# An Analysis of the Medicare Hospital 5-Star Rating and a comparison with Quality Penalties

11 December 2016

J. Graham Atkinson, D.Phil.
Executive Vice President for Research and Policy
Jayne Koskinas Ted Giovanis Foundation for Health and Policy

## Executive Summary

Medicare is now publishing star ratings of hospitals with the intent to provide the public with an easy way to compare the quality of inpatient care being provided by hospitals. The paper provides a brief description of the methodology used to construct this rating. Some concerns regarding the design and biases included therein are discussed, and data to support the concerns is presented. The concerns fall into two categories: 1) the biases that are evident in the results of the rating system; and 2) conceptual problems in the design of the method used to combine individual quality scores.

It is demonstrated that biases against larger hospitals and against hospitals with a higher level of disproportionate share (DSH) patients are present in the overall quality reward/penalty system as well as the 5-star rating system, and that these biases are highly statistically significant.

An important feature of the 5-star rating methodology is the use of latent variable models to construct seven intermediate scores for categories of quality measures. It is argued on conceptual grounds that the use of such a model is inappropriate, and that the results published by CMS demonstrate some of the deficiencies of these models. The latent variable models are based on an invalid assumption, i.e., that the various observed quality measures within a category are projections of a single underlying (and unobserved) variable. In addition, they provide an excessive weight to some of the initial quality measures and virtually zero weight to others.

It is of interest to note that there are some 5-star hospitals that are hit with penalties for low quality, while there are some 1-star hospitals that receive aggregate rewards for quality. This is symptomatic of the fact that different messages are being sent by the different quality programs.

# Background

Medicare is now publishing star ratings of hospitals that are intended to provide the public with an easy way to compare the quality of inpatient care being provided by hospitals. A brief description of the methodology used to construct this rating is provided in this section and a more complete description can be found on the qualitynet website[1]. Some concerns regarding the design and biases included therein are outlined and these concerns are expanded in later sections of this paper. The concerns will be presented in two forms: 1) the biases that are evident in the results of the rating system; and 2) conceptual problems in the design of the method used to combine individual quality scores.

## The 5-star rating methodology

The following description is a simplified version of what is actually done. Complications involving the method used to select which quality measures would be included, how to standardize the measures and deal with trimming of outliers, as well as the handling of situations where hospitals have insufficient data to obtain a reliable result for any particular measure are omitted. These omissions do not affect the arguments presented.

The process starts with a set of over 60 individual quality measures. These are grouped into seven major categories: mortality, readmissions, safety of care, patient experience, effectiveness of care, timeliness of care and efficient use of medical imaging. The individual measures within each of these seven categories are combined using a technique known as latent variable modeling resulting in seven composite measures. These seven intermediate composite measures are then further combined to form a single composite measure. Four of the categories received a weight of 22% each and the other three categories received a weight of 4% each. This results in a single summary score for each hospital. A clustering method is then used to classify the hospitals into five groups based on these scores, and the clusters are labeled with star ratings from one to five. About 20 percent of hospitals do not receive any star rating, and certain classes of hospitals that do not participate in the Medicare quality programs are excluded, for example the critical access hospitals.

The analysis that follows was done using a combination of data from various sources. Data on hospital characteristics, such as bed size or level of disproportionate share were taken from the inpatient prospective payment system impact file that is published by the Centers for Medicaid and Medicare Services (CMS). The star rating model was simulated using data from the HospitalCompare website. The resulting star ratings differed slightly from those published by Medicare for 30 hospitals, but the differences never exceeded one star, and were in hospitals that were close to the borders between star rating categories. These differences would not materially change the arguments or conclusions presented below. In addition, the Medicare quality rewards and penalties imposed on hospitals were accumulated and expressed as a percentage of the revenues. The dollar amount of the rewards and penalties on Value Based Purchasing, Readmission Reduction Program and Hospital Acquired Conditions were added and then the result was divided by the estimated amount of Inpatient Prospective Payment System operating dollars to obtain the percentage impact of rewards and penalties.

