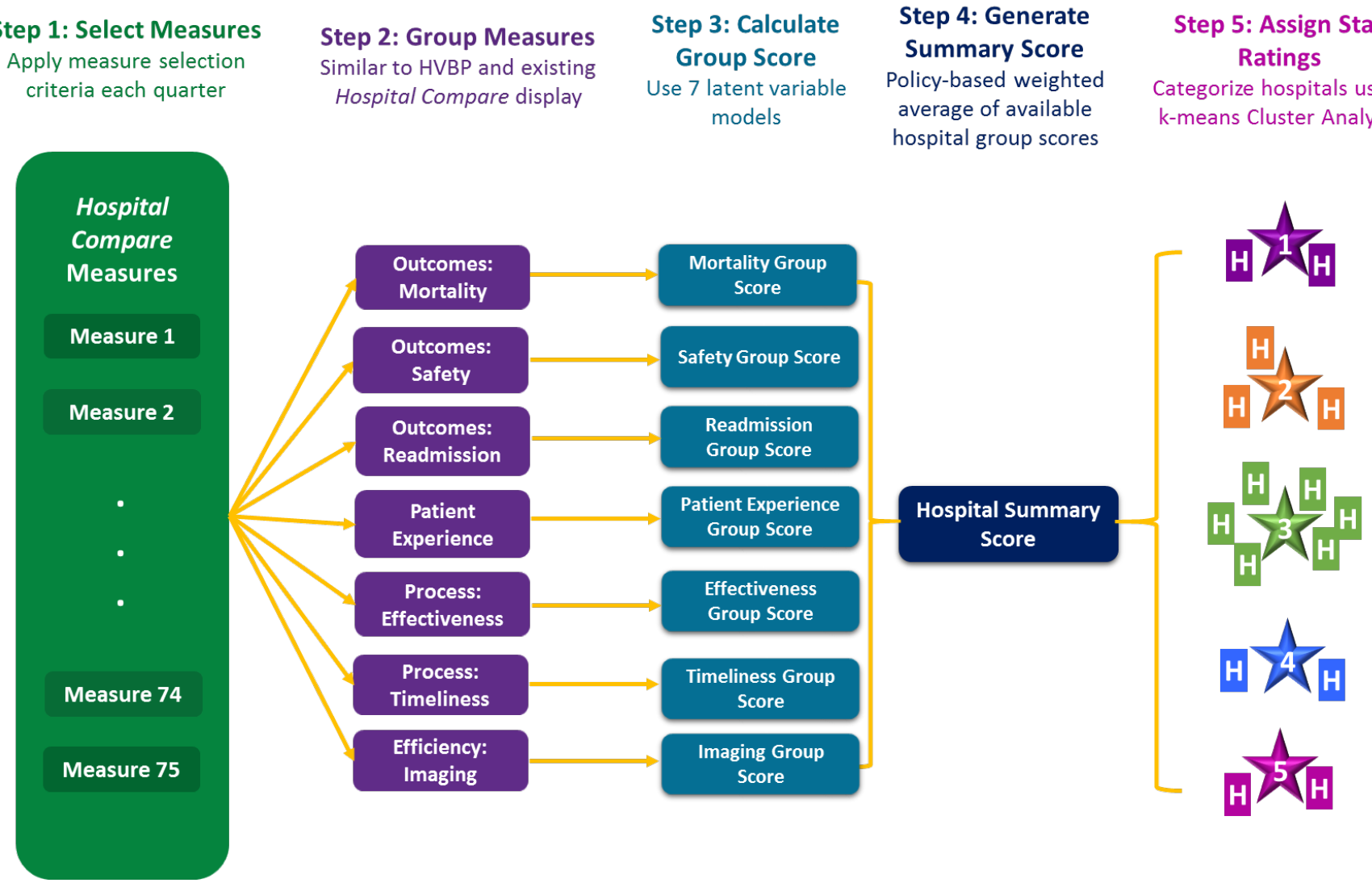
A weighted likelihood is used to account for sampling variation of each hospital for each measure. *L* is the likelihood function. N_{kd} is the total number of hospitals for measure *k* in measure group *d* and n_{khd} is the denominator for hospital *h* and measure *k* in measure group *d*. A hospital with larger denominator will be weighted more in the LVM.

Equation 3. Calculation of Hospital Summary Score from Group Scores

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

For Equation 3, the proposed weights, w_d, are listed in Table 2.

Figure 3. Process for Calculating the Overall Hospital Quality Star Ratings



Stage 1: Group-Specific Latent Variable Modeling (LVM)

The first stage of the two-stage approach utilizes LVM to calculate a group score for each of a hospital's measure groups. LVM derives an estimate for the hospital-specific score by modeling the individual measures within a group through a latent variable to reflect the aspect of quality represented by those measures. The LVM approach confers several advantages making it well-suited for the star ratings.

Figure 4. Detailed Path Diagram of Two-Stage Approach

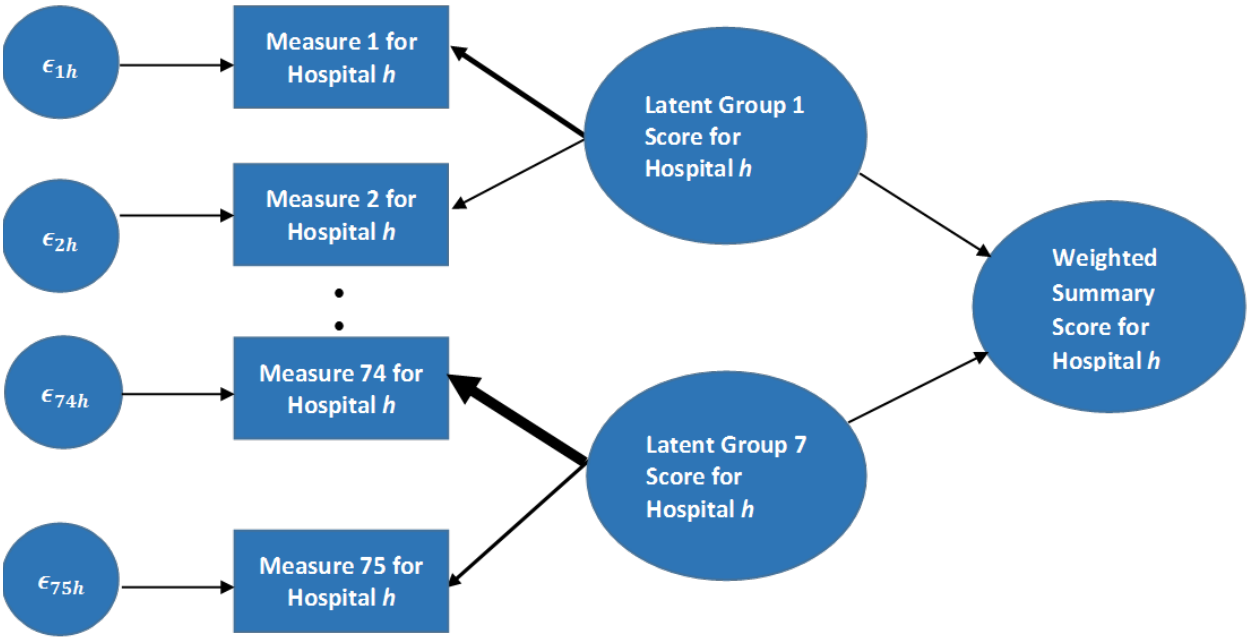


FIGURE DESCRIPTION: The ovals represent the group scores and the hospital summary score. The group score is the latent trait in Equation 1. This number is not directly observed, but is inferred from the individual measures Y_1, …, Y_{75}. The arrows between the group scores and each individual measure 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 75 measures. The residual error (ε) is the variation which could not be explained by the group score (random effect). This example latent variable model can be estimated using standard software, SAS Proc NLMIXED. The 'loadings' are estimated by maximum likelihood method and the group scores for each hospital are estimated by empirical Bayes estimates. A weighting scheme then is applied to the measure groups, and the group scores are averaged using weights to create the hospital summary score. Ultimately, the hospital summary score is used to organize hospitals into overall star rating categories.

Advantages of LVM to Calculate Group Scores

- Method is used for composite measures in healthcare quality literature.[11]
- The 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 only available information to generate a group score so that hospitals with limited information will not be calculated to have extreme group scores.
- The model can account for sampling variance, or reflect the differences in precisions for each hospital's individual measure score as a result of different hospital volumes counted for each measure.

Challenges of LVM to Calculate Group Scores

- The modeling technique may be challenging for patients and consumers to understand.
- Each LVM assumes that each group reflects a distinct aspect of quality. Each measure contributes to one group score even if it may potentially reflect more than one aspect of quality.
- Each included measure is a valid indicator of quality.

Measure Loadings

The LVM estimates a "loading" for each measure in a group that is associated with the hospital-specific group score (Appendix D). The loading is the extent of the measure's association to the group score (latent aspect of quality) relative to the other measures included in the group. Key considerations for measure loadings include:

- A measure's loading is the same across hospitals that are associated with hospital-specific group scores. The loadings are different by 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 measure 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.

Accounting for Measure Sampling Variation

As part of the statistical model to obtain groups scores, the development team also accounts for the sampling variation of each measure for each hospital. The development team uses the hospital's

measure denominator as an approximation of the sampling variation. A weighted likelihood is used to ensure that a hospital with a larger denominator, or a more precise measure score, would be weighted more in the LVM.

Stage 2: Weighted Average of Group-Specific Scores

The second stage of the two-stage approach applies weights to each of the seven group-specific scores to generate an overall summary score for each hospital. We aimed to determine weights for each group that represent the preferences of stakeholders, particularly patients and consumers.

CORE presented two potential options to the TEP and CMS; equal weighting and weighting modified from the FY 2017 HVBP program. To obtain consensus on the TEP's preferences for weighting, the development team administered an online survey asking the TEP to rank the following measure groups: Outcomes – Mortality, Outcomes – Safety, Outcomes – Readmission, Patient Experience, Process, and Efficiency. The results of this survey are summarized in Table 1. The TEP ranked the groups on a scale of 0-10 with higher numbers indicating higher importance. Please note, the development team separated the Process group into Process – Effectiveness and Process – Timeliness following the final TEP meeting to better account for the heterogeneity of the process measures included in the Overall Hospital Quality Star Ratings dry run.

Table 1. TEP Survey Results Ranking Measure Groups (N=13 TEP Members)

Measure Group	Mean Ranking by TEP
Outcomes – Mortality	7
Outcomes – Safety	8
Outcomes – Readmission	7
Patient Experience	6
Process	4
Efficiency – Imaging	5

We applied the following criteria in our approach to determining an appropriate weighting scheme:

- Measure importance
  - The weight of outcome measure groups should be greater than that of process measure groups.
  - The weight of the efficiency – imaging measure group should take into account the limited population captured by these measures.
- Consistency
  - The weights should align with the existing weighting schemes of other CMS programs to ensure consistent incentives.
- Policy priorities
  - The weights should reflect CMS's priorities as reflected in the CMS Quality Strategy.[12]

- Stakeholder input
    - The proposed weights should reflect the prioritization of measure groups by the TEP as well as future feedback via public comment periods, the hospital dry run, and additional sources of patient and consumer feedback.

Given the TEP's feedback and these criteria, CMS is considering using a policy-based weighting scheme modified from HVBP. We used the weighting below (Table 2) in alignment with HVBP for the dry run. An additionally proposed weighting scheme that is not used for the dry run is also presented in the Public Comment report.

Table 2. Proposed Policy-based Weighting Scheme Modified from FY 2017 HVBP

Measure Group	FY17 HVBP Weight	Dry Run Proposed Weight
Outcomes – Mortality (N=6)	25%	22%
Outcomes – Safety (N=8)	20%	22%
Outcomes – Readmission (N=7)	---	22%
Patient Experience (N=11)	25%	22%
Process – Effectiveness (N=30)	5%	4%
Process – Timeliness (N=8)	---	4%
Efficiency – Imaging (N=5)	---	4%
Efficiency – Cost	25%	---

Note: The FY 2017 HVBP program is set to include 22 measures and assigned a 25% weight to Medicare Spending per Beneficiary measure, which is not included in Star Ratings.

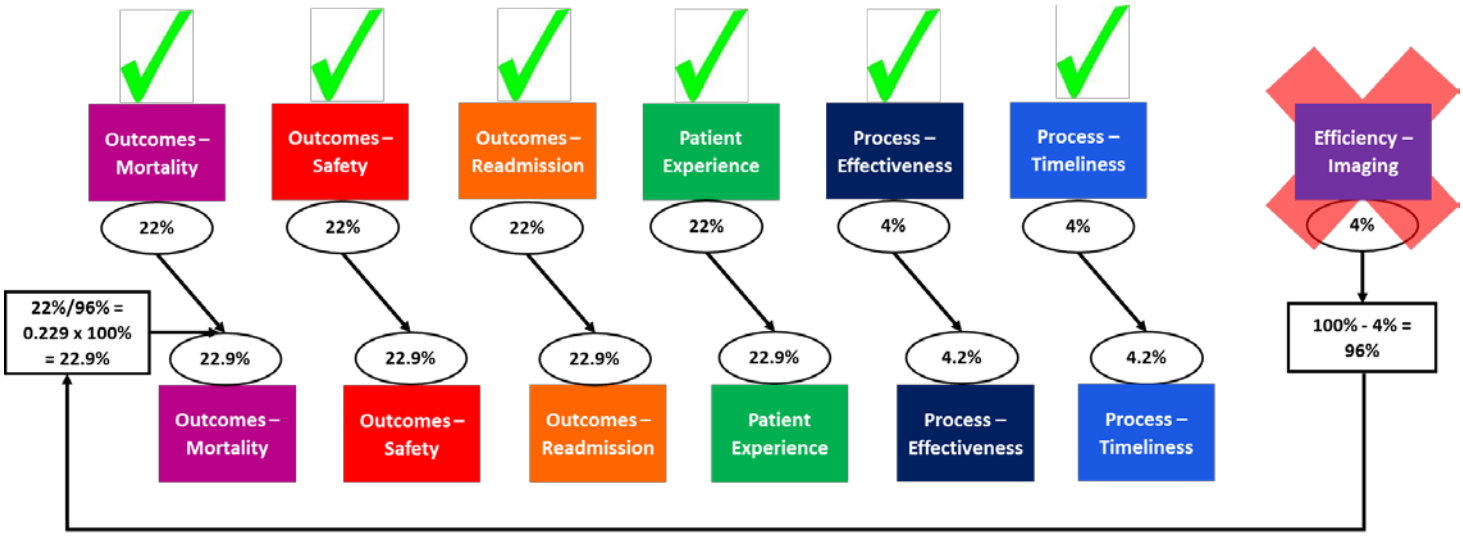
Proposed Method for Weighting When Missing Group(s)

In some cases, a hospital may not have measures reported for every measure group. The development team has recommended following the approach taken by the HVBP program in which weights of missing groups are re-proportioned across the groups that a hospital reports. If a hospital reports zero measures in a group, the group is considered missing. An example adjusted weighting scheme accounting for missing groups is shown in Table 3 and Figure 5.

Table 3. Example of Re-weighting Scheme when Missing Outcomes – Safety Measures

Measure Group	Dry Run Proposed Weight	Re proportioned Weight
Outcomes – Mortality (N=6)	22	22.9
Outcomes – Safety (N=8)	22	22.9
Outcomes – Readmissions (N=7)	22	22.9
Patient Experience (N=11)	22	22.9
Process – Effectiveness (N=30)	4	4.2
Process – Timeliness (N=8)	4	4.2
Efficiency – Imaging **(Missing; N=0)**	4	---

Figure 5. Example of Re-weighting Scheme when Missing Outcomes – Safety Measures



Section 6: Translating Summary Scores to Stars

This section describes the methodology for translating summary scores into star ratings once hospital summary scores are generated.

Assumptions:
There are several important assumptions to consider prior to determining the approach for translating hospital summary scores into one of five star categories:

- Hospitals will always have summary scores at the margin of a star category (in other words, 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 measure 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 overall star rating.
- Star ratings are not intended to guide specific hospital quality improvement efforts, but rather to make summary information available to the public.

CORE considered several options with the TEP for translating summary scores to stars. For the dry run, we present the use of k-means clustering to assign hospital star ratings.

k-Means Clustering

k-Means Clustering is the method used in the dry run for translating hospital summary scores into star ratings. This is a method for creating groups (or clusters) of hospitals so that scores in each cluster are closer to their group mean than to any other group mean. In the case of a summary score, *k*-means cluster analysis with k=5 can be used to categorize hospitals into a star rating category such that hospitals in each star category have an overall score that is 'more like' the other hospitals in that star category than it is like hospitals in different categories.

