July 7, 2016

Patrick Conway, M.D.
Deputy Administrator for Innovation and Quality
Centers for Medicare & Medicaid Services
200 Independence Avenue
Washington, DC 20001

Dear Dr. Conway:

On behalf of our hospital and health system members, the American Hospital Association (AHA), Association of American Medical Colleges (AAMC), America's Essential Hospitals and Federation of American Hospitals (FAH) are deeply disappointed that we have not been given the opportunity to work with the Centers for Medicare & Medicaid Services (CMS) to examine the serious concerns we have with its star ratings methodology, nor has CMS shared any data to demonstrate the validity of its methodology. In addition, our continued review of the limited information available to us raises serious questions about the ability of the proposed ratings approach to provide accurate and meaningful information to patients.

We urge CMS to share additional information with hospitals and the public about how accurately star ratings portray hospital performance. We also urge CMS to address several significant underlying methodological problems with its star ratings. Until CMS has taken the time to address these problems and share information with hospitals and the public demonstrating that its star ratings methods offer a fair and accurate assessment of hospital quality, we strongly urge the agency to continue to withhold publication of the flawed star ratings.

To be clear, hospitals strongly support transparency on the quality of care they provide. That is why we brought organizations such as CMS, The Joint Commission, the National Quality Forum and others together more than a decade ago to launch a groundbreaking effort to provide credible, important information on hospital quality through public reporting. This multi-stakeholder effort led to the development of the *Hospital Compare* website. Hospitals are investing significant resources to collect, report and use data on hundreds of quality measures – for CMS and other payers and regulators – to inform the public about quality and identify opportunities for performance improvement. Through these efforts, the hospital field has been able to reduce harm and improve patient outcomes, and we remain committed to continuing to improve.

On March 18, we sent CMS a letter raising significant concerns about its star rating methodology, and asked the agency to delay publication until it could more closely examine the methodology. We specifically asked that CMS:

- Analyze the impact on different types of hospitals and provide more transparent information so that the fairness and accuracy of the star ratings could be evaluated.
- Consider the need for a sociodemographic adjustment for readmissions and other outcome measures to create fair comparisons.
- Examine whether the flaws in the hospital-wide readmissions measure and the patient safety indicator (PSI-90) measure bias the rating against hospitals that care for more complex patients.

Since our March 18 letter, we have brought to your staff's attention other concerns, including whether differences in the data reporting requirements for Maryland hospitals around the use of present on admission coding may have affected the ratings and whether small hospitals that have worked hard to drive down their central line and catheter-associated urinary tract infections to zero or near zero are disadvantaged because they do not have enough cases to have these measures count toward their star rating.

Unfortunately, we received virtually no additional information from CMS on any of the issues listed above. The agency provided additional information on how it calculates and assigns star ratings, but far too little information on whether the methodology gives a fair and accurate appraisal of the true quality of care provided in America's hospitals. Since the sole purpose for creating the star ratings is to provide accurate information to the public to guide their decision-making about where to get their care, hospitals and patients alike must have meaningful information on whether the assessment is fair and accurate. The very fact that some of the nation's best known hospitals with the highest of ratings on other assessments and that serve large numbers of low-income and complex patients are slated to receive a small number of stars from CMS should make one question the validity and soundness of the methodology.

Failure to provide information that might offer insights on hospitals' concerns serves no purpose; it does not meet the intent of what Congress urged CMS to do nor what CMS told Congress it would do. Indeed, in letters signed by the majority of both the House and the Senate, lawmakers questioned the accuracy of the star ratings and urged CMS to "work with Congress and members of the hospital community to resolve these concerns." CMS responded by delaying publication of the star ratings. In its notice to Congress, CMS stated that it was "committed to working with hospitals and associations to provide further guidance about star ratings." We are disappointed that this has not happened.

In the absence of additional information from CMS, we have continued to examine the small amount of information that is publicly available, including the updated methodology document (Version 2.0) that was published on the Quality Net website several weeks ago. Further, we asked an independent expert, Dr. Frank Vella, who is the chair of economics at Georgetown University and well-versed in the core aspect of CMS's methodology – latent variable modeling – to review the available information. **While the published methodology provides an adequate**

description of what CMS's assumptions were and what the agency did to calculate the star ratings, it provides very little insight into how well the methodology worked.