# Examination of the results of the 5-star ratings

The analysis consisted of several different components. The first comparison was between the results of the star ratings and the aggregate quality rewards/penalties (RP) imposed on the hospitals. The next set of analyses looked at the summary score by different groups of hospitals to determine whether there were statistically significant differences in the mean scores by hospital size or level of disproportionate share. The analyses used data from the 3rd quarter 2016 release on HospitalCompare.

## Comparison of star ratings and rewards/penalties

In aggregate the star ratings and quality RP are consistent, but there are some aberrations. Looking at the mean percentage RP by star rating level, the mean RPs were statistically significantly different between the star levels, and were in the direction one would expect, i.e., the hospitals with higher star ratings had lower penalties or higher rewards. The distributions of the RP by star rating had huge overlaps, as is shown in Chart 1. There were 5-star hospitals that were hit with net penalties and there were 1-star hospitals that received net rewards for quality. The conclusion is that for some hospitals the RP and the star ratings are sending quite different messages about the quality of the hospital.

## Summary score by bed size quartile

The hospitals were sorted into four quartiles by bed size. The mean summary score was calculated for each bed size quartile. There were highly statistically significant differences between these means for all the bed size quartiles.

Table 1: Comparison of difference in mean summary quality score by hospital bed size quartile

| | | Unadjusted | | |
|---|---|---|---|---|
| SummaryScore | Contrast | Std. Err. | [95% Conf. | Interval] |
| bedquartile | | | | |
| 2 vs 1 | -.183009 | .0244683 | -.2309838 | -.1350343 |
| 3 vs 1 | -.2628444 | .0244833 | -.3108485 | -.2148403 |
| 4 vs 1 | -.3531509 | .0245059 | -.4011994 | -.3051025 |
| 3 vs 2 | -.0798354 | .0245498 | -.1279698 | -.0317009 |
| 4 vs 2 | -.1701419 | .0245724 | -.2183206 | -.1219632 |
| 4 vs 3 | -.0903065 | .0245873 | -.1385145 | -.0420985 |

The “contrast” is the difference between the mean summary score for the two bed-size quartiles listed in the left hand column. It can be seen from the fact that the 95% confidence interval does not include zero that all these differences are statistically significant. Chart 2 presents the data graphically and shows clearly the consistent pattern favoring smaller hospitals.

## Summary score by disproportionate share quartile

The hospitals were sorted into four quartiles by disproportionate share percentage. The mean summary score was calculated for each disproportionate share quartile. There were highly statistically significant differences between the means for all these quartiles.

Table 2: Comparison of difference in mean summary quality score by hospital disproportionate share percentage quartile.

| SummaryScore | Contrast | Std. Err. | Unadjusted [95% Conf. | Interval] |
|---|---|---|---|---|
| dshquartile | | | | |
| 2 vs 1 | -.2038801 | .0237767 | -.2504987 | -.1572615 |
| 3 vs 1 | -.2735082 | .0237695 | -.3201126 | -.2269037 |
| 4 vs 1 | -.4999311 | .0237767 | -.5465497 | -.4533125 |
| 3 vs 2 | -.069628 | .0237767 | -.1162467 | -.0230094 |
| 4 vs 2 | -.296051 | .0237839 | -.3426838 | -.2494182 |
| 4 vs 3 | -.226423 | .0237767 | -.2730416 | -.1798043 |

The "contrast" is the difference between the mean summary score for the two disproportionate share percentage quartiles listed in the left hand column. It can be seen from the fact that the 95% confidence interval does not include zero that all these differences are statistically significant. Chart 3 presents the data graphically and shows clearly the consistent pattern favoring hospitals with lower disproportionate share percentages.

## Regression models including both bed size and disproportionate share percentage

In order to account simultaneously for both bed size and disproportionate share percentage a regression model was constructed with the summary score as the dependent variable and bed size and disproportionate share percentage as independent variables. The results of this model are presented below:

```
. regress SummaryScore Beds DSHPCT

      Source |       SS       df       MS              Number of obs =    3290
-------------+------------------------------           F(  2,  3287) =  305.63
       Model |  136.238477     2  68.1192383           Prob > F      =  0.0000
    Residual |  732.609978  3287   .22288104           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1563
       Total |  868.848454  3289   .26416797           Root MSE      =   .4721

------------------------------------------------------------------------------
SummaryScore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Beds |  -.0003533    .000045    -7.86   0.000    -.0004415   -.0002652
      DSHPCT |  -.9937138   .0462655   -21.48   0.000    -1.084426   -.9030017
       _cons |   .3112314   .0169657    18.34   0.000     .2779669    .3444958
```

This model is highly statistically significant, as are the coefficients of both of the independent variables.

