

A Framework for Health Care Organizations to Develop and Evaluate a Safety Scorecard

Peter J. Pronovost, MD, PhD

Sean M. Berenholtz, MD, MHS

Dale M. Needham, MD, PhD

THE DEMAND TO IMPROVE PATIENT SAFETY IS INCREASING within health care organizations. Boards of trustees have a fiduciary responsibility to ensure patient safety, and senior management is often charged with evaluating and improving patient safety. External agencies such as the Centers for Medicare & Medicaid Services (CMS), the Leapfrog Group, and the Joint Commission have developed measures to evaluate patient safety and quality of care.

Many hospitals have responded to this heightened focus on patient safety by creating scorecards to evaluate and publicly report progress in improving quality and safety. Scorecards are attractive because hospital leaders and other interested parties can quickly obtain a broad overview of patient safety performance. These scorecards tend to include measures required by the CMS, The Joint Commission, and insurers, as well as measures developed by individual hospitals for local improvement. Scorecards are increasingly used to evaluate the performance of physicians and senior management.^{1,2}

Despite this increasing interest in scorecards, the science of measuring patient safety is immature. Many health care organizations lack scientifically sound measures to evaluate their progress toward improving patient safety and quality. Measures should be important, scientifically sound (valid and reliable), useful, and feasible.³ A model to differentiate measures that evaluate progress in patient safety and those that identify hazards has been described.^{4,12} This model includes 2 rate-based measures (how often patients experience harm and how often they receive evidence-based interventions) and 2 non-rate-based measures (whether an organization learns from its mistakes and whether it has a culture of safety). Because of substantial measurement and selection bias, measures obtained from patient safety reporting systems should not be used to evaluate progress in patient safety. Rather, these systems help to identify hazards, and the measurement should focus on whether the organization reduced the risk to future patients.

This Commentary presents a potential framework to help health care organizations develop their safety scorecard, evaluate its validity, and understand measures appropriate to present as rates. The term “safety scorecard” is used while ac-

knowledging an overlap between quality and safety. This framework is based on the premise that the goal of the scorecard is to monitor progress in improving patient safety over time or relative to a benchmark. Organizations need to stop conceptualizing safety as a dichotomous variable (ie, safe or unsafe) and start viewing safety as a continuous variable (ie, is it improving?).

Framework

This approach to evaluating safety scorecards is based on models in the Users' Guides to the Medical Literature,⁵ which includes 3 key questions: Are the results important? Are the results valid? Can the results be used to care for patients? Adapting these 3 questions to evaluate safety measures suggests 3 additional key questions: Is the measure important? Is the measure valid? Can the measure be used to improve safety in the health care organization? (BOX)

Is the Measure Important? When selecting a measure for a safety scorecard, health care organizations should ensure that the measure addresses an issue important in their organization. This could be a strategic priority for the organization or a measure required by an external agency such as the Joint Commission.¹¹ Leaders of the organization must prioritize what to focus on measuring and improving. Scorecard measures inherently have a high profile within an organization, and much energy will be devoted to ensuring that the results achieved present a favorable image. Consequently, measures must be worth the concentration of devoted effort.

The scorecard measures selected must be salient to those who will be required to improve them. This requires that leaders select outcome or process measures that represent the interests of frontline health care workers. For example, senior leaders may believe that data on the incidence of deep venous thrombosis can be collected retrospectively, whereas clinicians may question the validity of using discharge data to measure this indicator.⁴ Organization leaders must prioritize among competing demands regarding what to measure and improve and also must allocate sufficient resources to collect and analyze the data.

Author Affiliations: Departments of Anesthesiology and Critical Care Medicine and Surgery (Drs Pronovost and Berenholtz) and Pulmonary and Critical Care Medicine (Dr Needham), School of Medicine; and Department of Health Policy and Management, Bloomberg School of Public Health (Dr Pronovost), Johns Hopkins University, Baltimore, Maryland.