This approach is based solely on the hospital summary scores. Table 4 presents the frequency of star ratings using *k*-means clustering.

Example Categories for Clustering

★★★★★	Cluster of hospitals with highest summary scores determined by the sum of the square of distance between hospital's summary scores
★★★★	Cluster of hospitals with higher than average summary scores determined by the sum of the square of distance between hospital's summary scores
★★★	Cluster of hospitals with average summary scores determined by the sum of the square of distance between hospital's summary scores
★★	Cluster of hospitals with below average summary scores determined by the sum of the square of distance between hospital's summary scores
★	Cluster of hospitals with lowest summary scores determined by the sum of the square of distance between hospital's summary scores

Advantages of *k*-Means Clustering Approach

- *k*-means optimally designated 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.
- An analysis conducted for validation broadly demonstrates statistically different group scores between each star rating category in many groups supporting the ability of this approach to distinguish hospital performance across these five clusters. We found statistically significant differences in measure group scores between each star rating category for every measure group except Outcomes-Mortality, Efficiency, Process-Timeliness, and Process-Effectiveness. For these groups, we found statistically significant differences in measure group scores for vast majority comparisons between star ratings, except Efficiency group.
- We constructed box-plots to describe the hospital-level distribution of measure group scores for hospitals in each star ratings category (Appendix C).
- We further tested for a linear trend of measure group scores between each star rating category. We found a statistically significant (p=0.05) linear trend for each measure group confirming the notion that higher measure group scores are associated with higher star ratings.

Challenges of *k*-Means Clustering Approach

- The majority of hospitals will fall into the three-star cluster.
- The complexity of approach may be difficult for patients and consumers to understand.
- The approach is "tournament-based" in comparison to an "achievement-based" approach. In other words, the approach calculates a hospital's star rating relative to other hospitals based on the distribution of hospital summary scores. An alternative approach or "achievement-based" approach would utilize an absolute threshold for each star category

(for example, a hospital with high performance in six out of the seven groups would be a five-star hospital).

Table 4. Frequency of Star Ratings in Dry Run Using k-Means Clustering (April 2015 data)

Rating	Frequency (Number of Hospitals)	Minimum Summary Score in Cluster	Maximum Summary Score in Cluster
1 Star	5	-2.40	-1.58
2 Star	544	-1.53	-0.43
3 Star	2615	-0.43	0.38
4 Star	528	0.38	1.31
5 Star	17	1.37	2.21

Note: The total number of hospitals in the *Hospital Compare* dataset as of April 2015 is 4,746 hospitals. Results shown are for all hospitals meeting the reporting criteria discussed in Section 7 (N=3,709).

Section 7: Thresholds for Receiving a Star Rating during Dry Run

A few hospitals may not report many individual measures resulting in the generation of hospital summary scores of lower reliability and face validity. HVBP addresses this issue by setting several thresholds for measure groups in order for a hospital to be eligible for a Total Performance Score Calculation. The development team sought the TEP's feedback on setting measure and measure group thresholds in order for a hospital to be eligible to receive an overall star rating. Given HVBP's threshold and the TEP's feedback, the development team proposes setting a minimum measure threshold guided by a reliability calculation and expert input. Setting a minimum measure threshold of three for each measure group would exceed a desired reliability level of 0.75 for all measure groups (Table 5). The development team proposes a minimum measure group threshold similar to HVBP, requiring hospitals report at least three of the seven groups with one being an outcome group. Together, these individual measure and measure group thresholds results in 78% of hospitals receiving a star rating for the dry run (Table 6).

The minimum measure and minimum group thresholds are applied solely for dry run reporting purposes and have no effect on the calculation of the hospital summary score or the star categorization. In other words, if a hospital meets these thresholds and has additional measure groups, each with one to two measures, these group scores will be included in their star rating using standard weights.

Table 5. Minimum Measure Thresholds Using Reliability Calculation

Domain	Measures	R	Required N	R	Required N	R	Required N	R	Required N
HCAHPS	11	0.6	0.73	0.7	1.14	0.75	1.46	0.8	1.95
Readmission	7	0.6	1.21	0.7	1.89	0.75	2.43	0.8	3.23
Mortality	6	0.6	1.28	0.7	1.99	0.75	2.56	0.8	3.41
Safety	8	0.6	1.14	0.7	1.78	0.75	2.28	0.8	3.05
Efficiency	5	0.6	0.98	0.7	1.52	0.75	1.96	0.8	2.61
Process-Effectiveness	30	0.6	0.90	0.7	1.40	0.75	1.80	0.8	2.41
Process-Timeliness	8	0.6	0.80	0.7	1.24	0.75	1.60	0.8	2.13

Table 6. Hospitals (N=4,746) with reported Overall Star Rating based on minimum thresholds (April 2015 data)

	Minimum Measure Groups					
Minimum Measures	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 measure groups shown in Table 7 must include at least one outcome group.

Considerations for Proposed Minimum Thresholds

- Setting increasingly high thresholds for both measures and measure groups would have excluded more hospitals from the Overall Hospital Quality Star Ratings dry run.
    - The development team will seek patients' and hospitals' input regarding the face validity of these thresholds and how hospitals with fewer measures are displayed.

References

1. 42 CFR Parts 405, 413, et al. Medicare Program; Hospital Inpatient Prospective Payment Systems for Acute Care Hospitals and the Long-Term Care Hospital Prospective Payment System and Fiscal Year 2015 Rates; Quality Reporting Requirements for Specific Providers; Reasonable Compensation Equivalents for Physician Services in Excluded Hospitals and Certain Teaching Hospitals; Provider Administrative Appeals and Judicial Review; Enforcement Provisions for Organ Transplant Centers; and Electronic Health Record (EHR) Incentive Program; Final Rule. In: U.S. Department of Health and Human Services (HHS), ed. 2014-18545: Federal Register; 2014:597.
2. Centers for Medicare & Medicaid Services (CMS). CMS Measures Inventory. 2014.
3. Medicare Hospital Compare Quality of Care. Centers for Medicare & Medicaid Services, 2015. at http://www.medicare.gov/hospitalcompare/)
4. Centers for Medicare & Medicaid Services (CMS). CMS Hospital Inpatient Quality Reporting Program Measure Comparison Table Calendar Year 2015 (Fiscal Year 2017). In: IQR_CMS_MsrsComp_CY2015, ed. 2015.
5. Institute of Medicine (IOM). Performance Measurement: Accelerating Improvement (Pathways to Quality Health Care Series). Washington, DC 2006.
6. Measure Applications Partnership (MAP). Pre-Rulemaking Report: Input on Measures Under Consideration by HHS for 2012 Rulemaking. Washington, DC 2012.
7. Measure Applications Partnership (MAP). MAP Pre-Rulemaking Report: 2013 Recommendations on Measures Under Consideration by HHS. Washington, D C2013.
8. Measure Applications Partnership (MAP). MAP 2014 Recommendations on Measures for More Than 20 Federal Programs. Washington, DC 2014.
9. Measure Applications Partnership (MAP). MAP Families of Measures: Safety, Care Coordination, Cardiovascular Conditions, Diabetes. Washington, DC 2012.
10. Health Services Advisory Group I. National Impact Assessment of Medicare Quality Measures 2012.
11. 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. Medical care 2008;46:778-85.
12. Centers for Medicare & Medicaid Services (CMS). CMS Quality Strategy: 2013 - Beyond. In: U.S. Department of Health and Human Services (HHS), ed. Baltimore, MD 2013:34.

Appendix A: Introduction to Statistical Terminology

In this Appendix, we define the statistical terms relevant to our initial analyses. We intend for this section to help streamline communication and develop a common, foundational understanding of the approaches to be 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 (summary score). It indicates the strength of the relationship on the path from the latent variable to the indicator.
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: Measure Selection

Table B.1. Measures by Measure Group (April 2015)

Measure Name	Measure Group
MORT-30-AMI Acute Myocardial Infarction (AMI) 30-Day Mortality Rate	Outcomes – Mortality
MORT-30-COPD Chronic Obstructive Pulmonary Disease (COPD) 30-Day Mortality Rate	Outcomes – Mortality
MORT-30-HF Heart Failure (HF) 30-Day Mortality Rate	Outcomes – Mortality
MORT-30-PN Pneumonia (PN) 30-Day Mortality Rate	Outcomes – Mortality
MORT-30-STK Acute Ischemic Stroke (STK) 30-Day Mortality Rate	Outcomes – Mortality
PSI-4-SURG-COMP Death Among Surgical Patients with Serious Treatable Complications	Outcomes – Mortality
HAI-1 Central-Line Associated Bloodstream Infection (CLABSI)	Outcomes – Safety
HAI-2 Catheter-Associated Urinary Tract Infection (CAUTI)	Outcomes – Safety
HAI-3 Surgical Site Infection from colon surgery (SSI-colon)	Outcomes – Safety
HAI-4 Surgical Site Infection from abdominal hysterectomy (SSI-abdominal hysterectomy)	Outcomes – Safety
HAI-5 MRSA Bacteremia	Outcomes – Safety
HAI-6 *Clostridium Difficile* (*C.difficile*)	Outcomes – Safety
COMP-HIP-KNEE Hospital-Level Risk-Standardized Complication Rate (RSCR) Following Elective Primary Total Hip Arthroplasty (THA) and Total Knee Arthroplasty (TKA)	Outcomes – Safety
PSI-90-Safety Complication/Patient Safety for Selected Indicators (PSI)	Outcomes – Safety
READM-30-AMI Acute Myocardial Infarction (AMI) 30-Day Readmission Rate	Outcomes – Readmission
READM-30-COPD Chronic Obstructive Pulmonary Disease (COPD) 30-Day Readmission Rate	Outcomes – Readmission
READM-30-HF Heart Failure (HF) 30-Day Readmission Rate	Outcomes – Readmission
READM-30-Hip-Knee Hospital-Level 30-Day All-Cause Risk- Standardized Readmission Rate (RSRR) Following Elective Total Hip Arthroplasty (THA)/Total Knee Arthroplasty (TKA)	Outcomes – Readmission
READM-30-PN Pneumonia (PN) 30-Day Readmission Rate	Outcomes – Readmission
READM-30-STK Stroke (STK) 30-Day Readmission Rate	Outcomes – Readmission
READM-30-HOSP-WIDE HWR Hospital-Wide All-Cause Unplanned Readmission	Outcomes – Readmission
H-CLEAN-HSP Cleanliness of Hospital Environment (Q8)	Patient Experience
H-COMP-1 Nurse Communication (Q1, Q2, Q3)	Patient Experience
H-COMP-2 Doctor Communication (Q5, Q6, Q7)	Patient Experience

Measure Name	Measure Group
H-COMP-3 Responsiveness of Hospital Staff (Q4, Q11)	Patient Experience
H-COMP-4 Pain management (Q13, Q14)	Patient Experience
H-COMP-5 Communication About Medicines (Q16, Q17)	Patient Experience
H-COMP-6 Discharge Information (Q19, Q20)	Patient Experience
H-HSP-RARTING Overall Rating of Hospital (Q21)	Patient Experience
H-QUIET-HSP Quietness of Hospital Environment (Q9)	Patient Experience
H-RECMND Willingness to Recommend Hospital (Q22)	Patient Experience
H-COMP-7 HCAHPS 3 Item Care Transition Measure (CTM-3)	Patient Experience
HF-2 Evaluation of LVS Function	Process – Effectiveness
IMM-2 Influenza Immunization	Process – Effectiveness
IMM-3 Healthcare Personnel Influenza Vaccination	Process – Effectiveness
OP-22 ED-Patient Left Without Being Seen	Process – Effectiveness
OP-23 ED-Head CT or MRI Scan Results for Acute Ischemic Stroke or Hemorrhagic Stroke who Received Head CT or MRI Scan Interpretation Within 45 Minutes of Arrival	Process – Effectiveness
OP-4 Aspirin at Arrival	Process – Effectiveness
OP-6 Timing of Antibiotic Prophylaxis	Process – Effectiveness
OP-7 Prophylactic Antibiotic Selection for Surgical Patients	Process – Effectiveness
PC-01 Elective Delivery Prior to 39 Completed Weeks Gestation: Percentage of Babies Electively Delivered Prior to 39 Completed Weeks Gestation	Process – Effectiveness
PN-6 Initial Antibiotic Selection for Community-Acquired Pneumonia (CAP) in Immunocompetent Patient	Process – Effectiveness
SCIP-Card-2 Surgery Patients on Beta-Blocker Therapy Prior to Arrival Who received a Beta-Blocker During the Perioperative Period	Process – Effectiveness
SCIP-Inf-1 Prophylactic Antibiotic Received Within One Hour Prior to Surgical Incision	Process – Effectiveness
SCIP-Inf-2 Prophylactic Antibiotic Selection for Surgical Patients	Process – Effectiveness
SCIP-Inf-3 Prophylactic Antibiotics Discontinued Within 24 Hours After Surgery End Time	Process – Effectiveness
SCIP-Inf-9 Urinary Catheter Removed on Postoperative Day 1 (POD 1) or Postoperative Day 2 (POD 2) with day of surgery being day zero	Process – Effectiveness
SCIP-VTE-2 Surgery Patients Who Received Appropriate Venous Thromboembolism Prophylaxis Within 24 Hours Prior to Surgery to 24 Hours After Surgery	Process – Effectiveness
STK-1 Venous Thromboembolism (VTE) Prophylaxis	Process – Effectiveness
STK-10 Assessed for Rehabilitation	Process – Effectiveness
STK-2 Discharged on Antithrombotic Therapy	Process – Effectiveness
STK-3 Anticoagulation Therapy for Atrial Fibrillation/Flutter	Process – Effectiveness
STK-4 Thrombolytic Therapy	Process – Effectiveness
STK-5 Antithrombotic Therapy By End of Hospital Day 2	Process – Effectiveness
STK-6 Discharged on Statin Medication	Process – Effectiveness
STK-8 Stroke Education	Process – Effectiveness
VTE-1 Venous Thromboembolism Prophylaxis	Process – Effectiveness

Measure Name	Measure Group
VTE-2 Intensive Care Unit Venous Thromboembolism Prophylaxis	Process – Effectiveness
VTE-3 Venous Thromboembolism Patients with Anticoagulation Overlap Therapy	Process – Effectiveness
VTE-4 Venous Thromboembolism Patients Receiving Unfractionated Heparin with Dosages/Platelet Count Monitoring by Protocol or Nomogram	Process – Effectiveness
VTE-5 Venous Thromboembolism Warfarin Therapy Discharge Instructions	Process – Effectiveness
VTE-6 Hospital Acquired Potentially-Preventable Venous Thromboembolism	Process – Effectiveness
AMI-8a Timing of Receipt of Primary Percutaneous Coronary Intervention (PCI)	Process - Timeliness
ED-1b Median Time from ED Arrival to ED Departure for Admitted ED Patients	Process - Timeliness
ED-2b Admit Decision Time to ED Departure Time for Admitted Patients	Process - Timeliness
OP-18b/ED-3 Median Time from ED Arrival to ED Departure for Discharged ED Patients	Process - Timeliness
OP-20 Door to Diagnostic Evaluation by a Qualified Medical Professional	Process - Timeliness
OP-21 ED-Median Time to Pain Management for Long Bone Fracture	Process - Timeliness
OP-3 Median Time to Transfer to Another Facility for Acute Coronary Intervention	Process - Timeliness
OP-5 Median Time to ECG	Process - Timeliness
OP-8 MRI Lumbar Spine for Low Back Pain	Efficiency – Imaging
OP-10 Abdomen CT Use of Contrast Material	Efficiency – Imaging
OP-11 Thorax CT Use of Contrast Material	Efficiency – Imaging
OP-13 Cardiac Imaging for Preoperative Risk Assessment for Non-Cardiac Low-Risk Surgery	Efficiency – Imaging
OP-14 Simultaneous Use of Brain Computed Tomography (CT) and Sinus CT	Efficiency – Imaging

Table B.2. Measures Exclused from April 2015 Star Ratings (N=31)