A similar model was constructed, but with the total percentage quality reward/penalty as the dependent variable.

```
. regress TotalQualityImpact Beds DSHPCT

      Source |       SS       df       MS              Number of obs =    3290
-------------+------------------------------           F(  2,  3287) =  152.47
       Model |  .047767407     2  .023883704           Prob > F      =  0.0000
    Residual |  .514885809  3287  .000156643           R-squared     =  0.0849
-------------+------------------------------           Adj R-squared =  0.0843
       Total |  .562653217  3289  .000171071           Root MSE      =  .01252

------------------------------------------------------------------------------
TotalQuali~t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Beds |  -.0000174   1.19e-06   -14.61   0.000    -.0000198   -.0000151
      DSHPCT |  -.0080355   .0012265    -6.55   0.000    -.0104403   -.0056306
       _cons |  -.0004615   .0004498    -1.03   0.305    -.0013434    .0004203
------------------------------------------------------------------------------
```

Once again, this model is highly statistically significant, as are the coefficients of both of the independent variables.

## Conceptual concerns regarding the 5-star rating method

The first step in the assignment of the star rating is the calculation of the seven category scores from the 60+ individual quality measures. This uses a statistical technique known as latent variable modeling. The theory underlying latent variable models can be found in [3] and [4]. The construction of a latent variable model requires an initial assumption that the observed or manifest variables (the initial quality measures in this discussion) are projections of linear combinations of unmeasurable underlying or latent variables. In this particular instance, it is further assumed that they are projections of a single latent variable. Thus, in the case of the mortality measures, it is assumed that there is an underlying mortality rate for each hospital, and the mortality rates for acute myocardial infarction, coronary artery bypass graft, chronic obstructive pulmonary disease, heart failure, pneumonia, and acute ischemic stroke are all derived from that overall mortality rate (along with a random error term). This is a far-reaching assumption and unlikely to be valid. By combining the individual mortality measures in this way the methodology is throwing away a lot of information that is contained in the individual measures. It is quite a stretch to assume that a hospital that has a low mortality rate for pneumonia is going to also have a low mortality rate for stroke and cardiac problems and vice versa.

The results posted by CMS in their Updates and Specifications Report[2] prove that this is a valid concern. Looking at the scree plots provided in Appendix E of that report, Figure E.2 (Safety of Care Group) shows that the (first) latent variable (principal component) captures less than 20% of the variance in the measures and that even adding two more latent variables (or principal components) still captures less than 50% of the variance. An examination of the scree plots proportion of the variance explained should convince any informed and objective reader that a single latent variable is not adequate to capture the information in the individual quality measures.

The individual quality measures within each of the seven categories of measure are combined using "loading coefficients", which can be thought of as relative weights. Looking at the "Efficient Use of Medical Imaging" category, two of the five quality measures have small negative weights and of the other three, one makes up 2/3 of the total. In other words, the measure for this category is being largely driven by a single quality measure – "abdominal CT use of contrast material".

