Improving Quality Measurement

Design Principles for Quality Measures

J. Graham Atkinson, DPhil

Abstract: The use of quality measures to adjust health care payments and to rank providers is growing rapidly, but there are many problems with the quality measures that are currently being used. This article discusses some of these problems and then lays out some principles and procedures that should be used in the development and combination of quality measures. Many of the problems with existing quality measures would have been avoided had these principles been applied as they were developed. **Key words:** *composite measure, health care, hospital, quality measure, regression, risk adjustment, statistics*

THE use of quality measures to adjust the payments to hospitals and other health care providers and to rank them is growing rapidly, but insufficient thought has been devoted to establishing principles of how to appropriately measure quality. The purpose of this article is to lay out some principles that should underlie the development and combination of quality measures. The principles fall into 5 categories:

1. what cases, if any, should be excluded from the process;
2. how to choose variables that should be used in a risk adjustment process;
3. how to quantify the impact of the variables selected in step 2;
4. validation of the results of step 3; and
5. appropriate methods to combine individual measures into composite quality measures.

I will generally try to provide examples of situations in which these steps were done inappropriately, or not at all, illustrating the importance of implementing such principles. While the examples of flawed quality measures provided in this article relate to hospitals, the procedures and methods suggested in this article are equally applicable to quality measures being developed for all health care providers.

The implementation of these principles will not necessarily be easy. They require an understanding of statistical methods beyond simply applying them as if they were cookbook recipes, and they require an examination of the results to see whether they make sense and are consistent with any assumptions that were made in constructing the measure.

Author Affiliation: Jayne Koskinas Ted Giovanis Foundation for Health and Policy, Highland, Maryland.

The author has disclosed that he has no significant relationships with, or financial interest in, any commercial companies pertaining to this article.

This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Correspondence: J. Graham Atkinson, DPhil, Jayne Koskinas Ted Giovanis Foundation for Health and Policy, PO Box 130 Highland, MD 20777 (JGAtkinson@aol.com).

DOI: 10.1097/JAC.0000000000000321

GENERAL DISCUSSION

In order to make quality measures meaningful to consumers, outcome measures are required rather than process measures. Process measures tend to become ends in and of themselves, to the detriment of other factors that may be of equal importance, but neglected because they are not being measured and so not rewarded or penalized. However, it is necessary to ensure that the outcome measures cannot be gamed, and that they do not provide incentives to select against severely ill patients, or patients with low socioeconomic status (SES). This requires a carefully designed risk adjustment mechanism. It is also essential to have some estimate of the reliability or confidence level of the measures. For example, mortality rate is an objective measure, but because of small numbers of deaths and difficulties in making risk adjustments it has a large confidence interval, and different mortality measures provide quite different results and rankings, as has been demonstrated by Shahian et al. (2017). Also, many of the current outcome measures are inadequately adjusted for SES and possibly for other risk variables for which the hospitals, or other providers, should not be held accountable. Atkinson and Giovanis (2014) provide a discussion of the lack of an adjustment for SES in the original Medicare readmission penalty methodology.

It is good that people are now concerned about health care quality and its measurement. However, more work is required on continually planning how to improve quality measures, both through incremental improvement in the measures themselves and in the risk adjustment mechanisms, but also on the overall structure of the quality measurement process, shifting to less emphasis on process measures and more on outcome measures as they become fairer and more reliable.

In this article I am not attempting to provide a prescription for exactly how to implement the various steps discussed, just to point out the need to consider each of these steps and to enumerate some principles that should be

taken into account in design decisions regarding quality measurement.

