The Methodological Problems with Pay for Performance

Linda Gorman, Ph.D.

For two decades, Americans have been told that the American healthcare system is broken.

Ideologues in favor of expanding the government-run systems to take over private-sector medical arrangements have attributed rising medical costs to patients who want too much care, and to physicians who are too willing to fatten their pockets to provide it. The solutions they typically propose ignore the problems with existing government-run systems and propose to cure the problems with more extensive regulation to control patient and physician behavior.

Pay-for-performance systems have long been championed as a way of eliminating medical practices that regulatory control advocates have identified as wasteful. Although the specific terms change with time, the current efforts champion value rather than volume, emphasize the use of evidence-based treatments, and push for comparative effectiveness research. Despite doubts about their effectiveness, these initiatives are making their way from discussions to policy.

In 2015, the Centers for Medicare and Medicaid Services (CMS) will begin adjusting payments for Medicare service based on the value of the care provided. "Value" is defined as a combination of the quality and cost of care. There are two problems with Medicare's value measure: the paucity of reliable quality measures and, thanks to Medicare's administered prices, the fact that prices for medical services are likely to be inaccurate outside of the cash markets for medical care.

The Education Model for Pay for Performance

The scarcity of vetted value measures in medicine have propelled some researchers to look at their use in other fields. For example, a Mathematica Policy Research Working Paper by Peterson and Schone argues that medicine can learn a lot from the value-added models developed for education.[1] This is surprising. In education, where research interest in applying the technique to teacher salaries has been an extremely active area for more than a decade, it is increasingly clear that serious statistical problems stand in the way of making pay for performance work. Even a cursory comparison of the two fields suggests that the health policy literature trails the education policy literature in its appreciation of the problems inherent in pay for performance.

Jesse Rothstein,[2] an economics professor at the University of California at Berkeley who specializes in school and teacher accountability, has shown that short-term value-added methods are inaccurate and do a poor job of identifying good performers. In *The New York Times* he noted that "it is quite possible for a value-added evaluation system to do more harm than good. Today, pilot studies have not been promising. We should not race ahead without careful policy development and testing."[3] One problem is that accountability policies relying on measures of short-term value-added measures do an extremely poor job of rewarding teachers who are best for students' longer-run outcomes.

In evaluating educational value-added models for medical use, Peterson and Schone argue that three criteria must be met to allow accurate measurement of a physician's contribution to an outcome. First, the outcome should be measurable on a continuous scale. Second, the outcome should be measured repeatedly for each person treated, and the score at the start of each measurement period should be reasonably predictive of the score at the end of the period. Third, a physician should be able to influence the changes in the score.

They suggest several types of outcome measures that meet these criteria, including glucose levels for patients with diabetes, cholesterol or blood pressures for patients with coronary artery disease, measures of functional status, and survey-based measures of health status. While acknowledging many difficulties, they conclude that value-added models for Medicare could provide a powerful way to control for patient and system-level characteristics, especially by rewarding providers for improvements in patient health.

Unfortunately, a review of the education policy literature suggests that pay for performance is unlikely to play more than a minor supporting role in education reform because it faces a number of unsolved methodological problems. The first problem is understanding the difference between inputs and outputs. Inputs are often easier to measure, but it is outputs that we really care about. Program designers often use convenient data to measure performance, with the result that they end up with pay-for-performance programs that are actually paying for inputs.[4]

In education, inputs are things like the time teachers spend in the classroom, how many minutes are devoted to math, how many minutes are devoted to vocabulary, and how much a school district spends on books. These inputs may or may not be related to how much children learn. In healthcare, inputs are things like whether a medical history was taken, whether the results of an examination are recorded electronically, the number of days in the hospital, and the number of nurses per patient. The inputs may not or may not be related to whether patients actually get well and stay that way.

Educational researchers find that while achievement gains are systematically related to observable teacher and school characteristics (inputs), the differences are small. The medical literature has reported similar findings. For example, a randomized trial of a quality improvement initiative of the effect

of pay-for-performance incentives on measurable outputs in small medical practices with electronic health records found statistically insignificant differences in improvement in rates of appropriate antithrombotic prescription, blood pressure control, and smoking cessation interventions.[5]

Although good outcomes are difficult to define in both education and medicine, education has long had the advantage of generalized agreement on the use of academic tests as a composite measure of students' ability to master what schools are supposed to be teaching. Unfortunately, as test scores began to be used to score school performance in addition to individual performance, the state standards movement pressured educators to move away from nationally normed tests. At one time, tests like the Iowa Test could be used to determine how well a student performed relative to every other student in the country. Now state tests are based on state standards that were not normed to anything, and even if one were to agree that tests are a good measure of educational outcomes, it is no longer clear what outcomes are being measured. In medicine, the problem is even more intractable because the important outcome will vary from case to case, with technological advancement, and with what the patient wants to do.