In fact, the minimal data available to us do not offer substantive proof that the methodology works as it was intended, and raise many more questions and concerns about the methodology than they answer. The independent analysis (attached) gave us strong reason to believe that the assumptions on which the current model is based are flawed in a number of ways:

- **The model fails to account for other factors that cause substantial variation in performance, and instead attributes the variation solely to quality.** Despite the fact that the methodology document asserts on page 27 that the latent variable analysis is valid, the data displayed in Appendix E actually show that in six of the seven categories, a single variable accounts for less than half of the variation.

- **The assignment of star ratings implies that hospitals have been measured on essentially an equal, or at least an equivalent, basis so that the comparisons are fair. However, that is not true.** In the methodology report, CMS indicates that a quarter of hospitals were assigned a star rating based on 18 or fewer measures, while other hospitals were assigned a star rating based on two or three times that number of measures. We note that this discussion in the methodology report refers to CMS having used 75 measures to assign star ratings, but, at other places in the report, CMS refers to and identifies only 64 measures to be used in assigning stars, so we admit to being a bit confused as to how many measures were actually used to assign stars to any hospital, and what the real variation is in the interquartile range of stars used. Still, it is clear that there is a large difference in the number of measures comprising a star rating for smaller and less complex hospitals versus the number used to assess larger, multispecialty hospitals. CMS provides no information that would show that these disparate bases for judging performance lead to fair and equitable comparisons.

  Further, it is not clear how many groups were used to assign the star ratings for smaller hospitals versus larger hospitals, nor is there any information that would allow us to understand if lacking results for any particular measure or for any particular group would make it more or less likely that a hospital would receive a lower (or higher) star rating. **Such a difference would represent a bias in the methodology that is attributable to decisions by CMS and its contractor rather than to the actual performance of the hospital. This is deeply troubling because the use of star ratings would make it appear as if the differences simply were attributable to the hospital's quality.**

- **The assignment of weights to measures and to groups of measures is completely arbitrary, and yet it likely has a significant impact on the number of stars assigned to each hospital.** At the very least, CMS should examine how the use of various weights contributes to the likelihood that a hospital would receive a particular number of stars. Further, CMS should explore how this weighting system affects the results when a number of hospitals have too few measures in a particular category to have a score from that category.

- The model groups the measures CMS selected into seven categories to perform the latent variable analysis and assumes that the categories are independent of each other, and that there is a common factor among the measures within each of the categories. **CMS offers no proof that either of these assumptions, which are vital to the use of a latent variable model, is true.** In fact, simply by looking at the measures lumped into each of the categories, it is clear that there exists at least some commonality among the categories. To the extent to which there is a commonality across the categories, CMS is essentially double-counting the influence of some performance on the overall star rating.

- **CMS crafted a methodology designed to maximize the difference in scores between hospitals in each of the star rating categories, yet its own data show this is not happening in some cases, particularly for Efficient Use of Medical Imaging.** In other words, CMS sought to ensure that those hospitals receiving one star had a different level of performance than those with two stars, those receiving two stars were different than those with three stars, and so forth. So it is not surprising that Pairwise Comparison of Star Categories would reflect differences in scores. What is surprising is the fact that very few of the pairs shown for the Efficient Use of Medical Imaging category show any statistically significant difference in performance. Between the scores of the hospitals CMS rated as one star hospitals and those it rated as five star hospitals, the difference in scores in this category was only 0.35, and that was not a statistically significant difference. In other words, CMS cannot tell the difference in performance among hospitals on Efficient Use of Medical Imaging, and yet the agency includes those measures in the star ratings.

In summary, we urge CMS to share additional information with hospitals and the public about how accurately its star ratings portray hospital performance. We also urge CMS to consider the several significant underlying methodological problems with star ratings laid out in this letter. Until CMS has taken these steps, and engaged in additional work with hospitals to validate the methodology, we strongly urge the agency to continue to withhold publication of the flawed star ratings.

