Evidence-Based Performance Measures: Preventing Unintended Consequences of Quality Measurement

Quality measurement and pay-for-performance programs can provide strong incentives to improve quality (1–3). However, performance measures can also have unintended consequences.

For example, consider performance measures related to antibiotic timing for patients presenting to hospitals with community-acquired pneumonia (4). The Medicare National Pneumonia Project adopted “first antibiotics within 8 hours of hospital arrival” as a national quality measure in 1998 and shortened the time window to 4 hours in 2002. The Centers for Medicare & Medicaid Services adopted this more strict measure as one of their initial core quality measures (PN-5b) for public reporting. In 2006, the PN-5b measure became part of a set used by several pilot pay-for-performance programs (5), and some hospitals gave bonuses to nonphysician staff who met a performance goal of 90% (6). However, the diagnosis of community-acquired pneumonia is often unclear during the initial evaluation, and the appropriate management of a stable patient is often to withhold antibiotics pending a more certain diagnosis. Yet, with public reporting and financial pressures in place, the PN-5b measure became the sword of Damocles hanging above physicians trying to choose between the best treatment for patients or following a national performance standard. Concerns arose that the measure was leading to unnecessary and inappropriate prescribing of antibiotics (7). Subsequent studies confirmed these fears (8, 9) and failed to confirm that early administration of antibiotics decreased mortality in stable patients with community-acquired pneumonia (10). In response, the Joint Commission relaxed the time window to 6 hours and created a new data element of “diagnostic uncertainty” that can be used to exclude patients from the measure.

In this issue, an American College of Physicians clinical practice guideline (11) and background review (12) raise similar concerns about performance measurement for prophylaxis of venous thromboembolism (VTE) for patients hospitalized for medical conditions or stroke. The guideline concludes that evidence does not support routine VTE prophylaxis in all medical patients. Risks for VTE and bleeding vary substantially according to individual risk factors, and prophylaxis may actually harm patients who are at low risk for VTE or at high risk for bleeding. The guideline recommends that clinicians assess the individual patient’s risk and benefit before initiating VTE prophylaxis. Moreover, it recommends against the use of performance measures that do not incorporate risk assessment. Similarly, current recommendations from the American College of Chest Physicians (ACCP) (13) and the National Quality Forum (14) specify that risk assessment guide the decision regarding prophylaxis.

There is a disturbing disconnect between the evidence-based guidelines and the Joint Commission’s current performance measure for VTE prophylaxis (VTE-1) for surgical and hospitalized medical patients, which states, “This measure assesses the number of patients who received VTE prophylaxis or have documentation why no VTE prophylaxis was given the day of or the day after hospital admission” (15). Although very-low-risk patients are excluded (children and patients hospitalized for <2 days), the measure does not consider other patients who are unlikely to benefit from VTE prophylaxis or for whom the risk for harm exceeds the likely benefits. We believe that the measure implicitly encourages prophylaxis for all patients. Many hospitals have implemented it for medical patients as an internal quality measure. We believe that this well-intentioned mandate may be causing some low- to moderate-risk patients to receive prophylaxis, wasting resources, and perhaps harming some patients.

How can we mitigate unintended consequences from performance measurement?

First, it is essential that national performance measures be based on high-quality scientific evidence (16). The process for reviewing the evidence and developing measures must be protected from financial and nonfinancial conflicts of interest. Developers of performance measures and organizations that endorse such measures should use a rigorous system for grading the strength of evidence, such as the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) system (17). In addition, the methods used to develop a measure must be transparent to help stakeholders understand its validity and the expected health benefits of improving performance—merely citing references is not enough. The description of the Joint Commission’s VTE prophylaxis measure references the ACCP guideline; however, as discussed, the ACCP guideline advocates VTE prophylaxis only for high-risk subgroups, whereas the Joint Commission measure includes most patients hospitalized with medical conditions.