Measure Name	Exclusion Criteria
OP-15 Use of Brain CT in the Emergency Department (ED) for Atraumatic Headache	Retired or suspended from public reporting (N=11)
SCIP-VTE-1 Surgery Patients with Recommended Venous Thromboembolism Prophylaxis Ordered	
AMI-2 Aspirin Prescribed at Discharge	
AMI-10 Statin Prescribed at Discharge	
SCIP-Inf-4 Cardiac Surgery Patients with Controlled Postoperative Blood Glucose	
SCIP-Inf-10 Surgery Patients with Perioperative Temperature Management	
HF- 1 Discharge Instructions	
HF-3 ACEI or ARB for LVSD	
PN-3b Blood Cultures Performed in the Emergency Department Prior to Initial Antibiotic Received in Hospital	
SM-PART-STROKE Participation in a Systematic Clinical Database Registry for Stroke Care	
IMM-1a Pneumococcal Immunization – Overall Rate	
OP-29 Endoscopy/Polyp Surveillance: Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients	Not yet publicly reported (N=3)
OP-30 Endoscopy/Polyp Surveillance: Colonoscopy Interval for Patients with a History of Adenomatous Polyps – Avoidance of Inappropriate Use	
OP-31 Cataracts – Improvement in Patient's Visual Function Within 90 Days Following Cataract Surgery	
CAC-1 Relievers for Inpatient Asthma	Too few hospitals reporting (N=6)
CAC-2 Systemic Corticosteroids for Inpatient Asthma	
CAC-3 Home Management Plan of Care (HMPC) Document Given to Patient/Caregiver	
AMI-7a Fibrinolytic Therapy Received Within 30 Minutes of Hospital Arrival	
OP-1 Median Time to Fibrinolysis	
OP-2 Fibrinolytic Therapy Received Within 30 Minutes of ED Arrival	
SM-PART-NURSE Participation in a Systematic Clinical Database Registry for Nursing Sensitive Care	Structural Measures (N=8)
SM-PART-CARD Participation in a Systematic Clinical Database Registry for Cardiac Surgery	
ACS-REGISTRY Participation in a Multispecialty Surgical Registry	
SM-PART-GEN-SURG Participation in a Systematic Clinical Database Registry for General Surgery	
OP-25 Safe Surgery Checklist Use	
OP-12 The Ability for Providers with HIT to Receive Laboratory Data Electronically Directly into their ONC-Certified EHR System as Discrete Searchable Data	
OP-17 Tracking Clinical Results between Visits	
OP-26 Hospital Outpatient Volume Data on Selected Outpatient Surgical Procedures	
MSPB-1/SPP-1 Medicare Spending per Beneficiary (MSPB)	Non-Directional Efficiency (N=3)
OP-9 Mammography Follow-up Rates	
PAYM-30-AMI Acute Myocardial Infarction (AMI) Payment per Episode of Care	

Appendix C: Latent Traits between Star Categories by Measure Group

Figure C. 1. Distribution of Hospital Mortality Group Score by Dry Run Star Rating

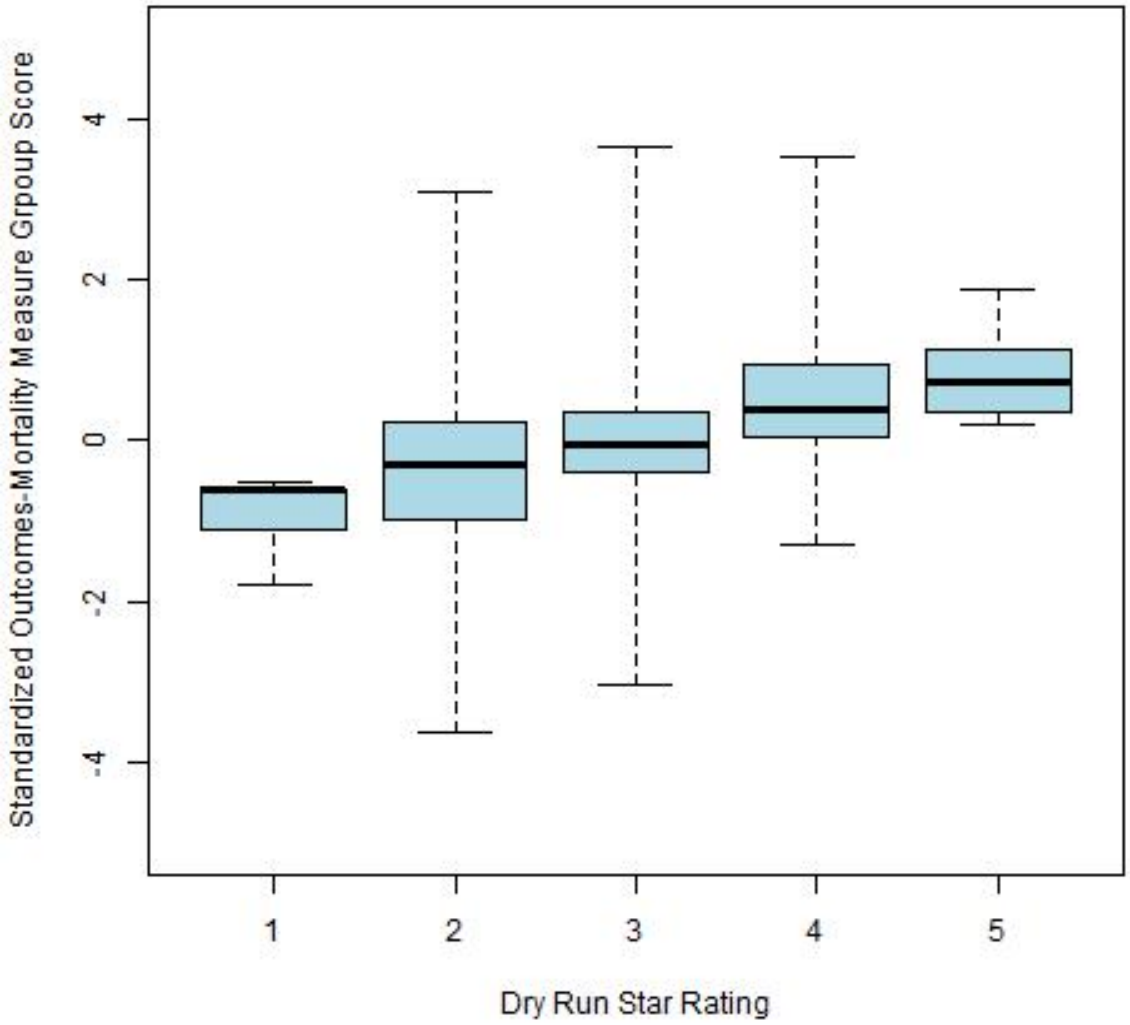


Figure C. 2. Distribution of Hospital Safety Group Score by Dry Run Star Rating

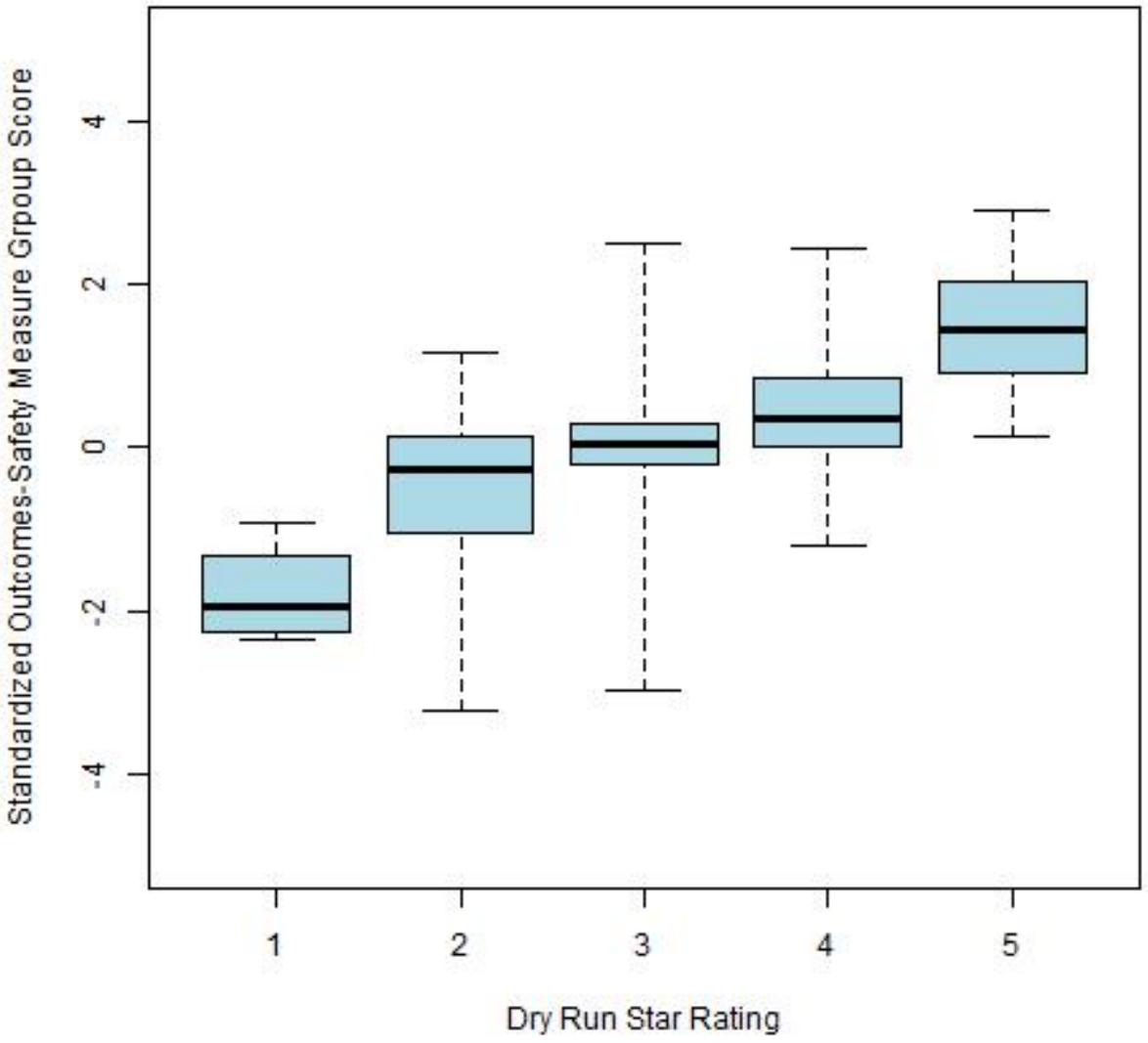


Figure C. 3. Distribution of Hospital Readmission Group Score by Dry Run Star Rating

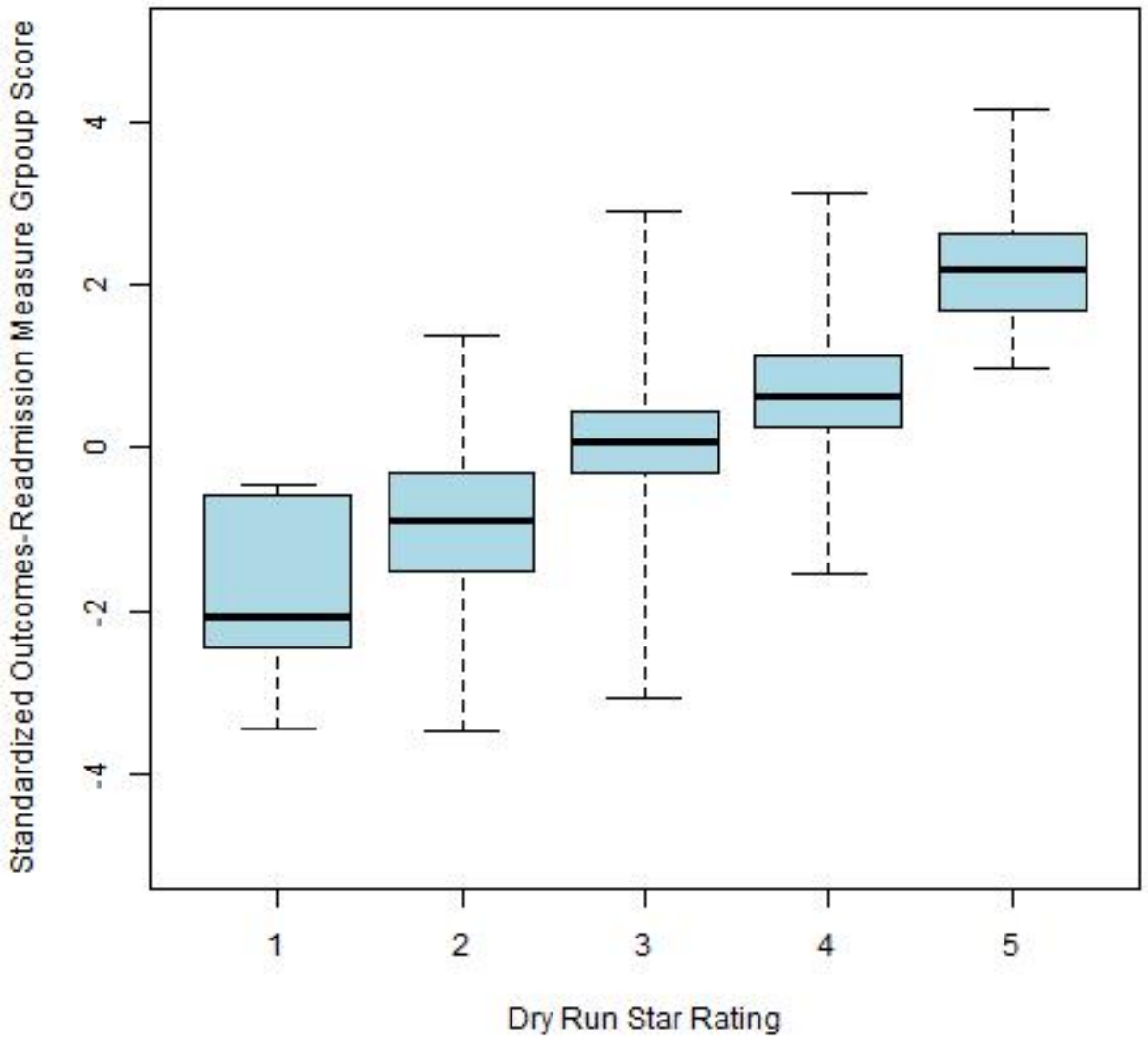


Figure C. 4. Distribution of Patient Experience Group Scores by Dry Run Star Rating

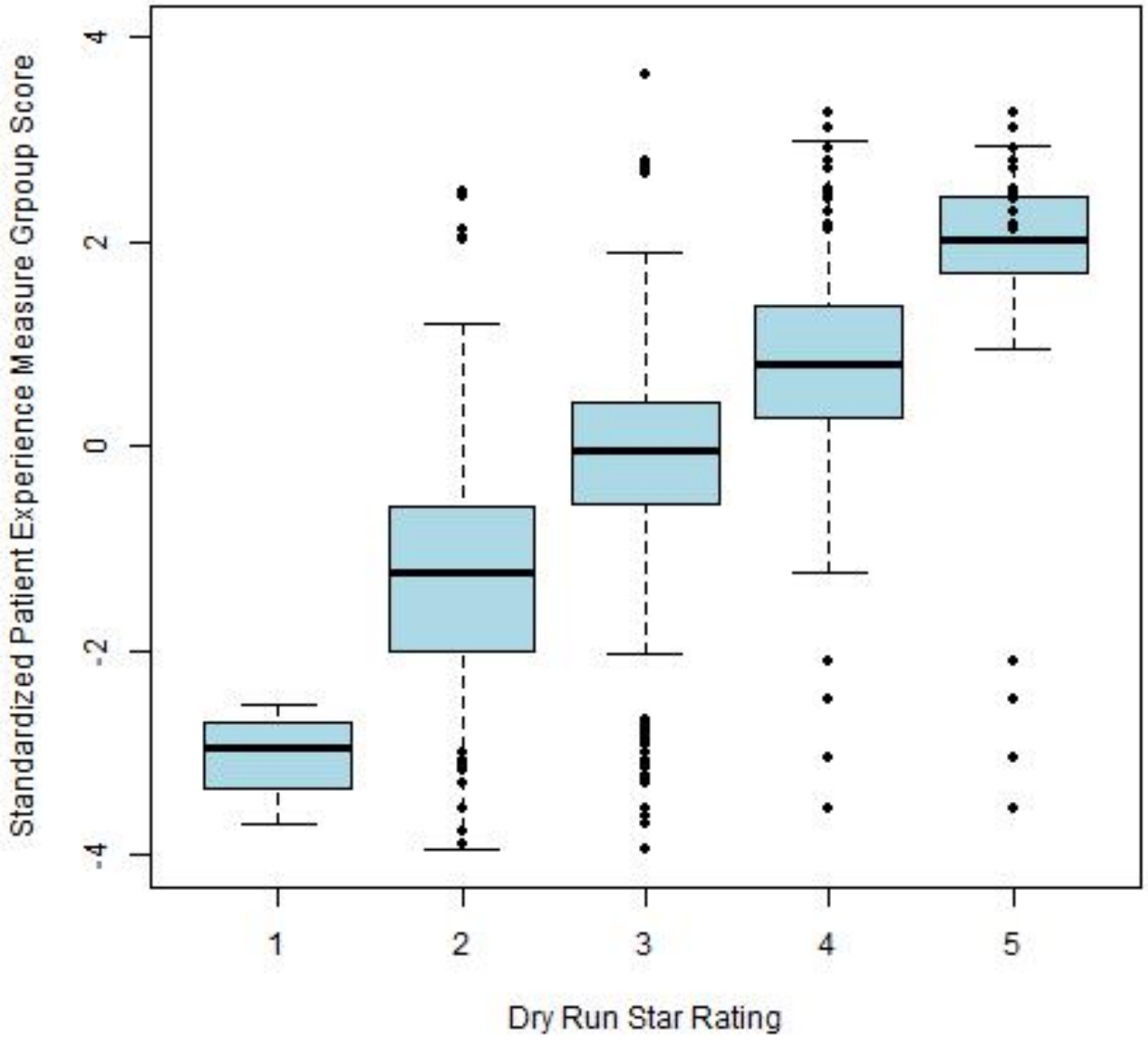


Figure C. 5. Distribution of Hospital Efficiency-Imaging Group Score by Dry Run Star Rating