Sincerely,

American Hospital Association
Association of American Medical Colleges
America's Essential Hospitals
Federation of American Hospitals

Attachment: Francis Vella's Critique of the Star Ratings Methodology as described in the Methodologic Report Version 2.0

GEORGETOWN UNIVERSITY

Francis Vella, Professor, Department Chair, and Villani Chair *fgv@georgetown.edu*
Department of Economics *Fax: 202-687-6102*
Washington DC 20057-1036 *Tel: 202-687-5573*

The objective of this report is to create and employ a methodology by which one can combine a number of various measures of hospital performance into a single measure which can then be used to rank hospitals via an ordinal ranking reflected by the number of stars. In addition to the rankings on the basis of the stars the measure is used to categorize hospitals as above average, average, or below average based on national averages.

The objective of my document is to evaluate what has been done rather than suggest an alternative. However, an objective reader might ask if potential patients are actually more informed when a scoring system takes a set of informative quality measures which are easily understood and aggregates them into a single measure which essentially has no underlying metric. Especially when the measures employed in the aggregation are somewhat arbitrarily chosen and weighted (and reweighted).

Another fundamental concern is that the approach adopted does not consider anything apart from quality outcomes. For example, it does not adjust for the nature of the patients or the different circumstances hospitals might encounter. I cannot see how one can ignore the implications of these factors on such measures such as mortality etc. Two (or more) identical hospitals could have very different outcomes depending on the type of patient they have, where they are located, the type of health issues they typically face and multiple other factors.

The first part of the project is to map the various measures of quality into a latent index. Before proceeding to the technical issues the authors need to decide on the measures chosen. This is clearly an extremely important part of the exercise. Clearly there are technical issues related to the most efficient use of information but that is ignored in the report.[1] However, the authors of the study need to show, given that the choice of measures is arbitrary, that the results are not sensitive to the inclusion or exclusion of any particular measures. Moreover, the failure to include measures when an insufficient number of hospitals report them introduces a possible bias. That is, if hospitals do not report measures on which they do poorly then failing to include the measure in the estimation procedure introduces a bias. Note that even when the hospitals do not have the capacity to fail to report such measures the implementation of such an approach may inadvertently introduce bias through the choice of measures employed.

I have no objections to the standardization of responses. However, the use of winsorization on the basis that the responses are "inaccurately reported" is troubling. The authors should report the results without winsorizing the data to see how important this process is to the final results. If they are very different this would raise grounds for concern.

The choice of groups seems reasonable until one inspects the assumption of the latent variable modeling (LVM) approach. Given the nature of the groups it seems difficult to argue that they each measure a distinct aspect of quality. This is important as it leads to double counting. That is, in the instance where two groups captured exactly the same aspect of quality including both groups would count the same measure twice. This is not innocuous as the same aspect of quality would then be contributing two times to the value of the latent value when it should be included only once. In fact, if one looks at the 3 assumptions underling the LVM approach listed on page 13 there is strong reason to argue each is violated in this setting. The authors need to check the robustness of these assumptions.

The latent factor model is presented on page 15. It is here that one sees how the quality outcomes should also be factors of other outcomes and that these should be included as additional regressors in the model. Failure to do so has implications for the model's estimates.

Although the model is not complicated I feel it is not well presented sufficiently clearly. Essentially the procedure explains the variation in the standardized outcomes as a function of a latent variable, capturing a common effect for each hospital for each measure in a group outcome, and an unobserved error. That is, if a hospital has similar high responses for all measures in a group category it is assigned a high value for a latent variable. Similarly, a hospital which has low responses for all outcome measures in a group category will receive a relatively lower value for the latent variable. The estimated coefficient maps the latent variable into the standardized outcome. As neither the coefficients nor the latent variable are observed one needs normalizations.[2]

Due to the manner in which the model has to be estimated it is necessary to impose distributional assumptions. Some of these are normalizations and as noted by the authors are fairly innocuous. However, the assumptions about the distributional assumptions regarding the equation error are important for determining the likelihood function. It seems to be a very strong assumption that these errors are not correlated for observations on the same hospital. It also seems implausible, given that other factors such as patient composition and regional location of hospital have not been considered, that the errors are not heteroskedastic. These are issues which should be tested for given they have implications for the model's suitability. This is because specification errors of this form can have serious implications in models estimated by maximum likelihood. If the specification error results in the estimates being inconsistent this will produce incorrect estimates of the latent variable.

