Center on HEALTH CARE EFFECTIVENESS

ISSUE BRIEF

JUNE 2013

Using Performance Measures to Promote Evidence-Based Care: A Bayesian Approach

Timothy F. Christian, M.D., M.P.A., Thomas W. Croghan, M.D., Myles Maxfield, Jr., Ph.D.
Mathematica Policy Research

The processes that clinicians use to deliver care to their patients often diverge from the evidence-based procedures recommended in clinical guidelines. Persuading clinicians to use evidence-based practices is vital, however, and many policymakers, insurance payers, and clinical leaders see performance measurement as an essential first step to accomplishing this. In performance measurement, researchers use precise measures to gather, analyze, and report data on the performance of physicians, hospitals, and medical groups. But despite the extraordinary range of performance measures that have been developed, tested, and used over the past two decades, the quality of care provided in the U.S. continues to lag (World Health Organization 2013).

In this issue brief, we explore a unique approach to performance measurement based on the Bayes theorum. Over the long term, this approach could strengthen the relationship between the measurement of provider performance and the growth of evidence-based, patient-centered clinical practices.

Introduction

Despite the widespread use of performance measures to help improve care delivery in clinical practices, there has not been an accompanying surge in the quality of U.S. health care. There are many reasons for this lack of progress, including concerns about the strength of the evidence underlying the performance measures (Tricoci et al. 2009), the ways in which measures are used to encourage providers to enhance their care quality, and limitations of the measures that could skew performance scores. The validity and usefulness of some measures have also been criticized. Even when measures are based on sound evidence, they may not adequately address differences in how the same treatment affects different patients. The measures also may not be suitable for clinically important subpopulations (Hayward 2007), or they may not account for a patient's or clinician's personal preferences for certain services (Tinetti et al. 2008).

In this issue brief, we discuss whether a Bayesian approach to performance measurement might motivate providers to use clinical practices that are more evidence based and patient focused. We begin with a short review of Bayes theorem—a simple math formula used to estimate conditional probabilities—as it might be applied to measure development and scoring. We then discuss three challenges to measure development, especially for measures used to assess individual providers: the small number of patients with a given condition in a typical practice, the need to identify the most appropriate measures for each patient subgroup, and the long lag time between provider action (or inaction) and the delivery of feedback to the provider. A Bayesian approach may be helpful in addressing these challenges.

Please note that, although we provide a brief summary of Bayes theorem, an in-depth statistical discussion is beyond the scope of this brief (see Spiegelhalter et al. [2000] for a detailed discussion). However, we hope to show that a Bayesian perspective is often a more intuitive approach for conceptualizing data, compared with other methods.

Bayesian Versus Frequentist Approaches

In recent years, comparative effectiveness researchers[1] have begun to explore how well the Bayesian approach works compared with the more commonly used

MATHEMATICA Policy Research

The Center on Health Care Effectiveness (CHCE) was founded in 2010 to inform health care policy and improve patient outcomes. Building on more than four decades of Mathematica's substantive expertise, methodological skill, and corporate infrastructure, CHCE is at the forefront of comparative effectiveness policy analysis and research.

"frequentist" approach.[2] Both approaches have been used to design and conduct randomized controlled trials comparing two groups of patients who are similar except for the treatment in question (the "treatment" group receives the treatment under study, whereas the "control" group does not). In a frequentist approach, any knowledge or evidence available before the measurement period is *not* used to generate conclusions, although it may be incorporated into the study design. In addition, the required sample size for the control and treatment groups will usually provide a result that will stand on its own statistically; a treatment will either be deemed effective at a statistically significant level, or it will not.

In contrast, the Bayesian approach *does* use prior knowledge or evidence to generate conclusions. The basic theorem asks, "What is the probability of a particular result given the evidence accumulated to this point?" In other words, the likelihood of getting a particular result from a clinical trial, formally known as the *posterior probability*, depends on the findings of the trial as well as the pre-trial probability of the result, formally called the *prior probability distribution*, or "priors." Figure 1 shows this approach used to evaluate two measures for hypercholesterolemia.

