The author argues against the use of pay-for-performance programs based on broad-based measures of patient outcomes in behavioral health care. He describes various problems with such programs. The purpose for collecting data is often not clear. Generic instruments do not measure improvement in specific mental disorders. Risk adjustment systems for behavioral health populations are not adequate. Mandated use of different instruments by payers is burdensome and precludes meaningful comparisons. The methodology for using outcomes-based approaches needs further development, and therefore the data collected will have little utility. The author discusses alternative approaches. (Psychiatric Services 60:108–111, 2009)

A great deal of attention has been paid in the past decade to performance improvement in health care. This reflects concerns expressed most clearly in a report from the Institute of Medicine (IOM) that documented large numbers of medical errors in the U.S. health care system (1). Subsequent IOM reports focused on ways to improve the quality of health care in the United States (2) and the quality of treatment for mental and substance use disorders (3).

Concerns about the quality of health care are not new. In the early 1900s Codman (4) exhorted surgeons to look at the outcomes of the operations they performed, which ultimately led to the development of the Joint Commission on Accreditation of Healthcare Organizations (JCAHO), now called the Joint Commission. In the 1960s Donabedian (5) divided quality assessment into structure, process, and outcome approaches, which all have their limitations. The Joint Commission initially focused on a structural approach that assessed an organization's infrastructure to support the delivery of care. This approach involves looking at the organization's governance and leadership structure, its credentialing practices, its policies, and so forth. Structural measures may indicate an organization's or practitioner's capacity to deliver good care but do not indicate whether the care was actually delivered.

More recently, the Joint Commission and other groups led by employers and insurers have worked with medical organizations to develop measures to assess clinical processes, such as the Healthcare Effectiveness Data and Information System (HEDIS) measures that have been widely used. Initially HEDIS included a measure of the use of beta blockers within seven days of hospital discharge for patients admitted to hospitals with an acute myocardial infarction; inclusion of this measure improved adherence to evidence-based guidelines for treatment of patients who received these drugs to close to 100%.

The hope and expectation has always been that improving adherence to evidence-based guidelines for care of patients with specific medical conditions would improve patient outcomes. This was the basis for the “100,000 Lives Campaign,” which was launched by the Institute for Healthcare Improvement with hospitals all over the country. The ultimate goal is to improve patient outcomes by improving the structure or process measures that determine the ability or actual delivery of health care services.

Health insurers, including Medicare, have tried to encourage such improvements by implementing pay-for-performance approaches, which tie rate increases or bonuses to reports of quality measures or performance on either process measures (Did you measure a HgA1C?) or outcome measures (Did you achieve an HgA1C level under 7%)? Financial incentives that get the attention of providers and lead to improvements in care or in patient outcomes are certainly a good thing. However, some practicing physicians have questioned the effort and expense of collecting, reporting, and analyzing such data and whether the measures and instruments employed are, in fact, evidence based and whether they will improve quality (7).

This Open Forum examines pay-for-performance approaches in behavioral health care that are based
on patient outcomes. Six problems with such programs are discussed, and alternatives are proposed.

Pay for performance in behavioral health care

**BCBSMA’s program**

Against the background described above, Blue Cross Blue Shield of Massachusetts (BCBSMA) announced that on May 1, 2007, it would launch “an outcomes measurement program in conjunction with our designated vendor, Behavioral Health Laboratories, Inc. (BHL), using their Treatment Outcome Package (TOP).” The emphasis on outcomes measurement represents a radical departure from more common approaches in mental health care that look at structure or process measures. The TOP consists of 70 questions written at the fifth-to-sixth-grade reading level, which takes five to eight minutes to complete. It was developed to follow the design specifications set forth by a Core Battery Conference convened by the American Psychological Association in 1994, and it provides scores in 12 clinical and functional domains (for example, depression or work) (8). In its May 1, 2007, letter to providers, BCBSMA stated that “Behavioral health has lagged behind most other clinical specialties in instituting standardized measurement processes for evaluating quality and monitoring outcomes. Responding to the recommendations of the IOM and many others, BCBSMA recognizes that implementing a measurement standard will enhance our ability to honor our commitments to our members and their health.”