Second, measure developers should carefully examine whether the benefits and risks of a test or treatment are homogeneous. Even if studies show that a clinical service is beneficial for an overall study population, the benefit may be large for some subgroups and negligible (or even harmful) for others. If the benefit–risk ratio is minimal or uncertain, use of a therapy should be left to the clinicians’ judgment and not be encouraged through performance measurement. For example, the Joint Commission’s VTE-1 measure could have been restricted to the subgroups of medical patients whom the ACCP designated as having “level A” evidence.
Third, all measures should routinely allow clinicians to report exceptions for not following a recommendation. The Joint Commission added the “diagnostic uncertainty” exception for the PN-5b measure for community-acquired pneumonia after concerns were raised about the measure, and the Joint Commission’s VTE-1 measure can be satisfied if there is a documented reason why no VTE prophylaxis was given. When exception reporting is allowed, studies show that most exceptions reported by clinicians are valid (18, 19). Ideally, exceptions should be recorded in electronic health records so that they can be easily audited to ensure accuracy, ensure that individual physicians or groups do not report very high rates of exceptions, and deter gaming. However, allowing documentation of a reason for not performing a recommendation in an unusual circumstance is not a substitute for accurate, evidence-based specification of the patient subgroups that should be targeted for measurement and quality improvement efforts.

Fourth, we must develop more sophisticated measures that can be programmed into flexible, electronic, clinical decision-support rules. These could take the form of more restrictive eligibility criteria, such as the patient subgroups described for the VTE prophylaxis measure. Corresponding point-of-care alerts or reminders designed to improve performance would only be triggered for patients in the same high-risk subgroups identified in the measure specifications. For some measures, more sophisticated multivariable risk rules can be applied, such as for identifying patients at high risk for cardiovascular events (for example, the Framingham risk score) who should be offered statin therapy. Several risk prediction methods have been proposed for VTE prophylaxis, but none has been endorsed for widespread use at this time. Further research is needed to refine and validate these VTE risk prediction rules, including assessment of the feasibility of implementing them into electronic health record systems.

Finally, to mitigate the risk for inducing unnecessary or harmful care, performance measure developers may need to allow multiple pathways to satisfy a measure. For example, when the Northwestern Medical Faculty Foundation sought to implement a performance measure for controlling low-density lipoprotein cholesterol levels (<2.59 mmol/L [<100 mg/dL]) in patients with diabetes, it recognized that some patients might not achieve this goal even if they were receiving a potent statin. Failure to recognize this in our performance measures and clinical decision-support tools could result in quality improvement efforts that encourage physicians to add additional medications to further reduce low-density lipoprotein cholesterol levels, even though such treatments have not been shown to improve outcomes. Therefore, the group decided that clinicians could satisfy the measure by prescribing a moderate to high dose of a potent statin, even if the low-density lipoprotein cholesterol level remains greater than 2.59 mmol/L (>100 mg/dL) and they used the same logic to suppress the electronic alert.

As we enter the age of personalized medicine and strive for high-value, cost-conscious care, we need flexible performance measures. Failure to develop and faithfully implement these more sophisticated measures will waste resources and harm some patients—the ultimate irony, as the rationale for performance measurement is to foster high-quality, high-value health care.

David W. Baker, MD, MPH
Feinberg School of Medicine at Northwestern University
Chicago, IL 60611

Amir Qaseem, MD, PhD, MHA
American College of Physicians
Philadelphia, PA 19106

for the American College of Physicians’ Performance Measurement Committee

References
10. Yu KT, Wyer PC. Evidence-based emergency medicine: critically appraised topic. Evidence behind the 4-hour rule for initiation of antibiotic therapy in


Current Author Addresses: Dr. Baker: Feinberg School of Medicine at Northwestern University, 750 North Lake Shore Drive, Chicago, IL 60611.
Dr. Qaseem: American College of Physicians, 190 N. Independence Mall West, Philadelphia, PA 19106.