Like randomized clinical trials, most performance measures are currently designed and scored using a frequentist approach. Because this approach does not allow the use of prior evidence to produce conclusions, most measures specify a "measurement period" that prevents researchers from incorporating this evidence. This means that providers must treat enough patients with the condition being assessed within a fixed time frame in order to have a statistically valid sample. Moreover, although most clinical guidelines account for different levels of risk and individual preferences when recommending that providers "consider" certain treatments, measure developers have taken a mostly dichotomous approach by simply excluding patients with certain low-risk characteristics, ignoring the largely multidimensional nature of clinical risks. The result is measures that may be inflexible, may not apply to certain populations, and may not translate into improvements in clinical practices.

Advantages of the Bayesian Approach

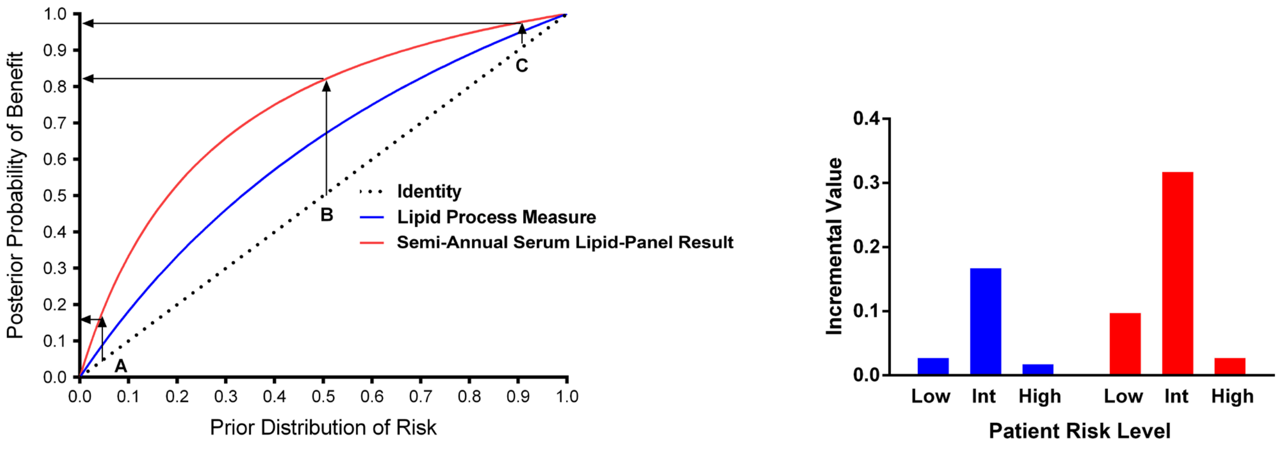
Insights into the use of the Bayesian approach in comparative effectiveness research suggest several ways this approach could enhance performance measurement. The benefits of this approach include the ability to handle smaller sample sizes, to measure value added, and to deliver rapid feedback to the providers being assessed. In the remainder of this brief, we will explore how these benefits can be used to tighten the link between health care performance measures and high quality care.

Minimizing the Effects of a Small Sample Size

A Bayesian approach can mitigate the challenges posed by a small sample size. For example, most providers see patients with a vast array of conditions, but except for some specialists, they only see only a few patients with any one condition (MedPAC 2007). Having a small sample of patients with a particular condition heightens the impact of those patients on a provider's

Figure 1.

Bayes Theorem Applied to Two Hypercholesterolemia Measures



Notes: In the line graph, the posterior probability—the likelihood that a patient will benefit if his or her provider complies with the measures—is plotted on the y-axis as a function of the pre-measure risk of adverse complications from coronary artery disease (x-axis). The incremental value of the measures (bar graph, right) is calculated as the difference between the prior distribution of risk and the posterior probability of benefit. The area under each measure curve in the line graph reflects the measure's ability to generate a benefit for the patient. Patients at low risk (A) have minimal benefit and, consequently, small incremental value for either the process measure or the more-intensive measure of lipid-panel results. Intermediate-risk patients (B) have the greatest benefit for either measure, as well as the highest incremental value from applying the more-intensive measure. High-risk patients (C) are a special case. In general, when the risk of adverse events is over 90 percent, measure compliance is unlikely to be effective without the concurrent use of more aggressive therapies. Incremental analysis allows researchers to compare different measures for similar diseases and to identify groups likely to benefit from measures that require extensive resources.

Identity = the point at which the value is the same on both the x- and y-axes.
Int = intermediate.

performance score, likely leading to a faulty score and major year-to-year changes in scores. Although mostly a problem when assessing individual providers, small sample sizes can also be a problem for groups of providers and hospitals (Schone et al. 2012).