To encourage participation, BCBSMA tied the entire provider fee increase for 2008 (3.5%–3.7%) to signing up for the program and announced that the entire 2009 fee increase of 3.5% would be tied to achieving participation rate targets. That meant that unless a sufficient proportion (60%) of new or returning patients agreed to answer these 70 personal questions (or 59 questions at follow-up visits) and submit the TOP to a for-profit company for scoring, the provider in question would receive no fee increase for 2009. BCBSMA subsequently modified its participation target by indicating that a patient could sign a form refusing participation and the provider could fax the form to BHL and get “credit” for that patient’s participation. BCBSMA also stated, “While we anticipate that the outcomes program may ultimately lead to a program of performance-based compensation at some point several years in the future, the current program offers increased compensation for providers who participate. We will not embark on such a pay-for-performance program until [our] Outcomes Scientific Advisory Council’s review of the data determines that we are ready for that step.”

**Problems with outcomes-based approaches**

With outcomes improvement the Holy Grail of quality improvement efforts and with outcomes measurement a necessary step in those efforts, what objections could there possibly be to such efforts? There are a number of problems with the approach described above.

First, the purpose for collecting the data needs to be clear. If, as implied, it is to improve the care that patients receive, what are the mechanisms through which quality improvement will be achieved? Or might the results be used to better manage utilization, or to demonstrate to purchasers that the plan is quality conscious and innovative in its approach?

Second, there is not much evidence that the use of standardized rating scales improves a patient’s continued participation in treatment, but it seems likely that the therapist will be perceived as more thorough and that the patient will appreciate having his or her care monitored systematically for improvement or deterioration. If that is so, is it better to use a broad-based generic instrument such as TOP or disorder-specific rating scales such as those used in clinical trials for medications? There is little empirical evidence to answer this question, but it seems likely that patients (and clinicians) will take more seriously changes in symptoms that are relevant to the presenting problems. This would suggest using briefer, more focused instruments such as the Personal Health Questionnaire-9 for depression screening (9) and the Hamilton Depression Rating Scale or Beck Depression Inventory for a presenting problem of depression, the Yale-Brown Obsessive Compulsive Scale for obsessive-compulsive disorder, or the Brief Psychiatric Rating Scale for psychotic symptoms.

Third, if the intent is to compare performance across the individual providers or group practices and then institute performance-based compensation, then the populations treated need to be carefully stratified by primary diagnosis and then risk adjusted by initial severity of illness, comorbid factors that may affect outcome (for example, concomitant substance abuse), and other specific demographic and clinical characteristics that influence outcomes independent of the quality of care provided. A recent review found that most published risk adjustment systems for mental health care lack sufficient explanatory power to be useful (10). The adequacy of BHL’s system for risk-adjusting TOP data needs to be demonstrated. It should also be noted that risk-adjusted comparisons require each clinician to have substantial caseloads of patients insured by BCBSMA in order to conclude with any validity that Doctor A is “better” than Doctor B for the treatment of elderly men with depression without comorbid substance abuse but “worse” than Doctor C for the treatment of younger women with anxiety and a history of childhood trauma. If such valid results could be achieved, then one might encourage Doctor A to treat more elderly men with depression or encourage Doctor B to get supervision, take a continuing education course on treatment of geriatric depression, or see only younger patients. At the level of system comparisons one might steer elderly patients to Clinic X rather than Clinic Y if outcomes were better at the former clinic for this age group (and if Clinic X had the capacity to treat...
every older person with depression). Such concerns are not unique to psychiatric care but have also been raised about efforts to collect and post data on cardiac surgical outcomes. Certainly patients would prefer to have their coronary artery bypass graft done at the hospital with the “best outcomes,” but are the currently available data able to distinguish the “best” hospitals or surgeons? And what happens when there’s not enough capacity at those institutions?

Fourth, if outcome measures are to be implemented by groups of providers or individuals, should they be mandated by individual payers? To incorporate such measures into routine clinical practice, it would clearly be preferable and administratively simpler to use the same instruments for all payers. Conversely, using different instruments for different payers would be a nightmare and require asking each patient at every visit whether their insurance had changed and if so to please fill out a new instrument. Switching instruments would also make it difficult to monitor the same patient’s care over time. Using different instruments also limits the usefulness of the results to the hospital’s or practice’s internal quality improvement efforts.

A big advantage of using HEDIS measures or those approved by the National Quality Forum or the Massachusetts Health Quality Partners is that they have broad acceptance by all payers on the basis of professional consensus. All of these organizations currently use process rather than outcome measures for behavioral health, reflecting the state of development of the field.

Fifth, is the outcome methodology scientifically defensible? The TOP instrument has been tested and validated for use with individual patients, so that it “has some ability to distinguish between behavioral health clients and members of the general population” (8). What is less clear is the plan for analysis of the huge number of forms that may be submitted. For example, what is the appropriate follow-up interval for completing a new form? Should it be resubmitted after a set number of visits or after a specified time has elapsed? There is no specific proposed interval, which complicates any attempt to compare providers. Similarly, response rates may vary widely between providers, raising questions about how representative the results are for an individual practitioner or provider. Will the data for patients who fill out only an initial form be eliminated from analyses, or to determine whether there are any systematic differences, will the patients in this group be compared with the group of patients who complete multiple forms?