RISK ADJUSTMENT

Risk adjustment is a critical component of any quality measurement system, as well as any case-based payment system. However, little consideration appears to have been given to the criteria used to select the variables that should be adjusted for, and the statistical or other method to be used to quantify the adjustment. In this section I discuss these issues and attempt to describe such criteria, as well as discussing the merits of the techniques that can be used to quantify the necessary adjustments. The appropriate adjustment factor to include varies depending on the particular dependent variable they are being used to adjust, and the purposes for which this dependent variable is to be used. For example, if analysis is being done to quantify disparities by race or gender, then race and gender may be included as independent variables to be studied, but not as factors to be adjusted for.

Probably the most important criterion that should be used to judge the appropriateness of a risk adjustment system is the answer to the question: How can I improve my performance score? The answer should *not* be:

By selecting a wealthier, more educated, or less sick patient population.

It should also *not* be:

By making some arbitrary reclassification, such as, defining more of my beds as intensive care unit (ICU) or coronary care unit (CCU) beds, as is the case for the current measure for hospital acquired infections, as demonstrated by Fuller et al. (2019). The risk adjustment method used in National Healthcare Safety Network standardized infection ratios uses the number of ICU beds as an independent variable but Fuller et al. reported that “across-hospital variation in reported ICU utilization was found to be unrelated to patient severity.”

Exclusions

The first step is to define what cases should be included in the analysis and which should

be excluded. The cases to be included are usually defined in terms of some medical procedures or diagnoses. The decision regarding what patients with the particular procedures or diagnoses should be excluded from the analysis is more subjective. The basic principle is that cases should be excluded if they are atypical in some way, and because of their atypical nature you would not want to hold the provider accountable for their quality outcome or to adjust for their atypical nature by including a variable that indicates or quantifies their atypicality. For example, if the measure of interest is the mortality rate then one might want to exclude patients who had a do-not-resuscitate order at their admission, since they are likely to be particularly high risk, and allowing them to die is following their expressed wishes. Thought should also be given to such factors as how to treat patients that are transferred in from, or transferred out to, another hospital.

Appropriate risk adjustment variables

The independent variables used in the risk adjustment process should be patient characteristics that influence the dependent measure, and that are objective and measurable. The independent variables used in the adjustment should be correlated with the dependent variable (ie, the quality measure) and there should be reason to expect that the correlation is causal with the direction of causality, being that the independent variable would be expected to influence the dependent variable.

Examples of specific potential independent variables could be age, gender, weight, admitting diagnosis, or diagnosis-related group.

Criteria for suitable variables should include:

- Can be measured objectively
- Are expected to causally influence the dependent variable
- Cannot be manipulated easily by the provider

You would not want the provider deliberately selecting not to treat patients with particular values of the variable simply to improve their measured performance. For

example, you would not want the provider to select less sick patients just in order to improve their performance score, although it would be acceptable for a provider to screen and exclude patients they did not have the capability to serve because of lack of staff expertise or equipment.

Patient or institution-level variables for risk adjustment

It is generally easier to obtain average values for some variables at the provider-level than patient-level data. The provider-level data are less likely to be confidential and the data sets are smaller and so easier to work with. However, much information is lost in moving from individual patient data to aggregated data at the provider level and many subtleties may be lost in the aggregation process. Whenever possible it is better to use patient-level data to retain the maximum amount of information. When aggregated or provider-level data are used, one should ensure that the provider-level measure is reflective of a real difference in the underlying patient characteristics and not simply the result of decisions or classifications being made by the provider, possibly to improve their apparent performance. However, some variables are not available at the patient level, for example, the teaching status of the hospital, and teaching status is believed to have an impact on the severity level of the patients admitted (with teaching hospitals attracting more difficult patients) that is not fully captured by current case-mix adjustment tools, so probably is a legitimate variable to include. The number of intensive care beds, on the other hand, is much more arbitrary. It would be much better to use a patient-level variable that measures the level of severity of the patients.

Discussion of some specific variables

It will sometimes be unclear whether particular independent variables should be included for risk adjustment purposes or not. In these instances one might ask what the incentives or implications for providers are of adjusting or not adjusting for the particular variable. As an example I have discussed the

case of SES and its use as an adjuster in quality measures, with a particular emphasis on hospital readmission rates.