A common frequentist solution to this problem is to extend the measurement period by a year or longer, which sometimes leads to more robust estimates of performance. The Bayes approach offers a more elegant solution, however, because it would not discard any potentially valuable prior evidence (Figure 2). For example, suppose a researcher is examining the quality of type 1 diabetes care. Type 1 diabetes is a condition treated by primary care providers, who over the course of a year see patients with nearly 400 unique diagnoses but only a few patients with type 1 diabetes (MedPAC 2007), as well as by endocrinologists, some of whom only see patients with diabetes. Using a fixed measurement period may therefore yield a stable estimate of quality for an endocrinologist, but not for a primary care provider. However, by incorporating all prior information—including data on care provided or performance scores on similar measures—a Bayes approach would produce a more stable, robust performance score for providers who only treat a few patients with diabetes. Researchers would therefore end up with a larger pool of potential providers to assess, as they would not necessarily need to exclude providers with only a small number of relevant patients.

Prior information can also be used to bolster the evidence underlying performance measures. In one study concerning this evidence, only 11 percent of clinical guidelines were based on high quality data from randomized trials (class A), and the remainder were based on less-robust study designs (class B) or expert opinion (class C) (Tricoci et al. 2009). The Bayesian approach addresses this issue by quantifying observational studies and converting them into a prior probability. New evidence can then be folded into the prior probability as it becomes available, and it need not be based on a certain sample size for consideration. This approach has been used for sequential clinical trials, when larger randomized trials were not practical, to gauge the increase in patients' survival chances associated with certain types of chemotherapy (Miksad et al. 2009). In this way, both prior and new evidence can be used in a measure.

Identifying Incremental Value

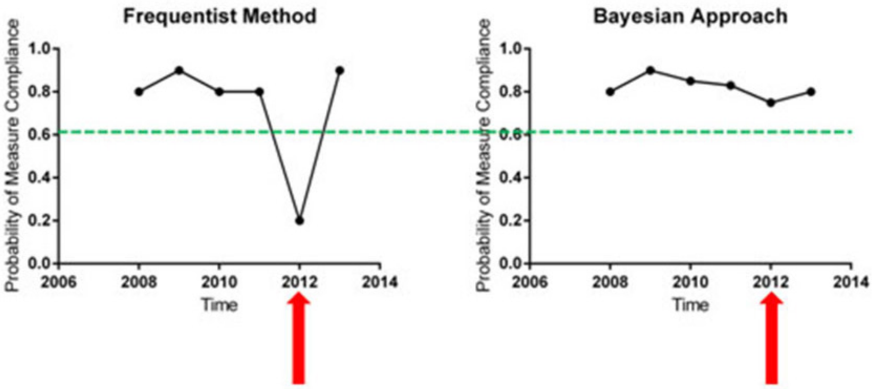
Another advantage of a Bayesian approach is that it allows researchers to calculate the incremental, or added, value of a given therapy for certain subgroups of patients. Formally, incremental value is the difference between the pre-intervention risk and the post-intervention probability of a benefit (Figure 1). Conceptually, it is a measure of how much a patient benefits from an intervention based upon his or her clinical risk. This type of analysis can help researchers and policymakers target performance measures to the therapies with the most value for patients at risk.

Good measures are sensitive to the fact that different therapies can be more or less effective for different patients. Even within groups of very similar patients, there are subgroups that benefit more (or suffer more risk) from a given therapy than others do. As the risks, benefits, and general effectiveness of a therapy can differ by subgroup, so too should the outcome measures. Taking these risks and benefits into account when developing measures helps ensure that clinical decisions that are truly based on each patient's needs are appropriately acknowledged and rewarded.

Consider aspirin therapy, for example. The U.S. Preventive Services Task Force has recommended that all men age 45 to 79 take aspirin "when the potential benefit due to a reduction in myocardial infarctions outweighs the potential harm due to an increase in gastrointestinal hemorrhage" (2009).

Figure 2.