Assuming that an outcome measure can be found, it is often difficult to figure out how much each individual contributed to it. Student performance depends heavily on parental characteristics, and students learn at different rates at different ages. Finally, the efforts of a teacher in one year may not bear fruit until years later, as in the case of a middle-school teacher who succeeds in making an attempt at difficult reading seem like a reasonable task, but whose students take several years of practice before they can demonstrate the ability to understand the expansive vocabulary and complex sentence structure needed to generate high test scores.

In medicine, a patient with a difficult condition might see many physicians for a variety of reasons over a course of years. Figuring out who was responsible for what is likely to be impossible, given that factors such as diet, illness severity, patient compliance, level of support at home, and genetic variation may also influence the outcome.

If individual contributions could be isolated, there are still difficulties in comparing the values of the individual contributions in producing the outcome of interest. Work in educational policy shows that students improve at different rates even if they have the same teacher, and that teacher performance depends upon the composition of the teacher's classes. Teacher and student achievement are also affected by their surroundings. There is evidence that working with high-performing peers improves performance, as does matching teachers with schools that fit their teaching styles. Given the likely variation across students, classrooms, and schools, it is difficult to see how the value added by a specific teacher can reliably be compared with that of another teacher.

In medicine, as in education, physician effort accounts for a relatively small share of outcome variations. One systematic literature review estimates that physician effort typically accounts for less than 20 percent of the variation in performance measures.[6] Patients vary as much as parents. Variations in patient compliance, environment, and home life may mean that improving outcomes by a lot in one patient is far easier than improving them by a little in another. Even measures that look objective, like hemoglobin A1c to track blood glucose management, may have a large genetic or patient compliance component. Without taking patient differences into account, pay-for-performance measures may actually end up paying for differences in physicians' patient populations.[7,8]

After isolating individual contributions and how they compare, it is important to consider the stability of performance indicators. If some teachers are better than others, one would expect that an indicator of teacher quality would be constant over time and that teacher rankings would change little between one year and the next. Unfortunately, this does not seem to be the case. Estimates using common measures of teacher performance suggest that 30% to 60% of the measured performance of a particular teacher results from transitory noise in the data. Year-to-year correlations of estimated teacher performance quintile rankings are quite low, ranging from 0.2 to 0.3 in one study.[9]

Medical performance measures have similar problems. In a study of one insurer, 7 years of data were analyzed for eight performance measures over 20 integrated medical groups. Only 45% of all medical group measurements had sample sizes sufficient for reliable measurements, and measured performance over time was inconsistent for most of the performance measures.[10] Small sample sizes are also likely to produce significant instability. Individual physicians often treat a relatively small number of patients with the conditions included in performance measures, making it impossible to reliably detect individual performance differences.[11] Even if the measures were stable, few studies of measured performance variation at the individual physician level consider whether the absolute variation is likely to have any clinical meaning.[6]

Assigning Value

The CMS emphasis on paying for value is simply pay for performance by another name. CMS designates certain metrics as quality measures that determine value. Hospitals that perform better on CMS quality metrics get paid more than those that do not. This isn't value, at least as economists describe it.

In market-oriented systems, people choose what to buy and how much to pay for it based on their tastes and preferences and the amount of money that they have to spend. The consumer-directed healthcare programs that were increasingly popular in the private sector before "ObamaCare" effectively outlawed them provided people with budgets and let them express their preferences by choosing what to spend their money on. CMS has fought them tooth-and-nail, and now one of the major problem with CMS programs is that its system of administered prices does not allow patients to show how much they would value a service. Value, at least in this case, is in the eye of the

bureaucratic beholder.

The problem is that there is clear evidence that bureaucrats spending other people's money on third-party healthcare have very different values than those that arise when the private sector spends private money on health. To see this, one needs to look no farther than the changes that occurred in the Oregon Medicaid priorities list after a state committee took over the job of ranking procedures.

The Oregon Health Plan ranks treatments for various diseases and conditions from 1-680, in order of priority, with the idea that procedures will be covered in order as the state's Medicaid budget permits. As program costs have grown, the list of covered procedures has become shorter. Between 2002 and 2009 there was a radical reordering of the priorities.

Thanks to the reordering, many life-saving procedures that ranked high in 2002 ended up in a much lower position in 2009. As things changed, the pattern that emerged was that the importance ascribed to routine and preventative care important to public health averages generally increased, while the importance given to life-saving treatment for individuals fell.