Socio-economic status is rather difficult to measure, so a proxy measure often used is Medicaid eligibility, which is generally available in the data sets being used for quality analysis. It has been demonstrated in many studies that SES is strongly correlated with hospital readmission rates, and a causal relationship can easily be postulated. Patients of lower SES have less money to fill prescriptions, may have poor living conditions, and may be less compliant with postdischarge care instructions. All of these could make a readmission more likely. However, for many years the Medicare program refused to account for SES in their risk adjustment of readmission rates. They argued that that would be setting lower standards of care for patients of lower SES. As a result, on average, safety-net hospitals were hit with higher readmission penalties than other hospitals, so had less resources available to deal with their problems. In fact, safety-net hospitals, like other hospitals, had a strong incentive to reduce readmissions with or without an SES adjustment. Following a recommendation from the Medicare Payment Advisory Commission (MedPAC), Congress has now mandated the use of SES peer groups in the calculation of readmission penalties, but no adjustment is made for SES in the other quality measures.

STATISTICAL OR OTHER ADJUSTMENT METHOD

The main decision to be made under this heading is whether the adjustment should be calculated using a statistical model such as, for example, a linear regression (for continuous dependent variable), logistic regression (for binary dependent variable), or using a categorical adjustment made by classifying the independent variables into multiple discrete categories and calculating means or medians for the various categories and then using these values as adjustment factors. When discrete categories are used these categories are nonoverlapping and usually, but not always,

encompass the universe of possible values. In some cases a categorical approach is the only one possible, for example, gender and race are naturally categorical variables, but there are variables that are naturally continuous, but for which a categorical approach might be constructively applied. An example might be age for a situation in which the dependent variable was not linearly related to the age of the patient or, worse, not monotonic with age. For example, mortality tends to be relatively high in the first year of life, then drops, and increases again with advancing age increasing more steeply with advanced age. A linear regression is not going to capture these effects, so it may be better to classify age into ranges and treat these ranges as values of a categorical variable. In this way one could account for U-shaped, or even more complex, distributions.

Model selection

Nearly all quality measures involve the use of a regression model of some sort, often linear or logistic. However, very little thought is given to the question of whether the relationships are, or should be expected to be, linear or whether the other prerequisites for applying such a model are satisfied, for example, an assumption that variables are normally distributed. After having decided what variables should be included in a model, the next question should be, but rarely is, what is the appropriate functional form of the regression equation? For example, would the variables be expected to interact additively or multiplicatively, or is some other function appropriate? For example, some variables may have an exponential or logarithmic effect.

A simple example of this lack of consideration of the appropriate functional form occurred during a statistics course I took. The students were to use a linear model to predict the weight of a particular type of fish given its length, breadth, and depth. The model that was expected was of the form:

$$\text{Weight} = A \times \text{Length} + B \times \text{Breadth} + C \times \text{Depth}$$

However, the weight is proportional to the volume of the fish and the volume involves cubic units, and is probably approximately proportional to the volume of the box with the same dimensions as the fish. Thus, the appropriate equation is

$$\text{Weight} = A \times \text{Length} \times \text{Width} \times \text{Depth}$$

And taking natural logarithms this yields:

$$\ln(\text{weight}) = \ln(a) + \ln(\text{length}) + \ln(\text{breadth}) + \ln(\text{depth})$$

The moral of this story is that before blindly applying a linear or logistic regression model, consider how the independent variables would be expected to interact, and how they would be expected to influence the dependent variable, and have some reasonable rationale for the particular model eventually used.