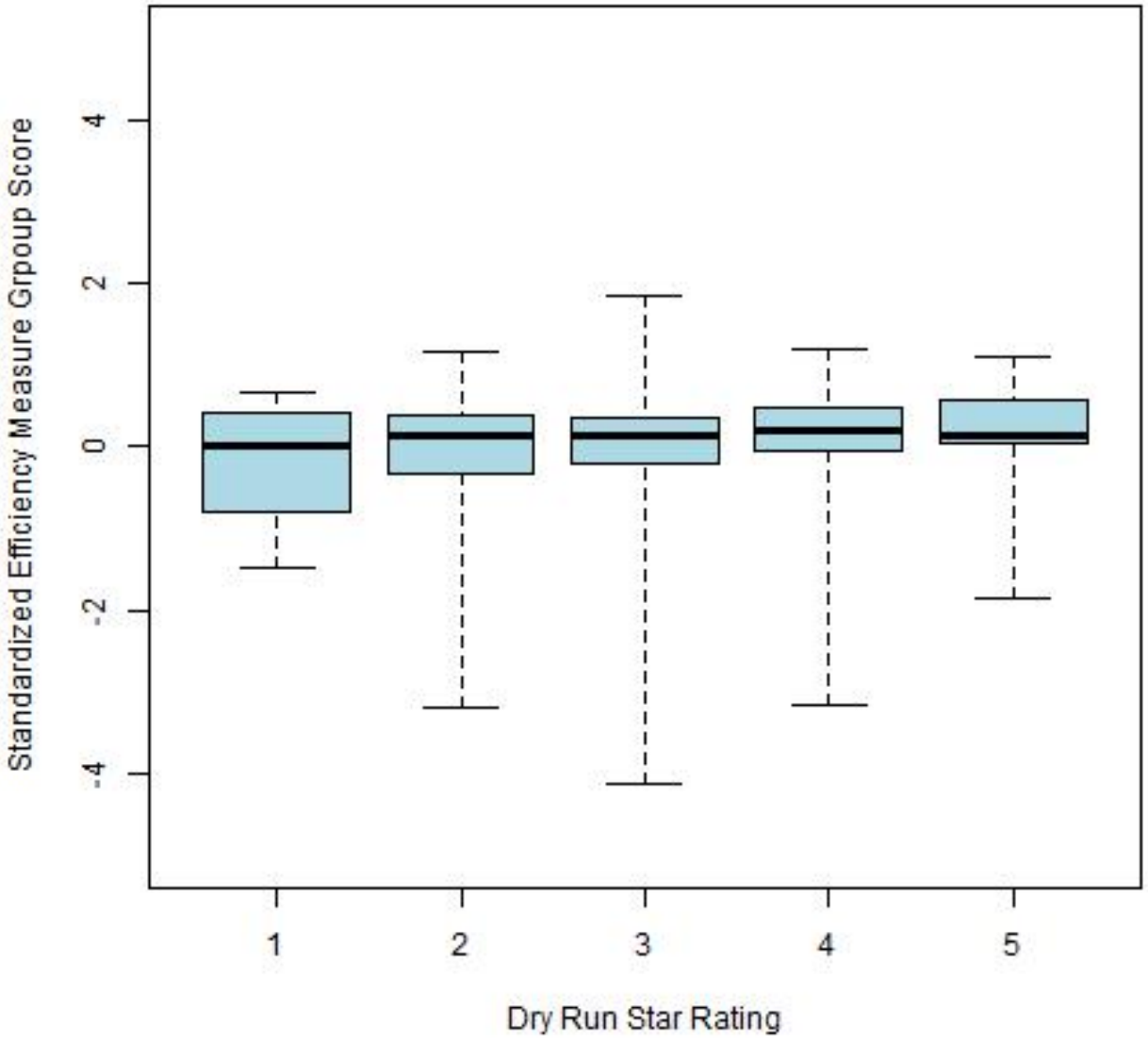


Figure C. 6. Distribution of Hospital Process-Timeliness Group Score by Dry Run Star Rating

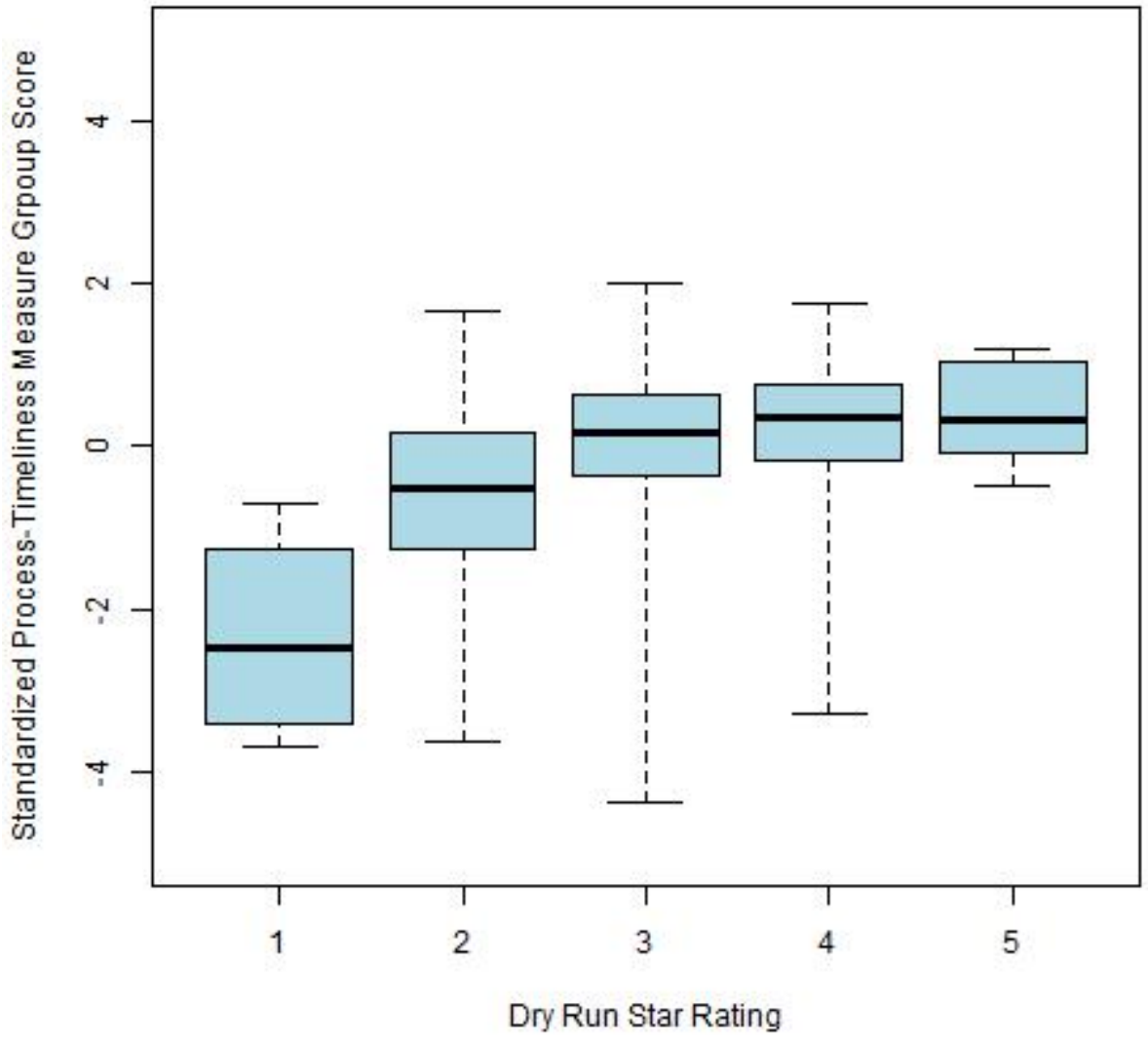
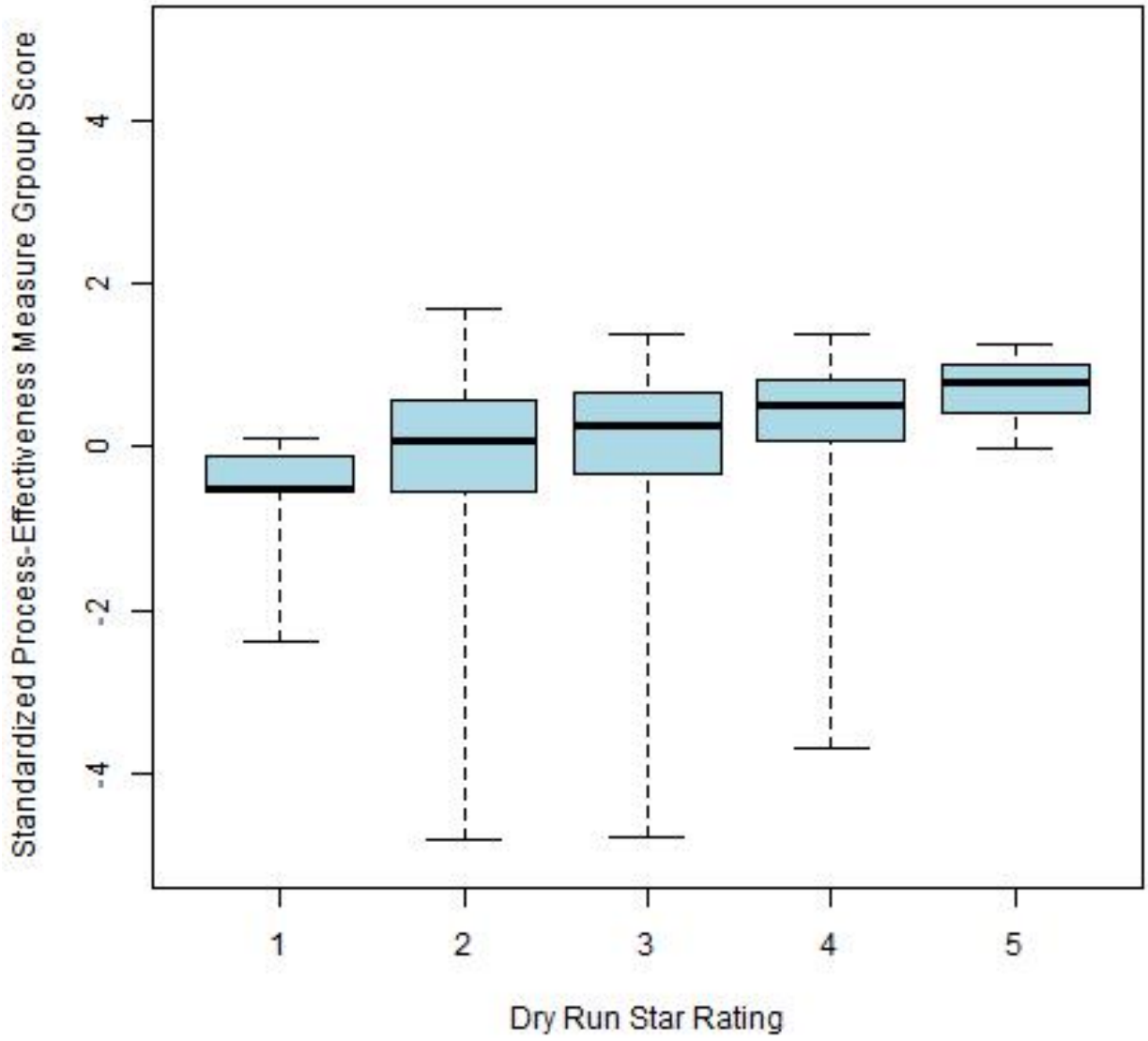


Figure C. 7. Distribution of Hospital Process-Effectiveness Group Score by Dry Run Star Rating



Appendix D: Measure Loadings

Figure D.1. Loadings by Measure in Outcomes – Readmission Group (April 2015 data)

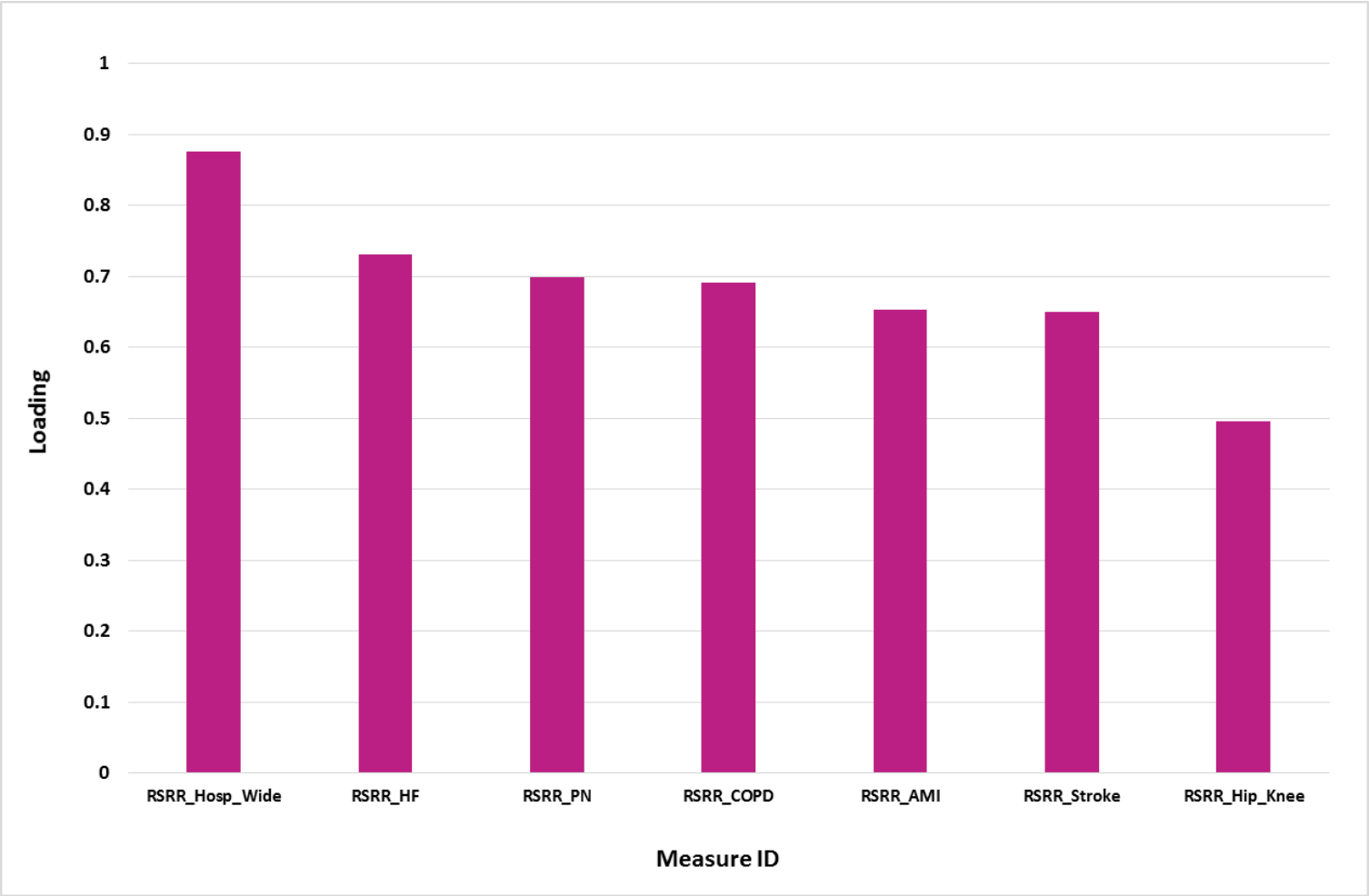


Figure D.2. Loadings by Measure in Outcomes – Safety Group (April 2015 data)

