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.