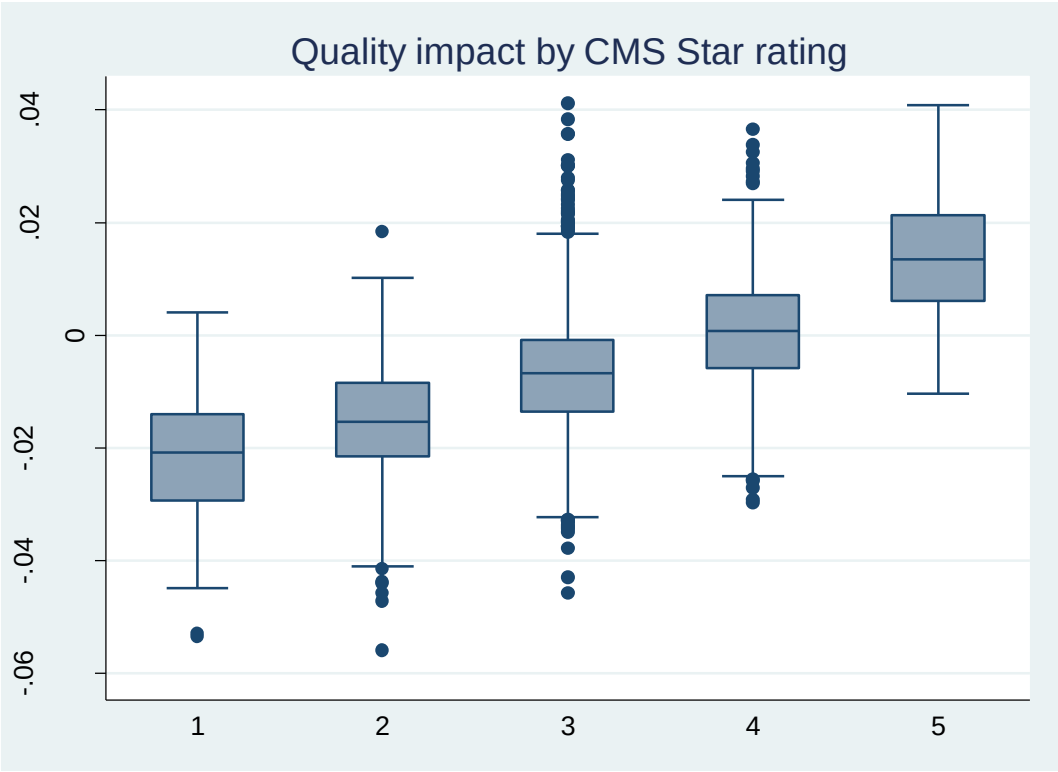
The Safety of Care category is also driven largely by a single measure – Complication/Patient Safety for Selected Indicators – that receives a loading coefficient of 0.93. The next highest loading coefficient in this category is only 0.17 and HAI-6, Clostridium Difficile, gets a loading coefficient of 0.001. This contributes negligibly to the category score, but it is clearly an important measure from a patient perspective. These are additional indicators of the lack of appropriateness of a latent variable model in this context.

## Summary and conclusions

The Medicare hospital 5-star rating system suffers from multiple problems. It is clearly biased against larger hospitals and safety net hospitals, i.e., hospitals with a high disproportionate share percentage. This exacerbates the problems caused for large and safety net hospitals by the quality related penalties imposed on them. In addition, the use of a latent variable model to construct the seven category scores from the individual quality measures is conceptually unsound and results in some of the selected quality measures contributing virtually nothing to the final rating (e.g., clostridium difficile) while others are given an unduly high weight (e.g., Complication/Patient Safety for Selected Indicators and abdominal CT use of contrast material).