One issue which is very important in evaluating the suitability of the model is the signal to noise ratio of the model. That is, how much of the variability in the responses can be explained by the model. To evaluate this the authors should provide model diagnostics so that the readers can judge for themselves. In fact, the absence of model diagnostics (or output for the model) makes it very difficult to assess whether the model is performing well.

I find the degree of technicality involved in generating the values from the LVM approach somewhat inconsistent with the subjective manner in which the weights are assigned on page 17. In fact, it is likely to be the case that one could generate any desired ranking on the basis of these weights almost irrespective of the outcomes from the first step. I think this degree of arbitrariness greatly reduces the value of the "rigor" in the first part.

I see no justification of the use of winsoring on page 18. This is presumably to make some outcomes look more similar when the data says otherwise. Once again we should see outcome of study without this approach being employed.

I do not find the mapping of scores to stars particularly insightful. As the authors point out in their discussion of Step 5 on page 19, two hospitals with exactly the same score can be extremely different because they achieved the same score via different methods (i.e. one may do very well on one criterion while another may do well on another, unrelated, criterion). Similarity should be based on the hospitals being equivalent on all dimensions and I think the introduction of minimum values to be in a category somewhat achieves this. However, ranking the hospitals by stars is somewhat misleading as it indicates a qualitative jump as one goes from one category to the other and this may be inconsistent with reality and only reflects the scoring algorithm.

Before turning to a discussion of the testing methods it is clear that the most important issue is how well this procedure explains the data. The report provides no measure of this and as a result one cannot easily draw conclusions about the capacity of the LVM approach to explain the data. Note that this is a first order issue and should be addressed to give the reader some confidence that the approach is at least able to explain the observed outcomes.

One should also do testing of the model to examine issues such as model misspecification, incorrect distributional assumptions, correlation within clusters, heteroskedasticity etc.

I am not sure I completely agree with the report's conclusion that there is only factor for each group. There appears to be a large reduction in the variance for additional factors. Moreover, even if there was only one factor there is nothing which suggests it is the latent index they have estimated. As I noted above, it would be useful to see the residual variance.

The remaining issues discussed on model reliability section seem to touch upon largely inconsequential issues. I suspect that one could easily generate comparable results using a much simpler and more transparent approach.

In conclusion the approach appears to have several shortcomings. For the sake of summary I repeat them here. First, I do not see the net benefit of taking a multiple dimensional problem and summarizing it with a single measure.

Second, I do not feel the methodology is well explained although it is essentially straightforward. While it appears to give the impression of being rigorous and objective the estimation aspect is highly dependent on choice of measures and the weighting scheme is entirely subjective and highly determinant of the final outcomes. Also, I feel ignoring other determinants of quality outcomes (such as location of hospital and patient composition) potentially biases the results. Finally, the use of a star system is providing the sense that substantial differences may exist across hospitals when they do not.

---

[1] The procedure employed by the study is based on deriving a common element across different responses for the same hospital to infer the value of a latent variable driving the similarity across responses. A technical issue in this approach is how many responses are required for the same hospital to infer this information and what is the gain or cost of using more or less responses? An insufficient number of responses may not give an accurate estimate of the latent variable while responses on some measures may be uninformative and simply introduce noise into the procedure. A rigorous analysis of such an approach would examine how the estimate of the regression coefficients, their standard errors, and the values of the latent variable respond to changes in the number of responses which are available and employed.

[2] The normalization is required because the contribution of each trait is first determined by the "price" of the trait, captured by the regression coefficient, and the quantity of the trait possessed by each hospital. The contribution of each trait is the product of price and quantity. As neither the price or the quantity is observed there is an infinite number of combinations of price and quantity which would produce the same contribution. By normalizing the quantity of the trait to come from a standard normal distribution this allows the procedure to estimate the price. Note that the total contribution of the trait to the latent variable is determined by the weighting scheme which is independent of the estimation process. However, the normalization employed here has no implications for the value of the latent variable.