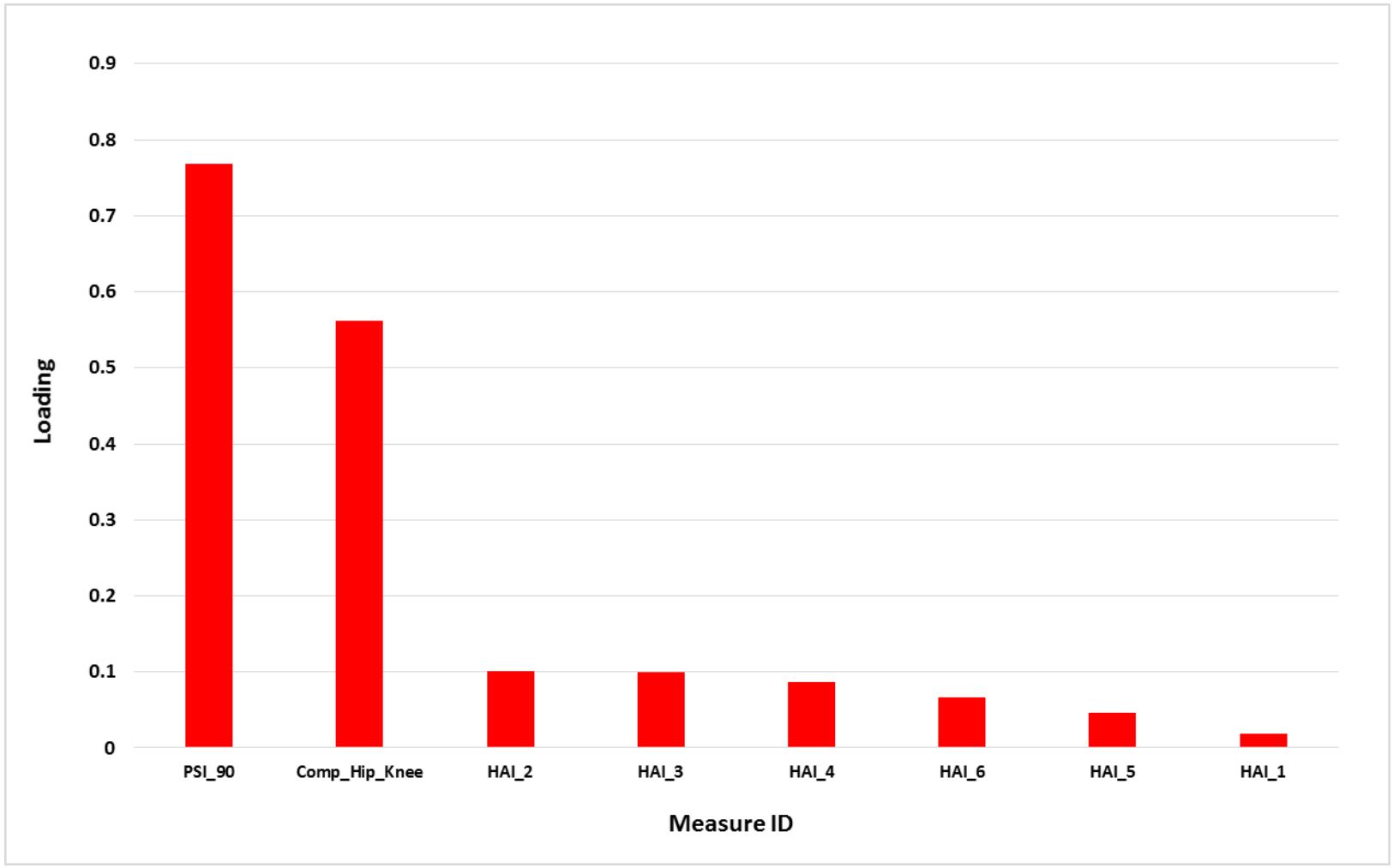


Figure D.3. Loadings by Measure in Outcomes – Mortality Group (April 2015 data)

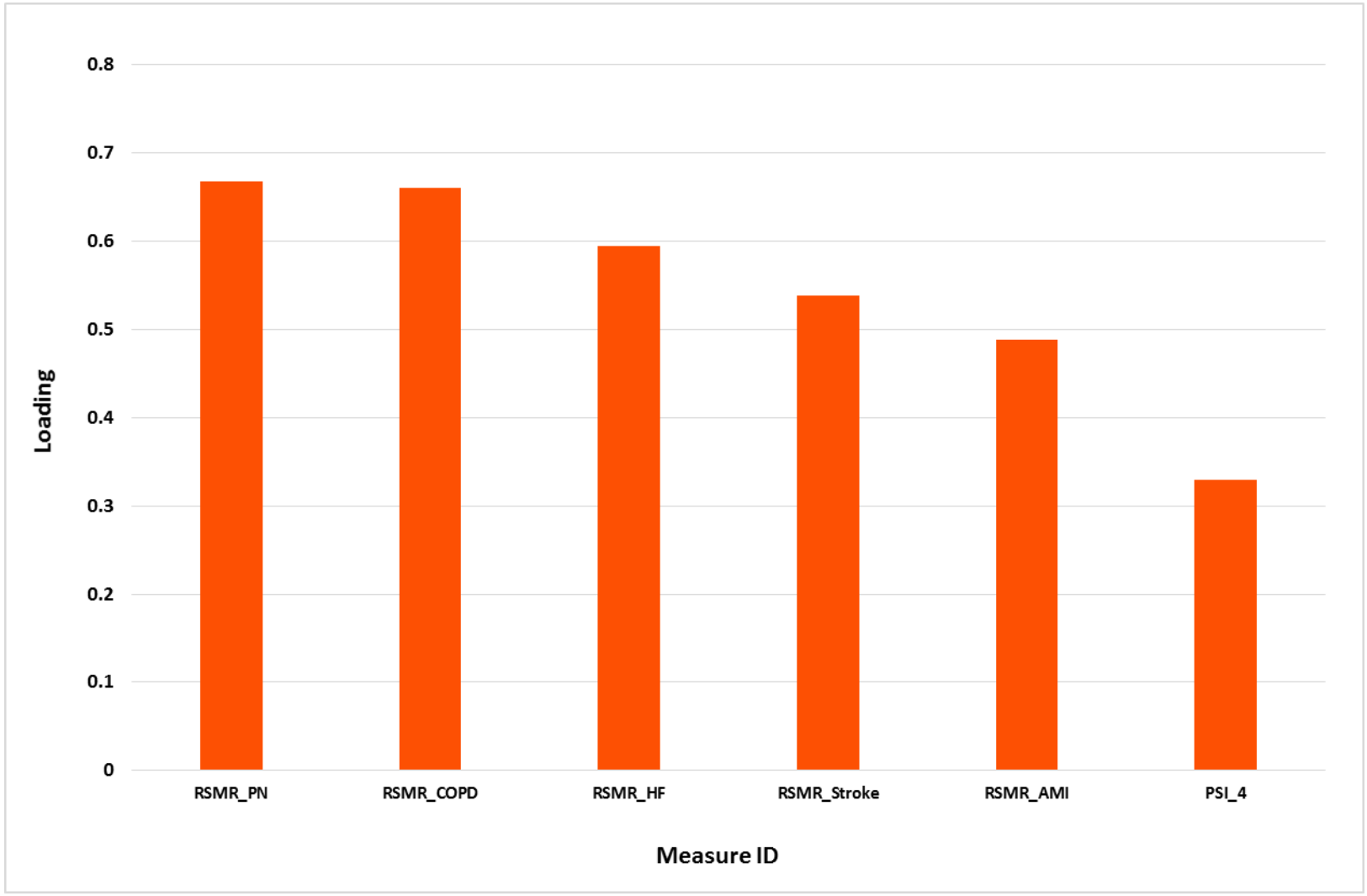


Figure D.4. Loadings by Measure in Patient Experience Group (April 2015 data)

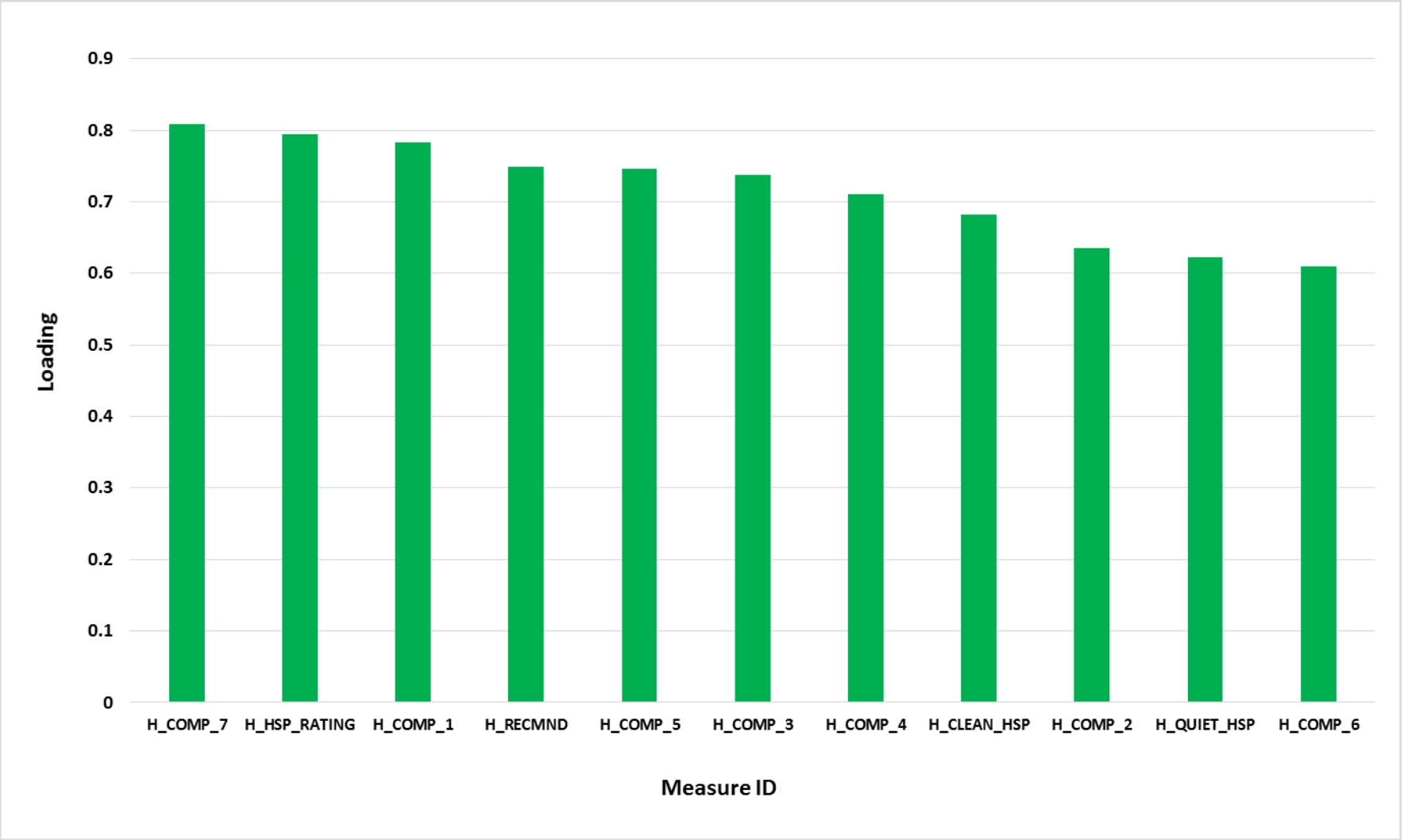


Figure D.5. Loadings by Measure in Process – Effectiveness Group (April 2015 data)

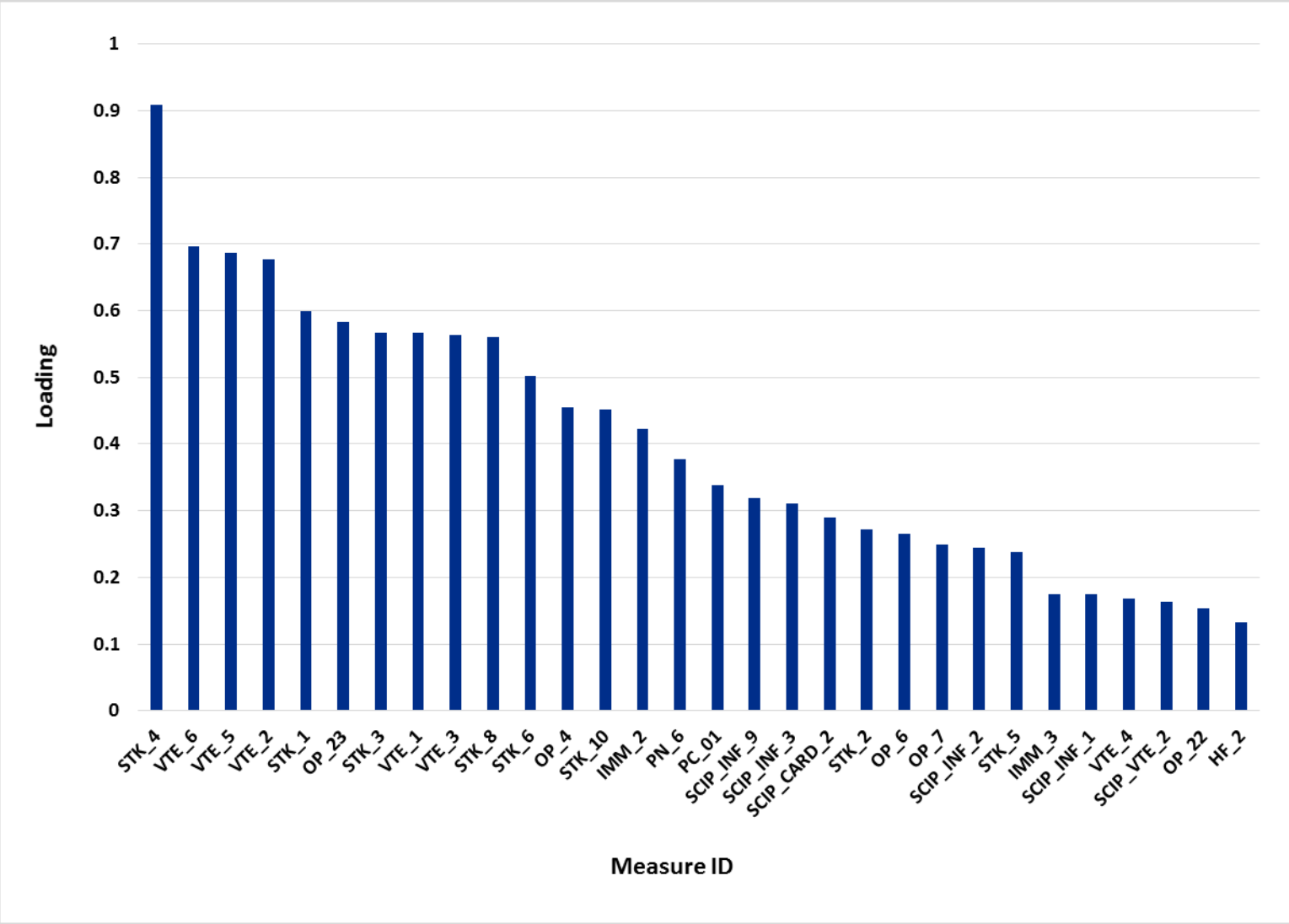


Figure D.6. Loadings by Measure in Process – Timeliness Group (April 2015 data)