## Charts

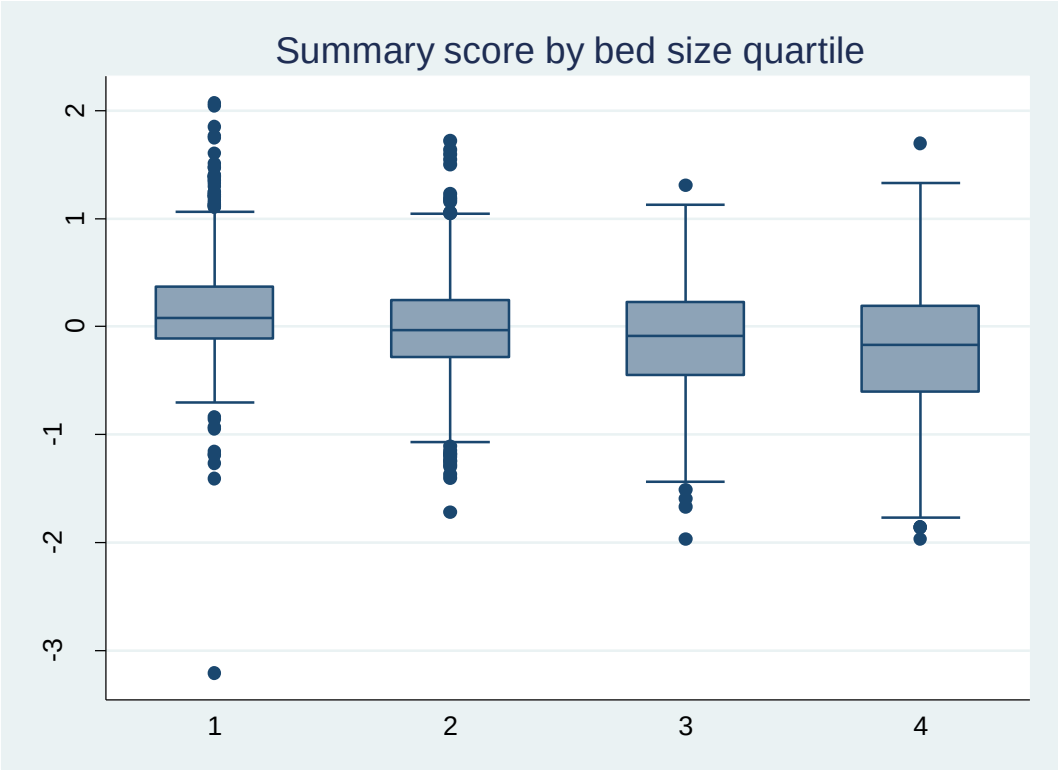
Chart 1: Box and whisker plot showing the range of the quality reward/penalty percentage by star rating




The shaded boxes show the range of quality reward/penalty for the middle 50% of the hospitals with any particular star rating. The lines above and below show the range of the rewards/penalties, except for outliers, and the dots represent the outliers.

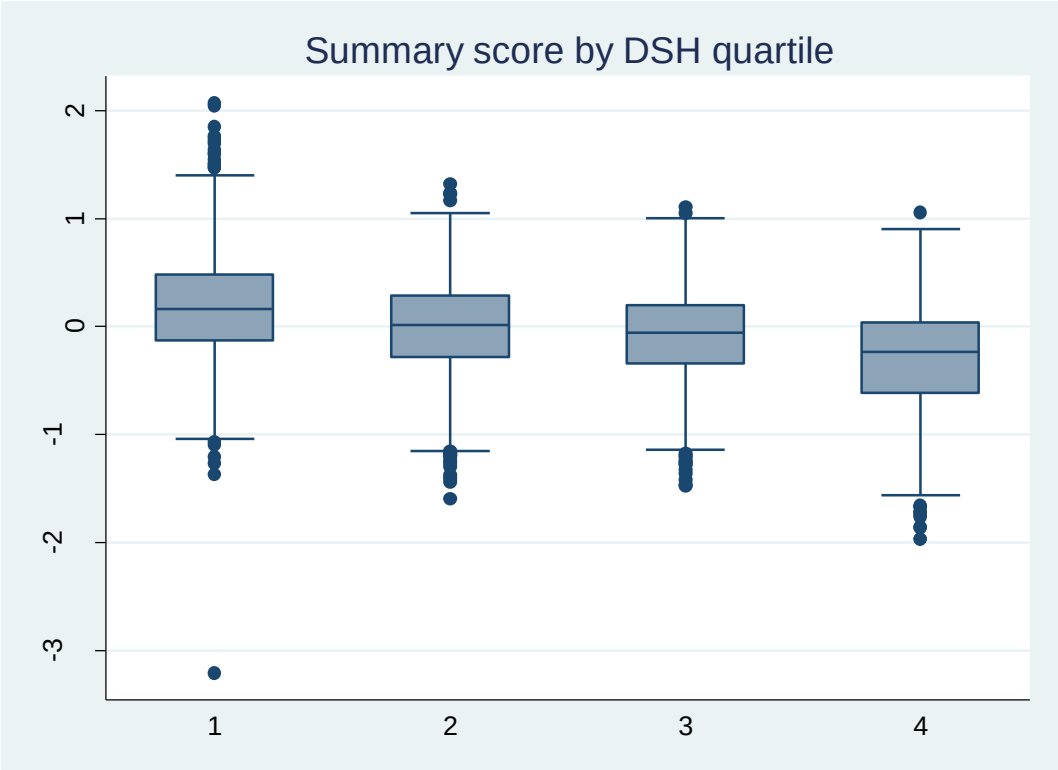
This chart demonstrates that the rewards/penalties vary enormously within the star rating categories and that some hospitals being classified as highest quality in the star rating system are being penalized for poor quality, while other hospitals classified as being of low quality are receiving financial rewards for their quality performance. It is easy to explain why this happens, but the fact that it does is indicative of a lack of cohesiveness in the quality measurement systems.