Comparison of Frequentist Vs. Bayesian Approaches for Evaluating a Single Provider



Notes: In the frequentist approach (left panel), each year constitutes a new independent sample. Consequently, a year must pass before a sample can be obtained and the data can be tabulated, and prior good performance is not considered. The provider in this example may therefore receive a reduced reimbursement for year 2012 based on a few adverse outcomes and a small denominator. In the Bayesian approach (right panel), the sample still requires a year for data collection, but the calculation includes the prior performance scores, with a cumulative average dating back to 2008. Reimbursement would therefore not be reduced for this single adverse year. The adverse year may establish a trend, but whether such trends should be used to reward or penalize providers remains controversial.

However, subgroups of men vary considerably in their potential benefit from aspirin. For instance, both men and women should receive chronic aspirin (or other anticoagulation) therapy following an acute myocardial infarction, as the risk of another cardiac event is high for this subgroup. The cardiac risk is lower but still significant for patients with peripheral vascular disease, and it is significantly lower for those without any predisposing diseases or risk factors. It therefore makes sense to tailor measures based on each subgroup's level of risk (or potential benefit).

To ensure measures are implemented as efficiently as possible, researchers could match the incremental value for each risk subgroup with the *intensity* of the measure. Intensity refers to the amount of effort or resources required to enact a measure. Process measures based on Medicare claims data are relatively easy to acquire, for example, whereas a specific laboratory measure may need to be extracted from a clinical record and thus would require more resources and effort. The more-intensive measures could be reserved for patients who will benefit the most from them. For instance, suppose a set of measures is graded for hypercholesterolemia by risk group. For patients at high risk of a cardiac event, serial determination of lipid levels would be the most intense form of measure application. For those at lower risk, process measures such as counseling or documenting the prescription of lipid-lowering medications may be sufficient. Those at very low risk, for whom the benefit of lowering cholesterol levels is uncertain, could be exempted from any measure (Figure 1). Matching incremental value and intensity will ultimately make measure implementation more efficient, especially because it may reduce the infrastructure, financial, and cognitive burden on clinicians using the measures. (It should be noted that defining risk groups for this particular clinical question is well-established but may be more challenging for other diseases.)

Incremental value can also be used to identify better measures when choices are available. This is particularly relevant considering the plethora of measures that will become accessible when electronic health records become widely available for query. For example, one process measure for osteoporosis is documentation that a clinician ordered a bone densitometry scan or prescribed a bone-loss prevention drug within the last year. This measure is recommended for all women over age 65 and for women under age 50 with a hip or wrist fracture. However, with the advent of the electronic health record, the actual results of the bone-density scan will be available as a potential measure. Will this add incremental value for both groups of women?

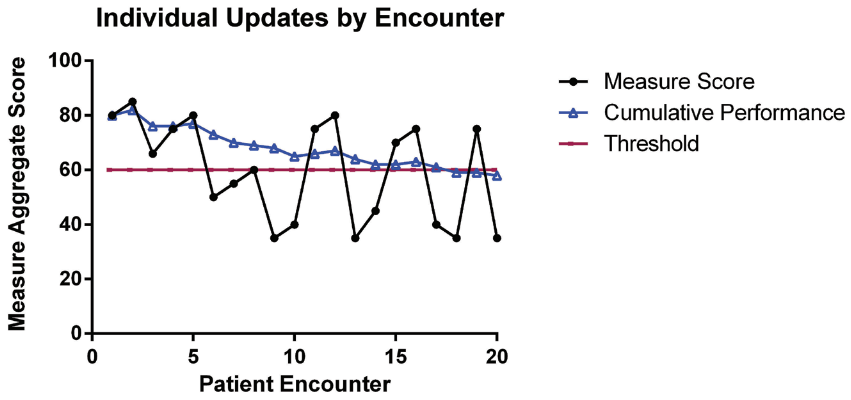
The answer is different for each group. For women over 65 with no prior fractures, the risk of osteoporosis is about 25 to 30 percent, and so a bone-density scan is appropriate. But because the scan result will provide more accurate information than the process measure, it would be wise to choose the new measure over the old one. The case is different, however, for women over 50 years with fractures, for whom the prevalence of osteoporosis approaches 100 percent. Because osteoporosis is almost inevitable for this group, the incremental value of a bone-densitometry result in this group is neglible. These women need therapy and prevention, not a diagnosis, and so neither measure would be appropriate for them.