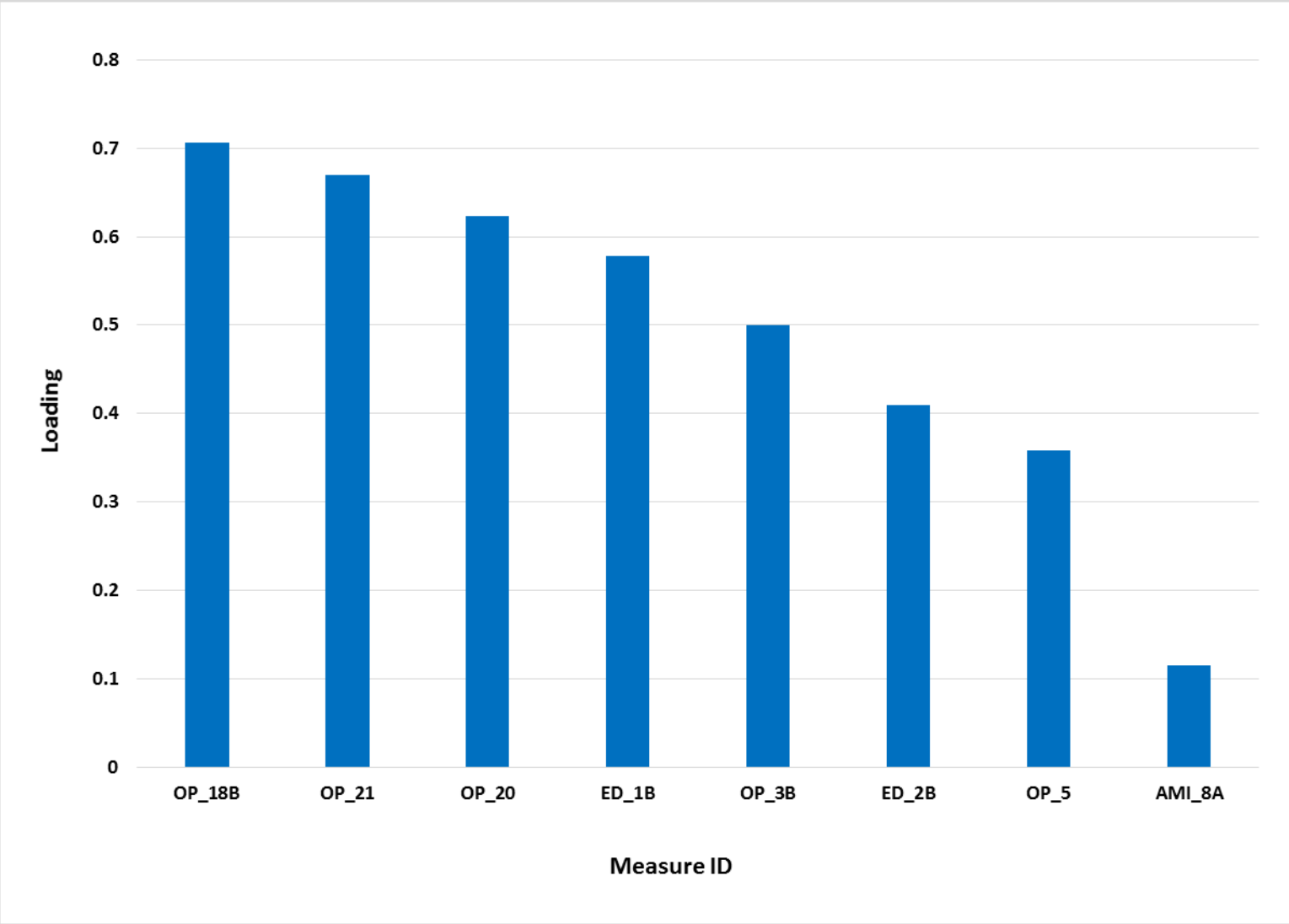
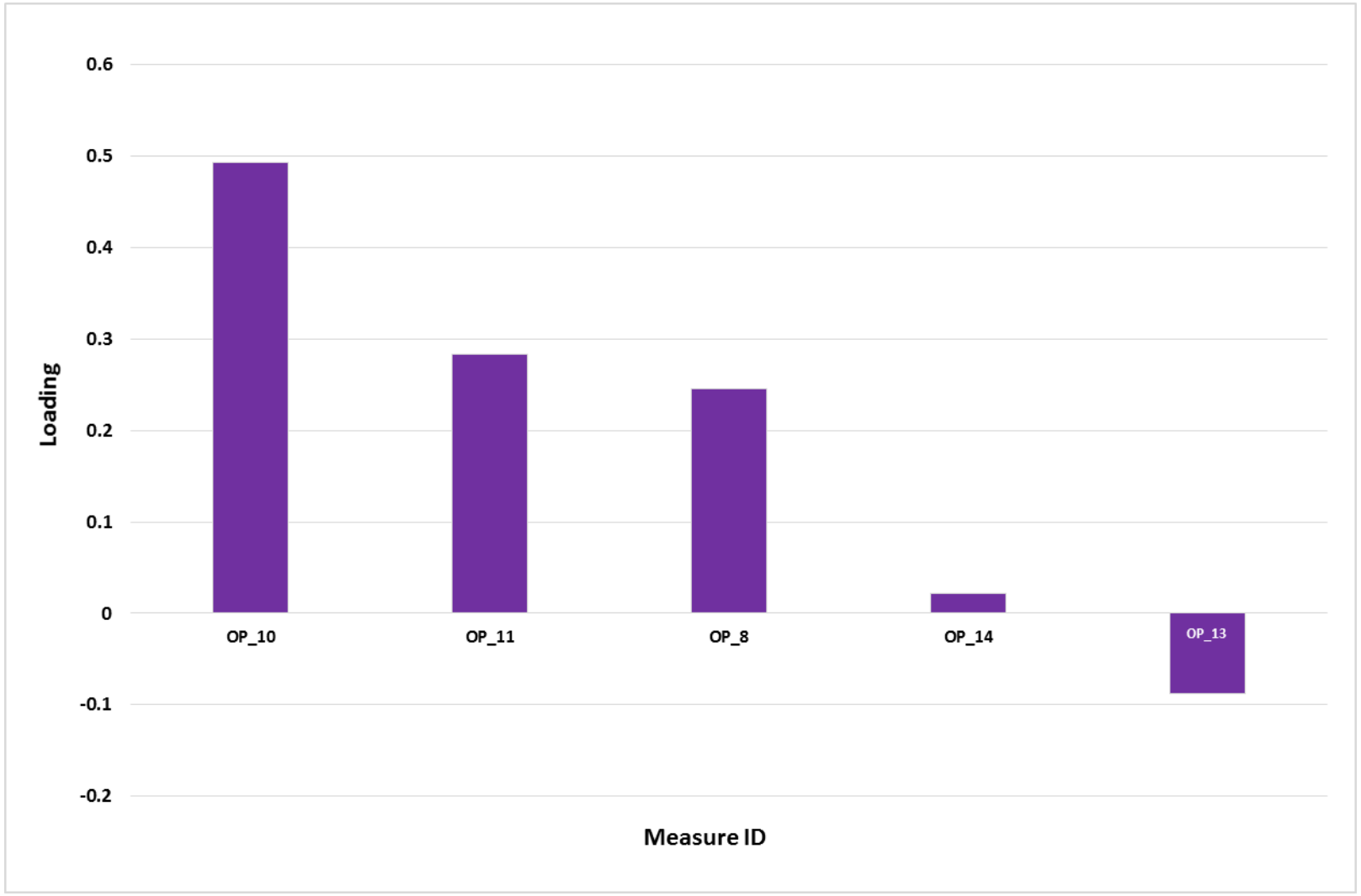


Figure D.7. Loadings by Measure in Efficiency – Imaging Group (April 2015 data)



Appendix E: Hospital Distribution by Reported Measures (April 2015)