If the BCBSMA project were submitted to a National Institutes of Health study section, these methodological questions about sample sizes, power calculations, sampling intervals, and so forth would have to be addressed, usually on the basis of some prior pilot data and before data were collected for thousands of patients. Furthermore, no journal would publish the results of such a study without being assured that an institutional review board (IRB) had determined that patients were asked to provide written informed consent after being told what use would be made of the data, what confidentiality protections were in place, what potential risks and benefits were involved, and what alternatives there were to participation. Although IRB approval is not needed for internal quality improvement activities, IRB review would be reassuring to patients given that the data are being sent outside the organization. It is not reassuring that an advisory counsel will be asked to help interpret the data after the data have been collected, rather than participating in the design before data collection.

Sixth, will the comparative data be of any use? Another large payer in Massachusetts, which serves a Medicaid population, attempted to mandate use of the TOP instrument but decided instead to let practitioners and providers choose from a list of approved instruments. However, the payer did analyze data collected from several thousand patients who had each completed the TOP instrument multiple times. Across diagnostic groupings, the payer found that for adults who had completed the TOP at five time points, depression scores declined somewhat from first to last administration. The utility of the results was limited because most of the patients were primarily (but not exclusively) substance abusers without a comorbid diagnosis of depression, and the multiple TOP administrations were compressed into a brief period after a detoxification admission.

The way forward
Currently, the outcomes initiative described in this Open Forum is limited to one payer in one state and affects one specialty. However, the implications should be of concern to all medical specialties. As a field, psychiatry and behavioral health have done considerable work over many years to develop quality measures (11). The measures that have received broad acceptance as part of HEDIS or from the National Quality Forum are all process measures—for example, the proportion of patients discharged from a hospital for a psychiatric condition who have a follow-up appointment within seven days and a medication follow-up visit within 30 days, or the percentage of school-age or adolescent patients treated with first-line medication in primary care for attention-deficit hyperactivity disorder whose medical record contains documentation of a follow-up visit twice a year. Adhering to process measures does not automatically lead to better outcomes. Other potential performance improvement initiatives could focus on screening for substance abuse or suicidality.

It would be highly desirable to ensure that whatever measures are chosen will have broad acceptance in the field and be clinically relevant to patients and practitioners. A study of early adopters of pay-for-performance programs suggests that there has been little evaluation of programs that are based on process measures (12). However, initial results have shown an inconsistent impact on quality improvement and very little evidence of cost savings but no important deleterious effects. Evaluation of these programs
will help determine whether they achieve positive results and are cost-effective. Such evaluations should be a required feature of any large-scale programs. Use of outcomes measurement for specific psychiatric illnesses should also be developed further and implemented more widely (13).

Pay-for-performance programs—whether they are based on process or outcome measures—should be used in addition to, not instead of, fee increases.

Acknowledgments and disclosures

The author is a board member and associate medical director of Health New England.

References

3. Institute of Medicine: Improving the Quality of Health Care for Mental and Substance-use Conditions. Washington, DC, National Academies Press, 2006
4. Codman EA: A Study in Hospital Efficiency: As Demonstrated by the Case Report of the First Two Years of a Private Hospital, Boston, Thomas Todd, 1920. Reprinted by the Joint Commission on Accreditation of Healthcare Organizations, Oakbrook Terrace, IL, 1995

---

**Psychiatric Services Invites Short Descriptions of Novel Programs**

Psychiatric Services invites contributions for Frontline Reports, a column featuring short descriptions of novel approaches to mental health problems or creative applications of established concepts in different settings.

Text should be 350 to 750 words. A maximum of three authors, including the contact person, can be listed; one author is preferred. References, tables, and figures are not used. Any statements about program effectiveness must be accompanied by supporting data within the text.

Material to be considered for Frontline Reports should be sent to one of the column editors: Francine Cournos, M.D., New York State Psychiatric Institute, 1051 Riverside Dr., Unit 112, New York, NY 10032 (e-mail: fc15@columbia.edu), or Stephen M. Goldfinger, M.D., Department of Psychiatry, SUNY Downstate Medical Center, Box 1203, 450 Clarkson Ave., Brooklyn, NY 11203 (e-mail: steve007ny@aol.com).