In 2002, medical treatment for Type 1 diabetes ranked second. In 2009 it was in 10th place. Even though refusing treatment for Type 1 diabetes is a death sentence, it now ranks behind smoking cessation, sterilization, and drug abuse treatment in importance. Bariatric surgery, thought to be of prime importance for those concerned with reducing obesity, is ranked at 33. Preventing obesity is now considered far more important than taking care of a closed hip fracture (89), or paying for surgery for a strangulated hernia (176).

The Oregon Health Services Commission Web site explains that the 2009 list emphasizes preventive care and chronic disease management because these services are less expensive and often more effective than treatment later in the course of a disease. Preventive care for the healthy is more popular than treating those who are actually ill, as is the treatment of diseases with active political constituencies. This drift appears to be unavoidable when political processes are given control over medical decision-making.

The problem that must be faced in any pay-for-performance program, and in utilitarian listings of medical cost effectiveness produced by bureaucratic processes like those that drove the Oregon rankings and the pronouncements of Britain's National Institute for Clinical Effectiveness (NICE), is that their values may not be the same as those of patients they are supposed to care for. In fact, the results produced by bureaucratic rankings often conflict with the "rule of rescue," the presumption that saving the life of an individual in imminent danger of dying is more important than improving the quality of life of someone whose life is not in immediate danger, or of saving hypothetical future lives through prevention.

In 2006, 21 of 27 participants representing the public on the NICE Citizens Council recommended that NICE consider the rule of rescue in making rationing decisions. They reasoned that death is final, and the purpose of medicine starts with saving life. In summer 2008, NICE officials rejected the advice of the Citizens Council, removing the rule of rescue from any status in its decisions about healthcare rationing, asserting, "NICE and its advisory bodies must use their own judgement [*sic*] to ensure that what it recommends is cost effective and takes account of the need to distribute health resources in the fairest way within society as a whole."[12]

Conclusions

As these examples make clear, the value that CMS or any other bureaucracy places on providing a particular treatment for an individual patient is unlikely to accurately reflect those of the people being treated, or of the physician providing the treatment. The existing measures of physician performance used in pay-for-performance programs account for relatively small fractions of total outcome variance, and they may end up punishing or rewarding physicians for simple variance in their patients' genetics, socioeconomic status, and co-morbidities. If that happens, it will be bad for patients, and bad for the physicians who treat them.

Linda Gorman, Ph.D., is director of the Health Care Policy Center at the Independence Institute in Denver, Colo. Contact: linda@i2i.org.

REFERENCES

1. Peterson G, Schone E. Rewarding physicians for their patients' health outcomes: what can Medicare learn from education's value-added models? Working Paper. Mathematica Policy Research, June 2012.
2. Rothstein J. Teacher quality and educational production: tracking, decay, and student achievement. *Quarterly J Economics* 2010;125(1):175-214. doi: 10.1162/qjec 2010.125.1.175.
3. Rothstein J. Let's not rush into value-added evaluations. *NY Times*, Jan 16, 2012.
4. Gorman L. What's wrong with pay-for-performance? John Goodman's Health Policy Blog, Feb 13, 2013.
5. Bardach NS, Wayne JJ, DeLeon, SF. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *JAMA* 2013; 310:1051-1059. doi: 10.1001/JAMA.2003.277353.
6. Fung V, Schmittdiel JA, Fireman B, et al. Meaningful variation in performance: a systematic literature review. *Medical Care* 2010;48:140-148.
7. Hong CS, Atlas SJ, Chang Y, et al. Relationship between patient panel characteristics and primary care physician clinical performance rankings. *JAMA* 2010;204:1107-1113. Available at: http://jama.jamanetwork.com/article.aspx?articleid=186551. Accessed Nov 12, 2013.
8. Chien AT. Do physician organizations located in lower socioeconomic status areas score lower on pay-for-performance measures? *J Gen Intern Med* 2011;27:548-554.
9. Sass, TR. Stability of value-added measures of teacher quality and implications for teacher compensation policy. National Center for Analysis of Longitudinal Data in Education Research, Urban Institute, Brief 4, November 2008. Available at: http://www.urban.org/uploadedpdf/1001266_stabilityofvalue.pdf. Accessed Nov 12, 2013.
10. Rodriguez HP, Perry L, Conrad DA, et al. Reliability of medical group performance measurement in a single insurer's pay-for-performance program. *Medical Care* 2012;50:117-123.
11. Hofer TP, Hayward RA, Greenfield S, et al. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA* 1999;281:2098-2105.
12. Gorman L. Rationing care: Oregon changes its priorities. Brief Analysis No. 645. National Center for Policy Analysis, Feb 19, 2009. Available at: http://www.ncpa.org/pub/ba645. Accessed Oct 23, 2013.