Delivering Timely Feedback

A Bayesian approach may also allow researchers to quickly deliver feedback to providers on how well they are complying with the measures. Most measures require a fairly long period of data accumulation in order to have enough statistical power to evaluate each provider's performance. This can prolong the feedback process and create a disconnect between the evaluation and the state of practice at the moment. However, if each encounter between patient and provider could be added sequentially to a running tabulation of measure performance, providers could receive their feedback more promptly. This could be done by including all the prior data on the provider for a specified period, as shown in Figure 3. Once more data sources become available from electronic health records, such a running evaluation could be done in real time, with each "encounter" providing information on the clinician's performance. This approach could also be used to estimate the probability that a clinician's performance will improve over time.

Figure 3.

A Bayesian System for Provider Evaluation, with Individual Updates by Encounter



Notes: In this system, each patient-provider encounter produces a measure score, and overall measure compliance is recalculated with each encounter. The feedback is delayed only for as long as it takes to calculate the score, which may not be long with an electronic health record. In this hypothetical case, the provider has enough adverse scores to trigger a reduction in payment by the 17th encounter but also receives prompt feedback on each encounter.

Cumulative performance is the average of all encounter scores, previous and current. The threshold is the minimum performance score a provider must receive to avoid penalties.

Conclusion

Persuading clinicians to use evidence-based practices is a critical first step toward improving health care in the United States. One key to this effort is measuring clinicians' current practices and giving them timely, accurate feedback—a process well-suited to the use of a Bayesian approach. This approach offers an elegant solution to several common issues, such as small sample sizes, the need to craft appropriate measures for different patient subgroups, and the long delay before a clinician receives feedback on his or her performance. The recommendations in this brief, such as incorporating prior performance and feedback into a clinician's performance score, identifying patients' risk factors before applying a measure, and varying the intensity of a measure based on patient risk, can help minimize these problems. Bayesian methods can also help researchers determine good candidates for new measures, quantify the value of each measure, and ultimately control the costs and reduce the burdens of implementing the measures. The result should be performance scores that accurately reflect a clinician's quality of care and encourage the use of better, more effective practices.

Endnotes

[1] Comparative effectiveness research focuses on comparing the real-world effectiveness of various health care treatments.

[2] Frequentism is based on the idea that an event's probability is the limit of its relative frequency in a large sample of patients.

References

Hayward, R. "All or Nothing Treatment Targets Make Bad Performance Measures." *The American Journal of Managed Care*, vol. 13, no. 3, 2007, pp. 126–128.

MedPAC. "Report to the Congress: Assessing Alternatives to the Sustainable Growth Rate System." Washington, DC: MedPAC, 2007.

Miksad, R.A., M. Gonen, T.J. Lynch, and T.G. Roberts, Jr. "Interpreting Trial Results in Light of Conflicting Evidence: A Bayesian Analysis of Adjuvant Chemotherapy for Non-Small Cell Lung Cancer." *Journal of Clinical Oncology*, vol. 27, 2009, pp. 2245–2252.

Schone, E., M. Hubbard, D. Jones, and M. Wrobel. "Reliability of Inpatient Quality Measures: Implications for Public Use." Presentation at the AcademyHealth Annual Research Meeting, Orlando, FL, 2012.

Spiegelhalter, D.J., J.P. Myles, D.R. Jones, and K.R. Abrams. "Bayesian Methods in Health Technology Assessment: A Review." *Health Technology Assessment*, vol. 4, no. 38, 2000.

Tinetti, M.E., G.J. McAvay, T.R. Fried, H.G. Allore, J.C. Salmon, J.M. Foody, L. Bianco, S. Ginter, and L. Fraenkel. "Health Outcome Priorities Among Competing Cardiovascular, Fall Injury, and Medication-Related Symptom Outcomes." *Journal of the American Geriatrics Society*, vol. 56, no. 8, August 2008, pp. 1409–1416.

Tricoci, P., J.M. Allen, J.M. Kramer, R.M. Califf, and S.C. Smith, Jr. "Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines." *Journal of the American Medical Association*, vol. 301, no. 8, February 2009, pp. 831–841.

U.S. Preventive Services Task Force. "Aspirin for the Prevention of Cardiovascular Disease." 2009. Available at [http://www.uspreventiveservicestaskforce.org/uspstf09/aspirincvd/aspcvdrs.htm]. Accessed June 11, 2013.

World Health Organization (WHO). *World Health Statistics 2013*. Geneva, Switzerland: WHO, 2013.