PAY FOR PERFORMANCE

The Measurement Conundrum

Kaveh Safavi, M.D., chief medical officer, Solucient, Evanston, Illinois

Today’s pay-for-performance dialog is characterized by cognitive dissonance. Healthcare leaders agree on the inevitable movement toward linking payment to some aspect of performance. Many even embrace pay for performance as a tool to improve whatever they see as the ills of the current system. However, providers have been forceful in their criticism of many pay-for-performance programs. Some suggest that current approaches are unsustainable—largely because of how performance is measured. Making sense of this contradiction—that pay for performance is good in principle but bad in execution—requires an understanding of the inherent limitations in any performance measurement process. Three axioms about performance measurement emerge from the debates:

1. Measurement matters only in context.
2. Measurement created for one purpose but used for another will cause problems.
3. Arguments about methodology are really about something else.

MEASUREMENT MATTERS ONLY IN CONTEXT

Performance measures can only be understood in context of the reason they were created; this is the intent-of-measurement test. At their heart, all performance measures compare actual performance to some ideal performance. The ideal is defined either normatively (i.e., the 90th percentile performance level) or based on clinical research best practices (i.e., based on evidence).

The original intent behind most contemporary healthcare performance measures is complex and varied. For example, some were first developed by professional liability carriers looking to assess and manage claims risk. Some were developed by public health researchers seeking to improve clinical quality. Others were developed in an effort to hold hospitals accountable for the public trust they enjoyed. At the center was a basic intent to use measures either to judge performance from the outside or to improve performance from the inside.

The basic data inputs and descriptive labels were the same (i.e., risk-adjusted complications). However, the measurement outputs were very different and reflected the intent behind their development. For example, in some cases a measure developer may only need to know the relative distribution of performance, while another may want to get at absolute levels of performance. Using the absolute value of a measure developed for relative ranking can be problematic. Consider a hospital board that ties compensation to an absolute improvement in mortality.
of 20 percent but uses a measure developed for relative ranking of hospital mortality. The hospital may experience significant improvement in rank without much improvement in absolute mortality, frustrating the executive team who fails to get a bonus. One reason for this apparent paradox may be that the hospitals whose results were considered in the first measurement period were different from the peer group considered in the second period. This is similar, for example, to high school class rank. A student ranked at the top of his or her class in any particular year may not compare, on the basis of absolute performance, to a student who ranked first in a different year, despite the fact that both students were first in class.

**MEASUREMENT CREATED FOR ONE PURPOSE BUT USED FOR ANOTHER CAUSES PROBLEMS**

Applying a measure from one situation to another can create problems. This is a variation of the first axiom. In this case, however, a measure developed to evaluate one type of performance is applied to another. For example, many of the current risk-adjusted measures of hospital performance have been applied to judge individual physicians in their care of patients. However, many current measures trace their roots to a whole-hospital, all-clinical-conditions screening tool such as Iezzoni’s Complications Screening Program (Iezzoni 2003, 110). The sensitivity and specificity of these measures were calibrated to large, heterogeneous general hospital populations. They were not designed to evaluate small patient populations such as those seen by an individual doctor. The model shows poor correlation with actual outcomes when applied to small or skewed populations, even though values can be computed. Hence, a model that is defensible in one setting is much less defensible in another.

**ARGUMENTS ABOUT METHODOLOGY ARE ABOUT SOMETHING ELSE**

This third axiom is a use-of-measurement problem. Measurement critics often argue about methodology because they are concerned about how a measure will be used. For example, will a performance measure allegedly developed purely for internal use be used to embarrass or financially harm the person or persons measured? The real concern is not that the measure has no value but that it will be used in an unacceptable way. The intensity of objection is directly related to lack of trust between the measurer and the measured. Given the lack of trust between payer and provider it is no wonder that providers have put significant muscle into challenging payer performance measures.

**THE ROLE OF ADMINISTRATIVE DATA**

Confounding this whole debate is the role of administrative data in performance measurement. Administrative data are elements created primarily for the purpose of tracking patients and receiving payment. The virtues of administrative data are well known. They are widely available, generally standardized, and relatively
inexpensive to collect for performance measurement compared to alternatives such as medical record abstraction. Also, a legacy of peer-reviewed research on the use of administrative data for performance measures exists. Criticism of this data is equally persuasive. These elements were created for payment, not for performance measurement; therefore, they are subject to biases that reflect financial incentives. They also lack clinical depth and nuance; therefore, one cannot always adjust for the underlying illness burden of patients.

However, ample evidence supports the use of administrative data in select applications of performance measures. A recent study of physician-directed care, by Catherine MacLean and her colleagues at RAND, demonstrates both the utility and limitations of this type of data (MacLean et al. 2006). They compared performance measures obtained from records to those calculated by administrative data. Physician data were used because physician bills contain information about procedures ordered; therefore, process-of-care measures can be evaluated. MacLean’s work, along with prior research on outcome measures, suggests five summary observations about the use of administrative data to measure performance:

1. Administrative data measure fewer dimensions of care than “clinical” data.
2. Administrative data are designed for population measures, not individual patient evaluation or evaluation of complicated subpopulations.
3. Administrative data are more likely to overstate quality, at least for process measures.
4. Administrative data yield similar results to medical records data for identical process-of-care measures.
5. Outcome measures from administrative data are more specific than sensitive.

There are two critical implications of these observations. First, clinicians who criticize administrative data and insist on clinical data as the only source of performance measures will likely look worse, not better, with clinical data. Second, there is an inherent bias toward specificity over sensitivity in outcome measures from administrative data. Administrative data provide an effective tool for finding cases that lend themselves to an improvement process. Something real is probably behind outlier performance. The preference for specificity was a deliberate consideration in the Agency for Healthcare Research and Quality’s Patient Safety Indicators, which were constructed exclusively from hospital billing data (Romano et al. 2003).

WHERE DOES THIS LEAVE US?

Successful pay-for-performance initiatives cannot start with a decision about what measures to use. The process must start with an agreement on the intent of the program. Is the program designed primarily to make payment more fair, or to create an incentive to improve quality, or to squeeze out waste and inefficiency?
From that starting point, a program sponsor can decide who will be the principal focus of the measure and what the measurement priority will be. For example, is the goal to establish relative rank or absolute levels of performance? If these assumptions are clear and the parties trust each other’s stated motives, it is much easier to reach agreement on a measure set that is appropriate, easy to use, and easy to explain.

Unfortunately, today’s pay-for-performance world lacks these precedent conditions. Therefore, we argue about measurement.

References

For more information on the concepts in this column, please contact Kaveh Safavi at ksafavi@Solucient.com.