Chart 2: Summary score by Bed Size Quartile



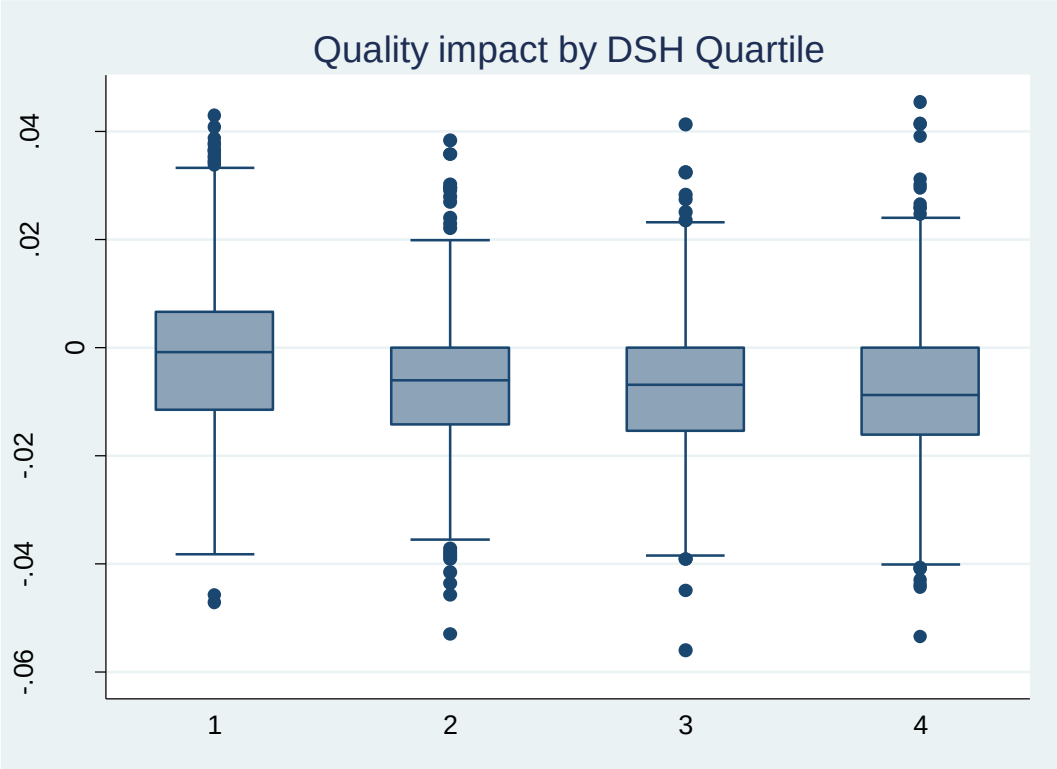

If large and small hospitals were similar in their summary star rating score, one would see the box and whisker plots to be more or less aligned horizontally. However, it can clearly be seen that the plots drop as one moves from the left to the right. That is, the smaller hospitals generally have higher scores than the larger hospitals.