Model assessment

An important, but again often omitted or inadequate, step is the evaluation of the model. In assessing a model the developers often look at the explanatory power of the model as a whole, perhaps using the R^2 or a C statistic. However, this does not take into account instability in the coefficients of the independent variables in the model. I am not suggesting that the coefficient of every independent variable must be statistically significant at the 95% level, but that totally insignificant variables should not be included, as they are likely to introduce volatility into the results. An examination of the individual coefficients can be instructive in pointing out structural problems with the model. A situation in which consideration of the individual coefficients should have led to rejection of the model is the latent variable model used by the Centers for Medicare & Medicaid Services (CMS) in its hospital 5-star quality rating system. Some of the quality measures were assigned negative coefficients by the latent variable model. This means that doing better on that quality measure lowers the score in

the 5-star rating—clearly not a reasonable result. Thinking about the underlying reason for this counterintuitive result should have resulted in a realization that a latent variable model with a single latent variable was simply not appropriate in this context, and that the different independent variables were actually measuring different aspects of quality and were not simply projections of some single unseen idealized measure of “quality” from Plato’s cave.

What I am advocating is a review of the results of the model, most of which are linear or logistic regression models, to assess whether the results make sense, both in terms of the sign and magnitude of the coefficients. In particular, it should be checked that small cell sizes are not resulting in extreme results for particular cases.

COMBINATION OF MEASURES

If quality measures are to be useful to consumers they must be comprehensible, and one component of that is ensuring that the consumer is not overwhelmed by a multitude of measures. One way in which that can be accomplished is by combining multiple measures into a single, easily comprehensible measure, such as a star rating. Star ratings are now commonly used by multiple organizations for ranking all sorts of sellers—restaurants, hotels, airlines, etcetera, and so are readily understood by consumers. There is a wealth of literature on ways in which composite measures can be constructed and a very readable review of these methods can be found at Schwarz et al. (2015). However, this extensive body of literature did not prevent the Medicare program from promulgating an absurd and inappropriate method for combining hospital quality measures into their 5-star rating system. The problems with this system are discussed in some detail in Graham Atkinson (2016), and are summarized in the next paragraph.

The first step in the assignment of the CMS 5-star rating is the calculation of 7 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 Everitt (1984) and Bartholomew (1987). 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 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 underlying “latent” 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. In fact, Nerenz et al. (2019) found that 4 underlying quality dimensions were required to adequately account for the variation in the various quality measures.

REFERENCES

- Atkinson, G., & Giovanis, T. (2014). Conceptual errors in the CMS refusal to make socioeconomic adjustments in readmission and other quality measures. *The Journal of Ambulatory Care Management*, 37(3), 269–272. doi:10.1097/JAC0000000000000042
- Bartholomew, D. (1987). *Latent variable models and factor analysis*. New York, NY:Oxford University Press.
- Everitt, B. S. (1984). *An introduction to latent variable models*. New York, NY:Chapman & Hall.
- Fuller, R. L., Hughes, J. S., Atkinson, G., & Aubry, B. S. (2019). Problematic risk adjustment in national healthcare safety network measures. *American Journal of Medical Quality*. doi:10.1177/1062860619859073
- Graham Atkinson, J. (2016). *An analysis of the Medicare hospital 5-star rating and a comparison with hospital penalties*. Retrieved from http://JKTGfoundation.org/data/an_analysis_of_the_Medicare_hospital_5-S.pdf
- Nerenz, D. R., Hu, J., Waterman, B., & Jordan, J. (2019). Weighting of measures in the safety of care group of the overall hospital quality star rating program: An alternative approach. *American Journal of Medical Quality*. doi:10.1177/1062860619840725
- Schwarz, M., Restuccia, J. D., & Rosen, A. K. (2015). Composite measures of health care provider performance: A description of approaches. *Milbank Quarterly*, 903(4), 788–825.
- Shahian, D. M., Wolf, R. E., Iezzoni, L. I., Kirle, L., & Normand, S-L. T. (2010). Variability in the measurement of hospital-wide mortality rates. *New England Journal of Medicine*, 363, 2530–2539. doi:10.1056/NEJMsa1006396