Figure E.1. Distribution of Hospitals for All Reported Measures

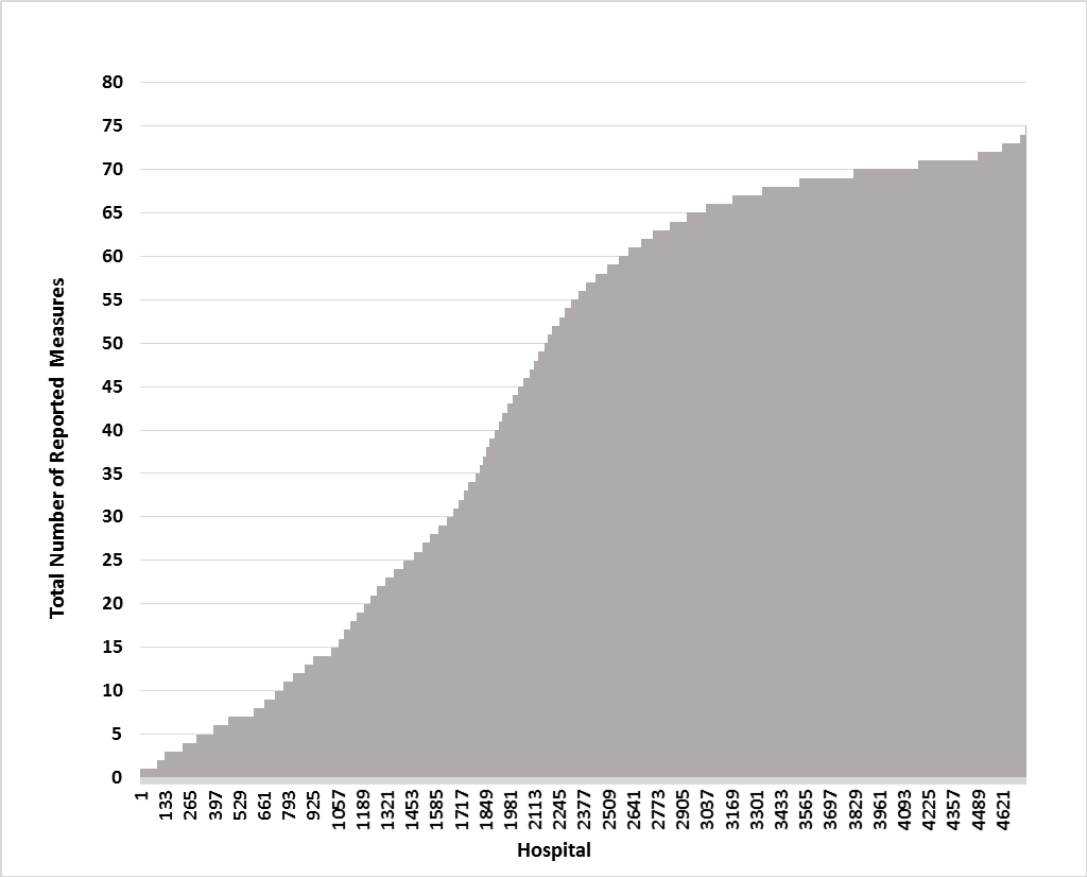


Figure E.2. Distribution of Hospitals for the Outcome-Mortality Measure Group

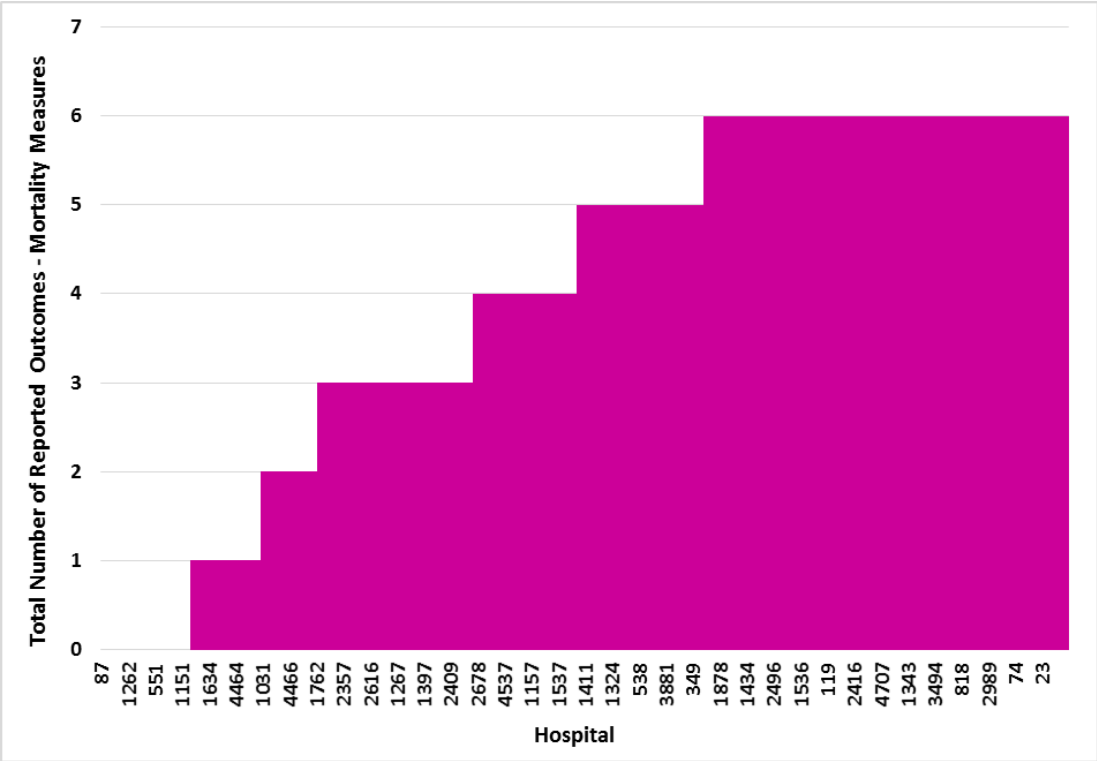


Figure E.3. Distribution of Hospitals for the Outcome-Safety Measure Group

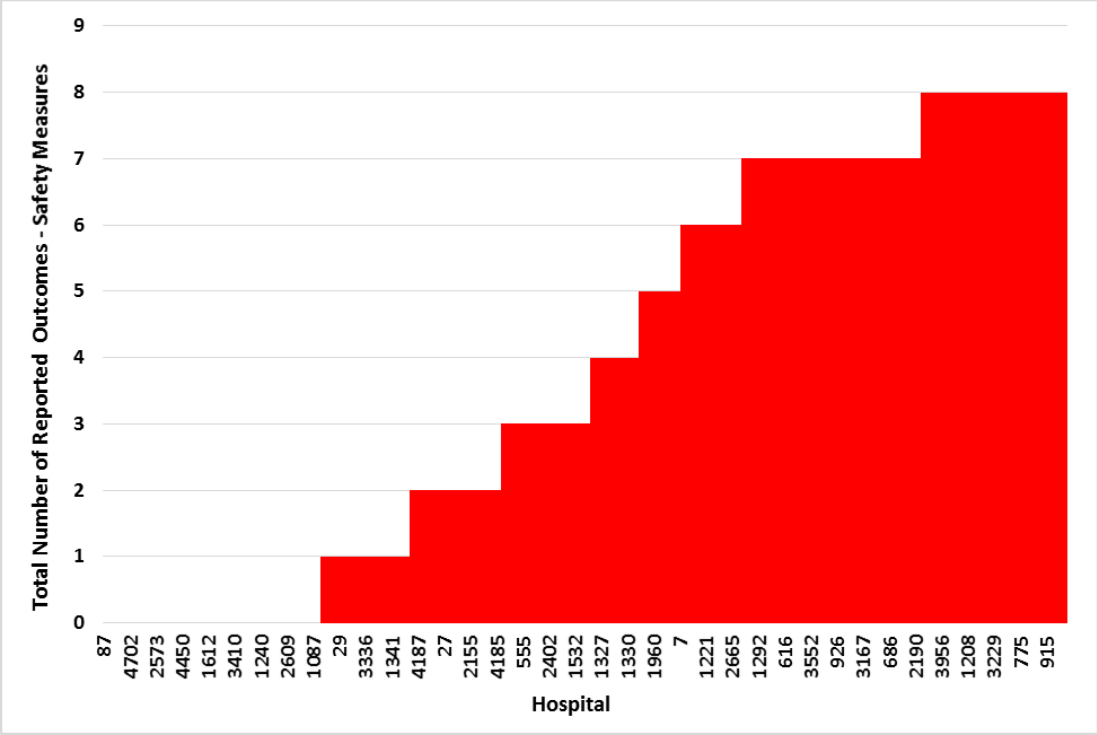


Figure E.4. Distribution of Hospitals for the Outcome-Readmission Measure Group

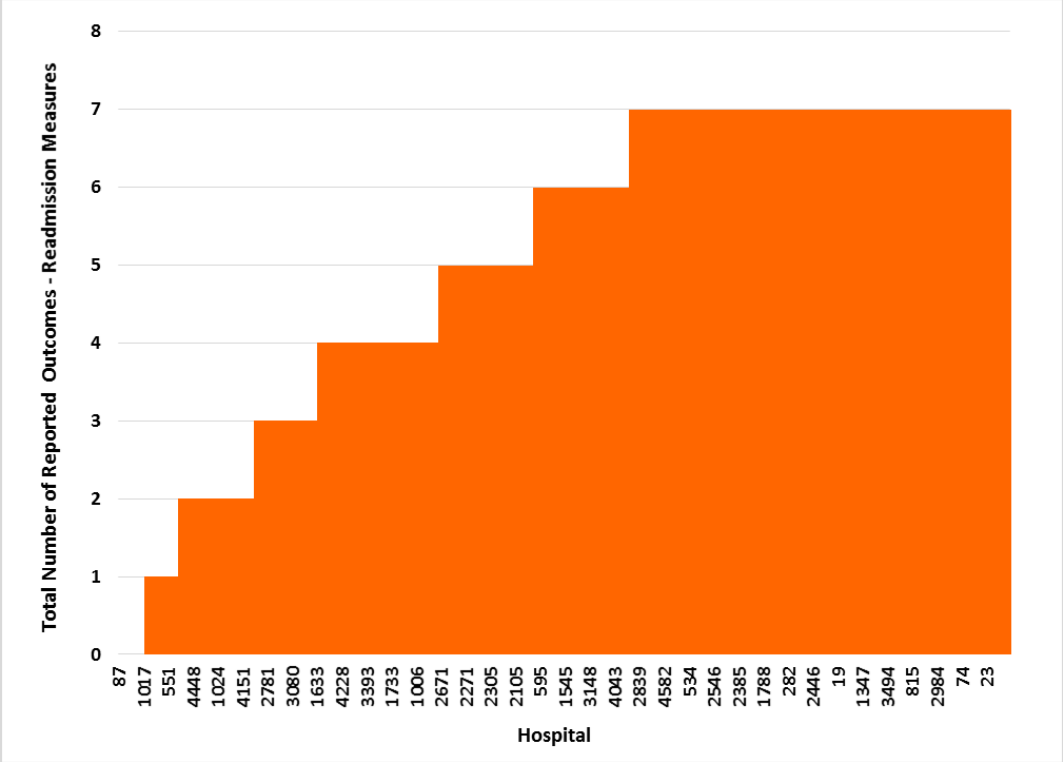


Figure E.5. Distribution of Hospitals for the Patient Experience Measure Group

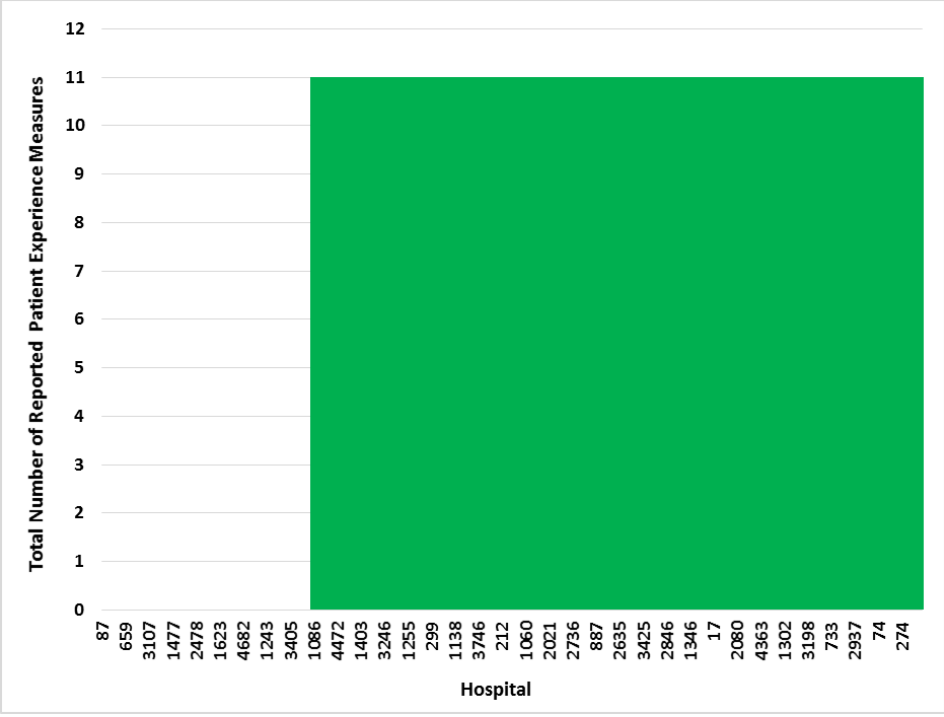


Figure E.6. Distribution of Hospitals for the Process-Effectiveness Measure Group

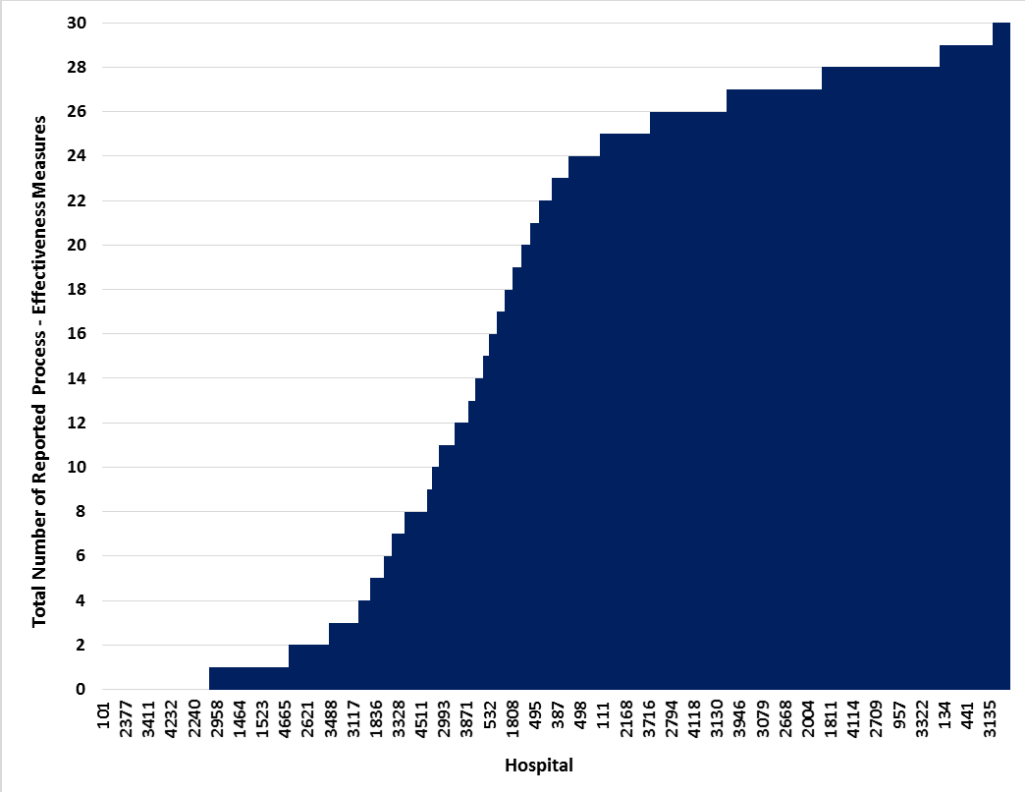


Figure E.7. Distribution of Hospitals for the Process-Timeliness Measure Group

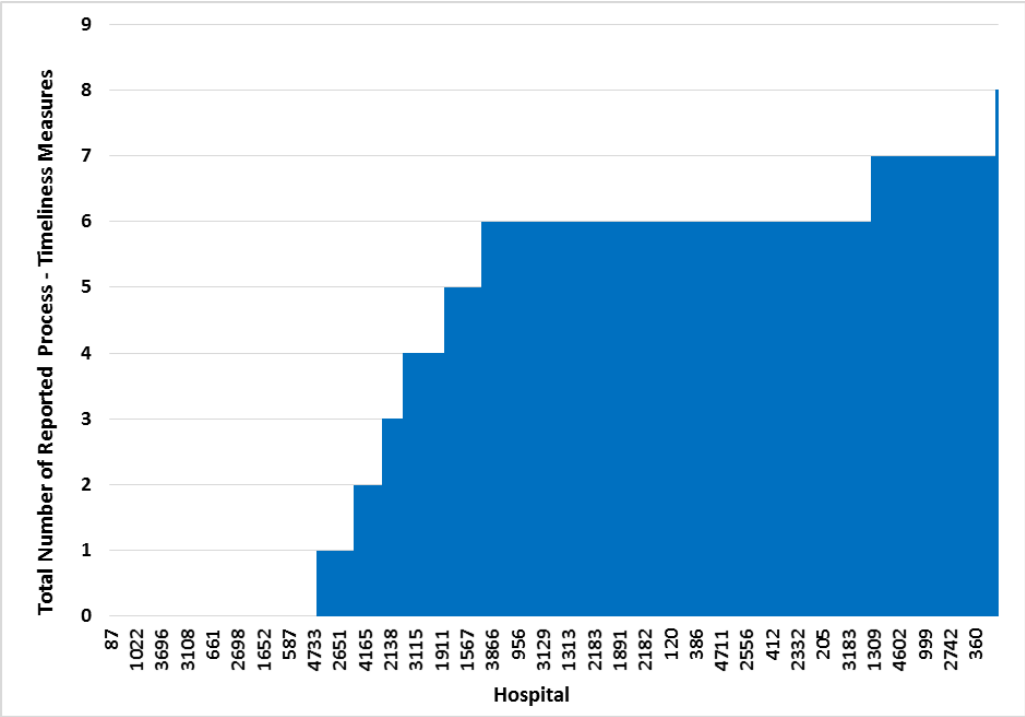


Figure E.8. Distribution of Hospitals for the Efficiency-Imaging Measure Group

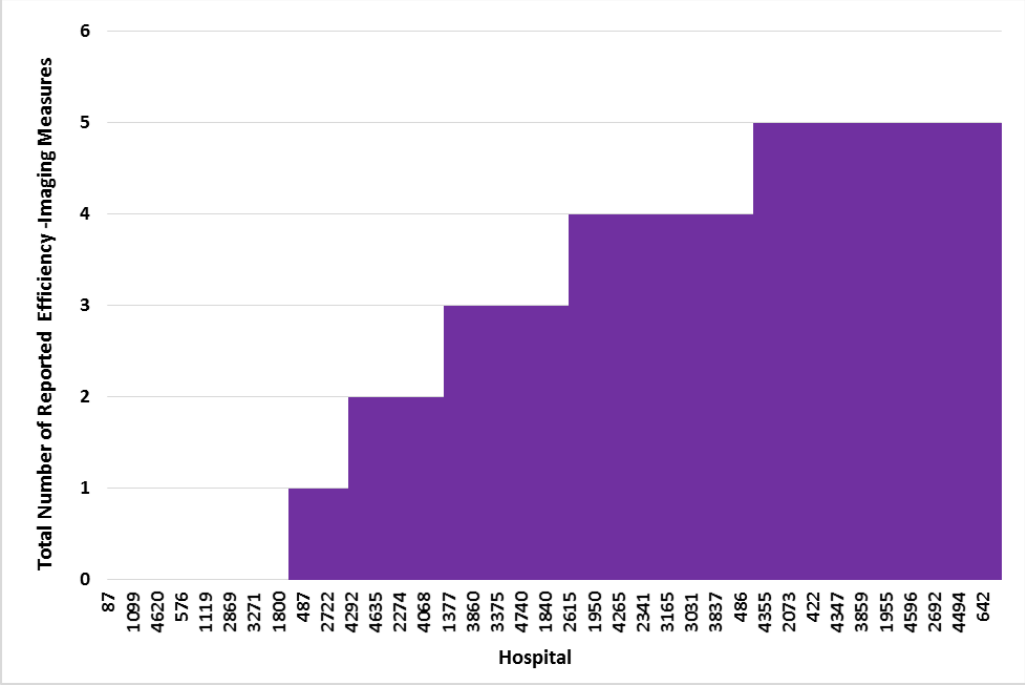


Figure E.9. Count of Hospitals By Number of Outcomes - Mortality Measures Reported, n hospitals (% total hospitals)

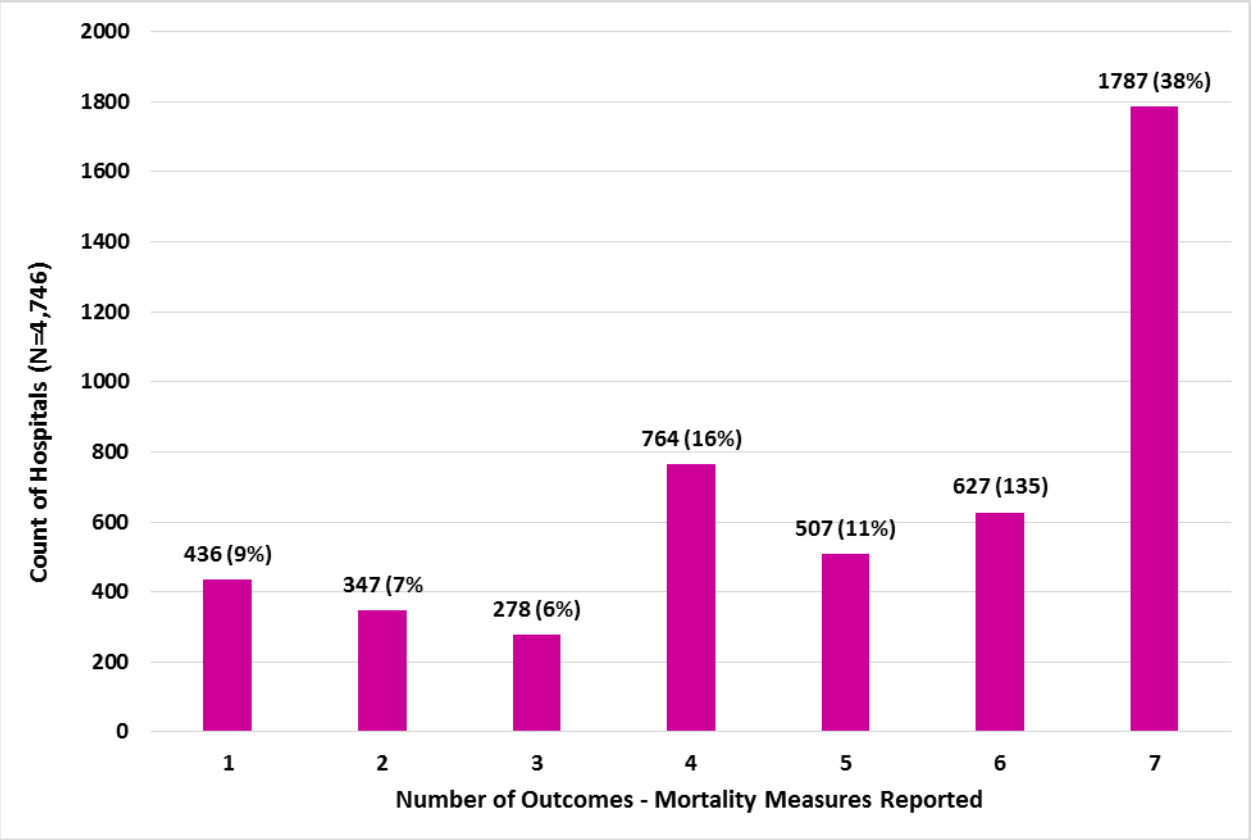


Figure E.10. Count of Hospitals By Number of Outcomes - Safety Measures Reported, n hospitals (% total hospitals)