Chart 3: Summary score by Disproportionate share quartile



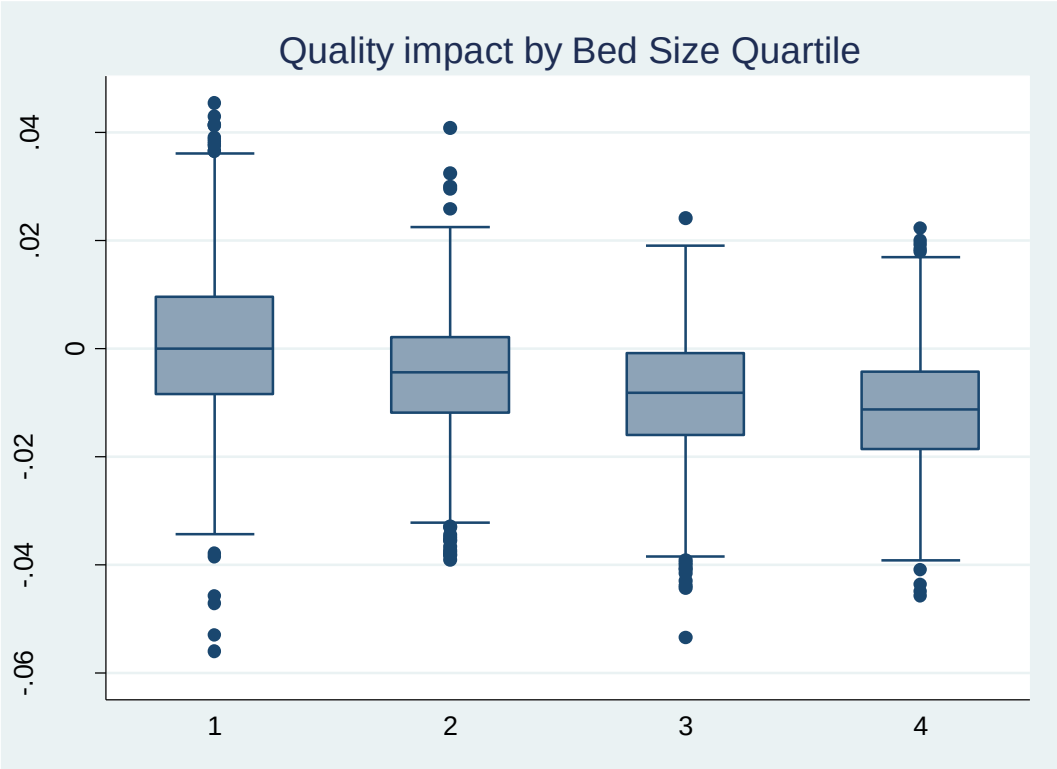

If hospitals with high or low disproportionate share percentages were similar in their summary star rating score one would see the box and whisker plots to be more or less aligned horizontally. However, it can clearly be seen that the plots drop as one moves from the left to the right. That is, the lower disproportionate share hospitals generally have higher scores than the hospitals with higher disproportionate share.

Chart 4: Percentage quality reward/penalty by disproportionate share quartile




If high and low disproportionate share hospitals were being hit with penalties equally one would see the box and whisker plots to be more or less aligned horizontally. However, it can clearly be seen that the plots drop as one moves from the left to the right. That is, the hospitals with lower disproportionate share generally have higher rewards/lower penalties than those with higher levels.

Chart 5: Percentage quality reward/penalty by bed size quartile




If large and small hospitals were being hit with penalties equally one would see the box and whisker plots to be more or less aligned horizontally. However, it can clearly be seen that the plots drop as one moves from the left to the right. That is, the smaller hospitals generally have higher rewards/lower penalties than the larger hospitals.

# Bibliography


1. Overall Hospital Quality Star Ratings on *Hospital Compare* Methodology Report (v2.0), retrieved November 2, 2016:

https://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890577152&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DStar_Rtngs_CompMthdlgy_052016.pdf&blobcol=urldata&blobtable=MungoBlobs

2. Overall Hospital Quality Star Rating on *Hospital Compare* October 2016 Updates and Specifications Report, retrieved November 2, 2016:

https://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890609207&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DStrRtg_Oct2016_QUS_Rept_083016.pdf&blobcol=urldata&blobtable=MungoBlobs

3. B.S. Everitt, An Introduction to Latent Variable Models, 1984, Chapman & Hall, New York, NY
4. David Bartholomew, Latent Variable Models and Factor Analysis, 18=987, Oxford University Press, New York, NY.