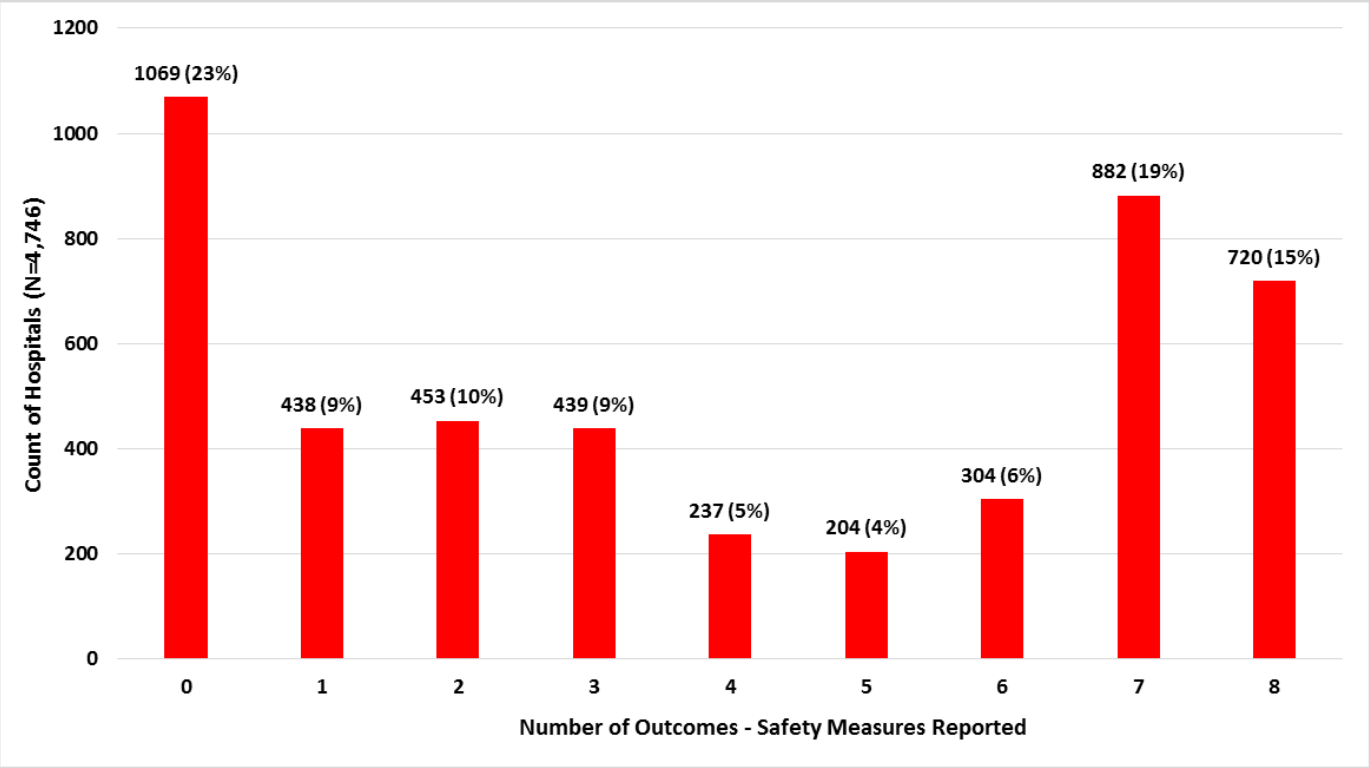


Figure E.11. Count of Hospitals By Number of Outcomes - Readmission Measures Reported, n hospitals (% total hospitals)

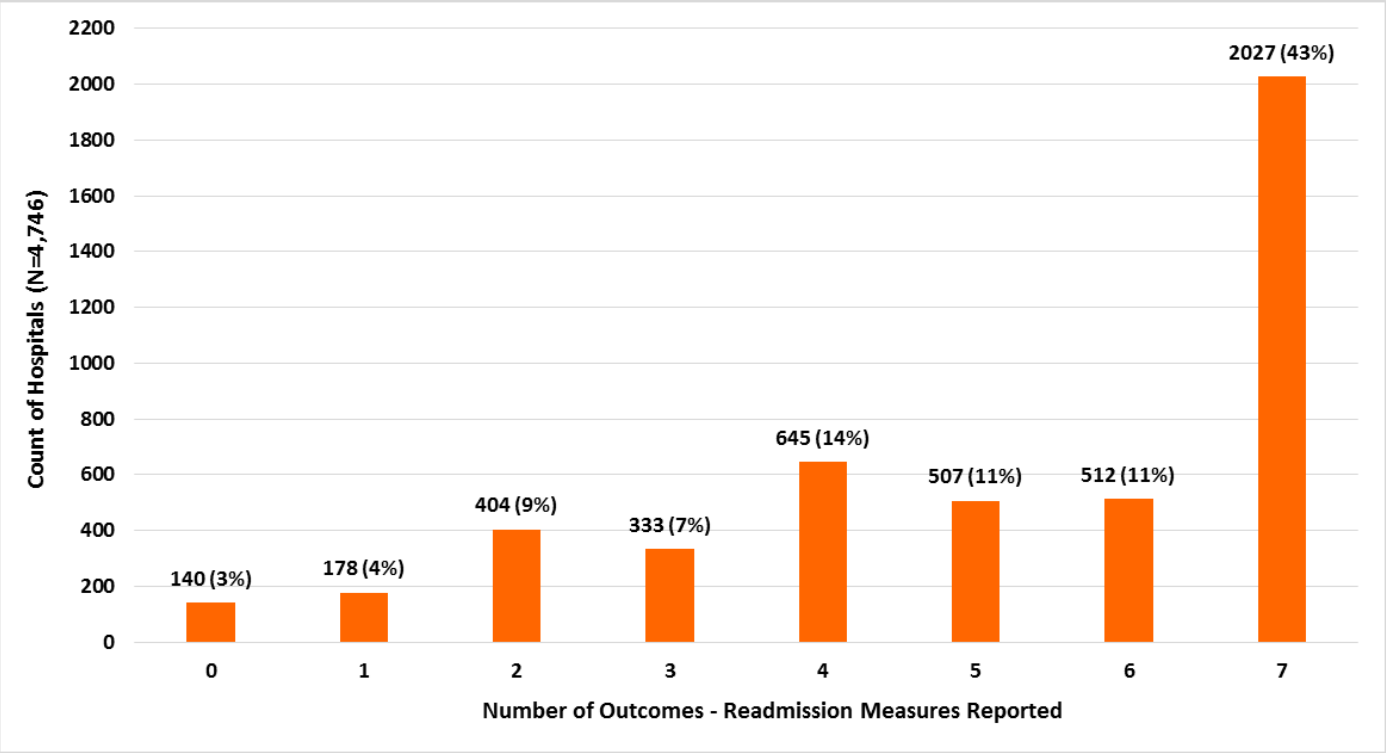


Figure E.12. Count of Hospitals By Number of Patient Experience Measures Reported, n hospitals (% total hospitals)

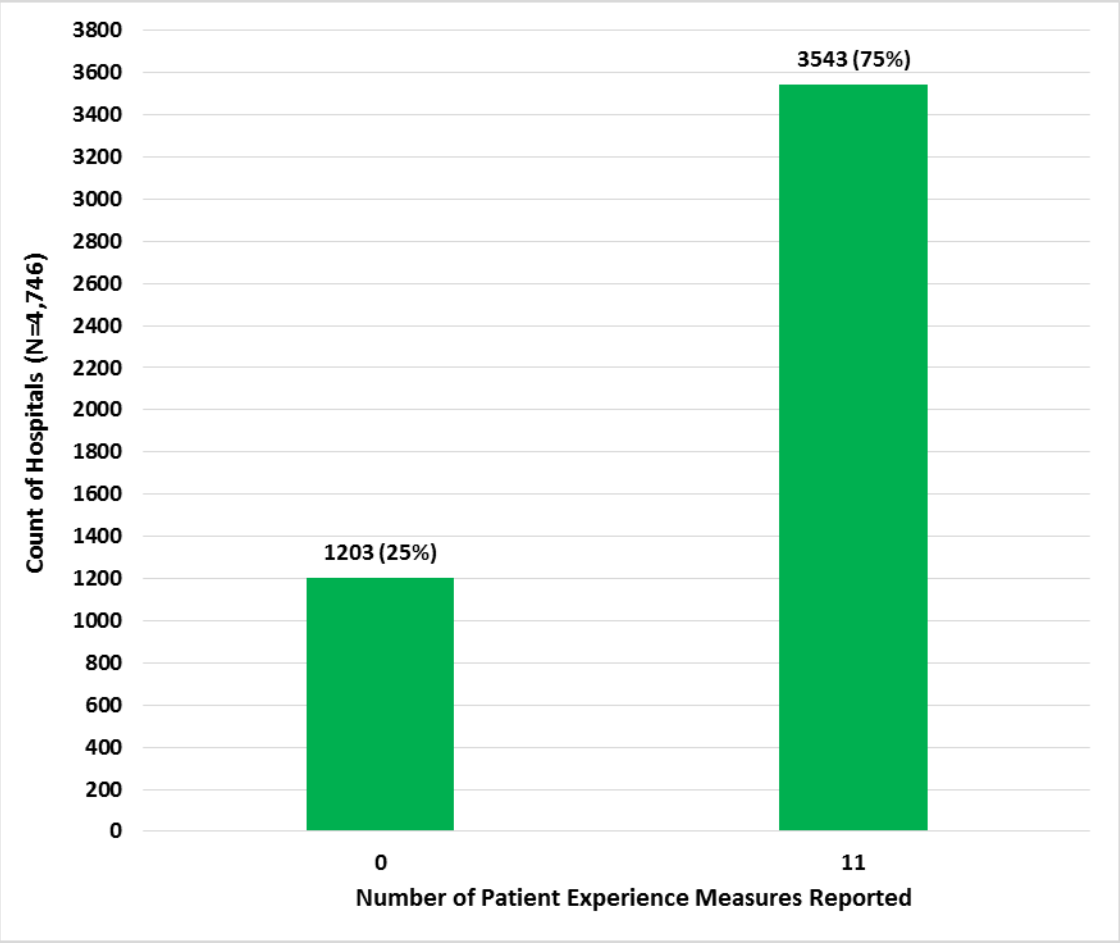


Figure E.13. Count of Hospitals By Number of Process - Effectiveness Measures Reported, n hospitals (% total hospitals)

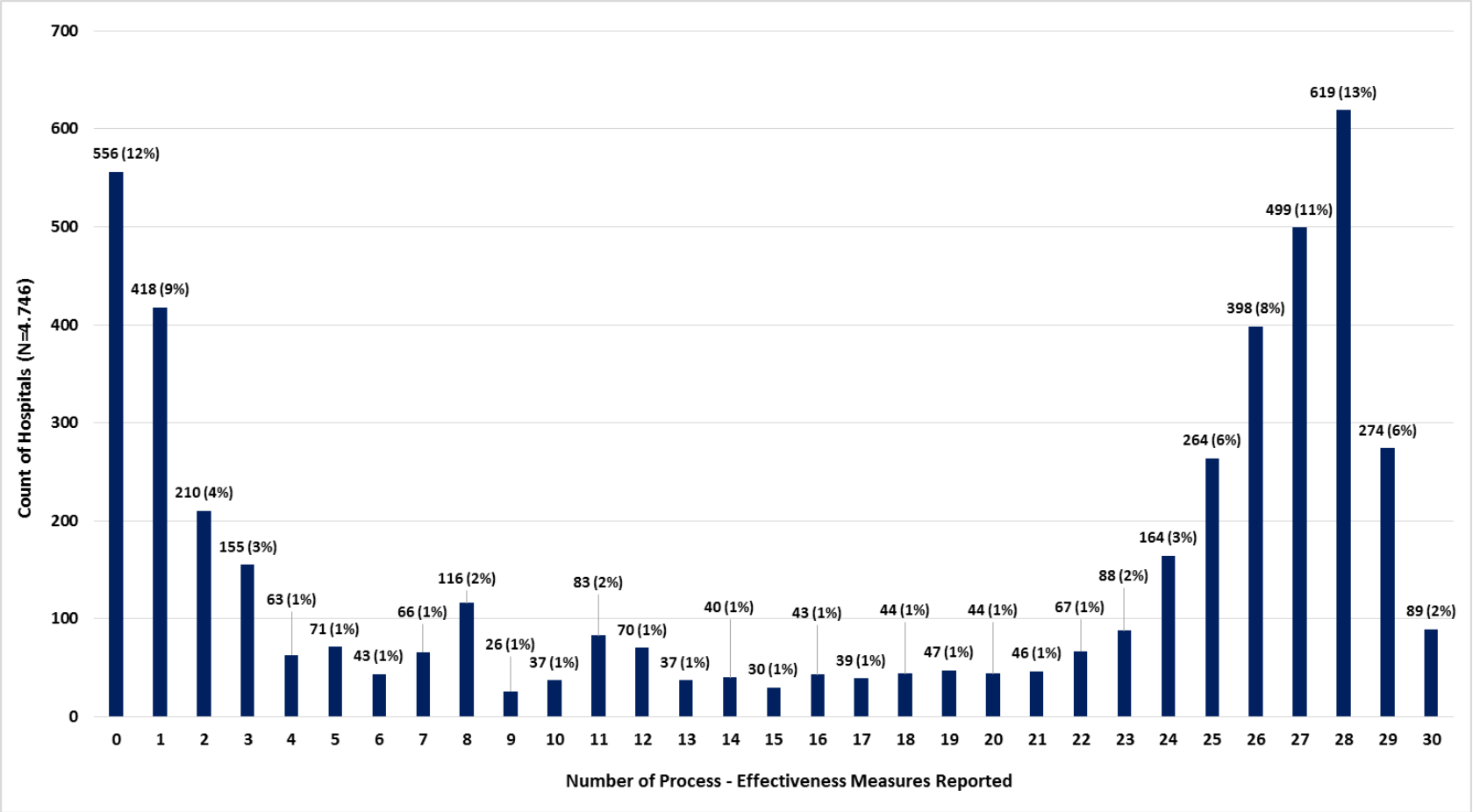


Figure E.14. Count of Hospitals By Number of Process - Timeliness Measures Reported, n hospitals (% total hospitals)

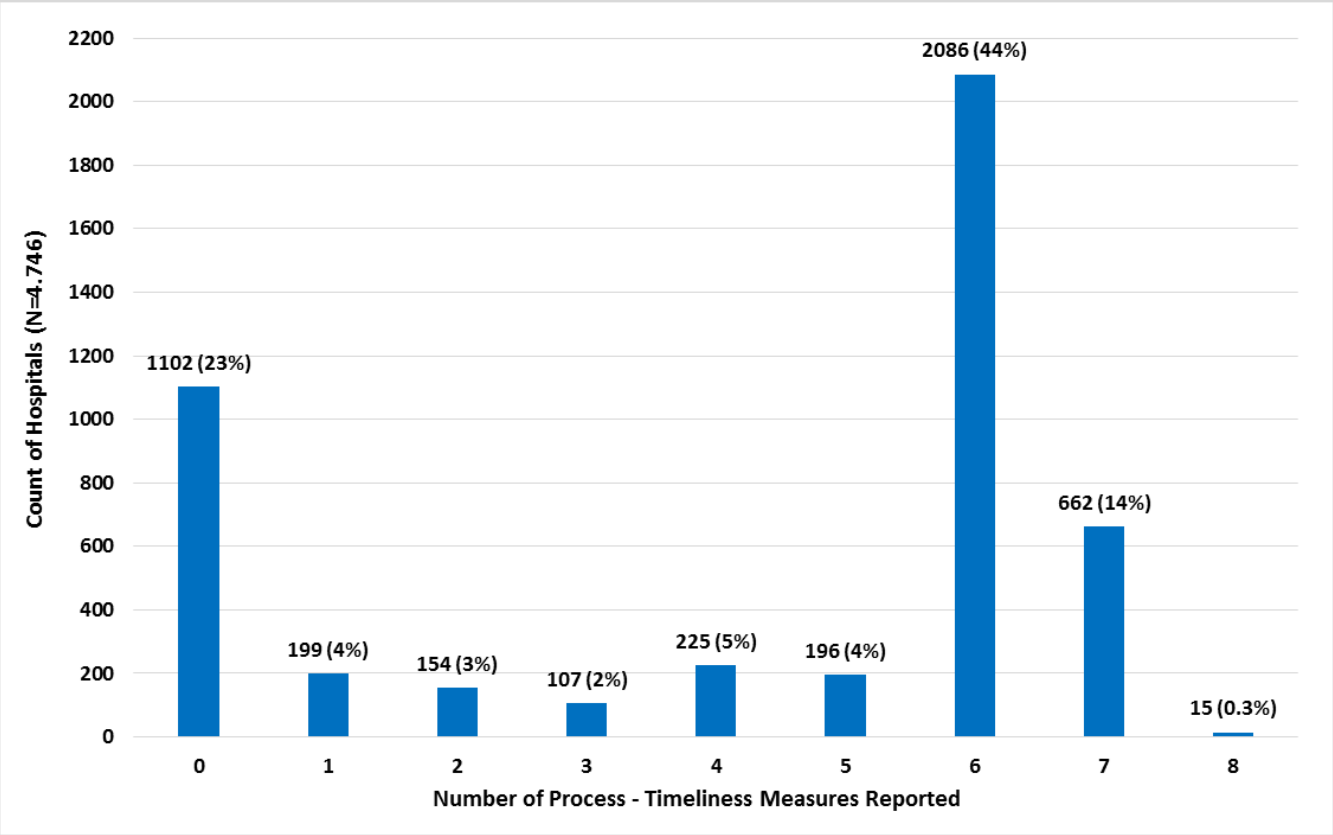


Figure E.15. Count of Hospitals By Number of Efficiency - Imaging Measures Reported, n hospitals (% total hospitals)