Corresponding Author: Peter J. Pronovost, MD, PhD, Johns Hopkins Medical Institution, 1909 Thames St, Second Floor, Baltimore, MD 21231 (ppronovo@jhmi.edu).

Box. Worksheet to Evaluate a Patient Safety Scorecard**Is the Measure Important?**

1. Does the measure address a strategic priority for the organization?
Does improved performance on the measure correlate with improved outcomes?
2. Is the measure required by an external group or agency?

Is the Measure Valid?

1. Is the measure supported by empirical evidence or a consensus of experts?
2. Does the measure have face validity—ie, do clinicians believe that improvement in performance on the measure will be associated with improved patient outcomes or, for performance measures, is the outcome preventable?
3. Is the risk for selection bias minimized?
Are explicit inclusion and exclusion criteria provided for the patient population being measured?
Is a similar patient population included in the measure during each period to allow comparability over time?
4. Is the risk for measurement bias minimized?
Is an explicit definition provided for the event (numerator)?
Is an explicit definition provided for those at risk for the event (denominator)?
Is a surveillance system in place to identify both the event (numerator) and those at risk for the event (denominator)?
Is a standardized data collection form(s) used with data collectors trained in appropriate use of the form(s)?
Is there an explicit plan for review of data quality during data collection?
Is there a plan to minimize, and report on, missing data for events (numerator) and those at risk for events (denominator)?

5. Is the risk for analytic bias minimized?

Are appropriate statistical methods used to estimate changes in performance over time?
Are differences in the patient population over time accounted for?
Are historical trends in performance accounted for?
Is clustering of events within a group(s) accounted for?
Are estimates of performance or changes in performance presented in a format that is meaningful for those who will use the data?
Are estimates of performance or changes in performance presented with an estimate of precision (eg, confidence interval)?
Have all potential biases in the measure been reported in a transparent manner?

Can This Measure Be Used to Improve Safety in the Organization?

1. Where does the measure fit within the organizational priorities?
2. Is valid and reliable data collection feasible within the organization?
Is there adequate infrastructure to collect the data completely, accurately, and reliably?
What does it cost (including people, time, and technology costs) to collect the data required for this measure?
Are there other measures that should be foregone to allocate resources for this measure?
3. Do the benefits of knowing the information provided by this performance measure outweigh the costs of data collection?
4. Will performance on this measure help focus quality improvement efforts?

Is the Measure Valid? To evaluate validity, the health care organization should assess the (1) level of evidence supporting the measure, (2) face validity of the measure, and (3) risk for bias. A valid safety measure has strong empirical evidence demonstrating that the intervention improves safety and also concisely evaluates the effectiveness of the intervention on its intended outcome. The validity of a given measure depends on context—thus, a measure may be valid for one patient safety issue on one unit but not valid for a different patient safety issue or unit. Consider a safety measure evaluating smoking cessation counseling after acute myocardial infarction. An organization could evaluate the evidence linking smoking cessation to improved patient outcomes after myocardial infarction and whether the measure effectively evaluates the appropriate use of that evidence. Yet, organizations frequently measure whether a nurse documented the provision of counseling to the patient. However, there is little evidence that this method of measur-

ing the intervention will accurately evaluate the patient's understanding of smoking risks or behavior change. In one study, reports of increased counseling for smoking cessation did not correlate with reduced patient mortality.⁶

The organization should also evaluate if the measure has face validity. Relative to the scorecard, organizations can evaluate face validity by asking staff members who will use the data if they believe improving performance on the measure (as it is defined and assessed) will improve patient outcomes or, for performance measures, the extent to which the outcome is preventable. If they believe the measure lacks face validity, they likely will not use it. For example, one health care organization measured readmission to an intensive care unit (ICU) within 30 days of discharge as an outcome measure of safety (P.J.P., unpublished data, August 27, 2007). The ICU physicians felt this measure did not reflect their performance because most patients readmitted

within 30 days had developed new medical problems unrelated to their earlier care in the ICU.

An organization should also evaluate the influence bias can have on a measure, particularly when assessing the hospital's performance over time or when benchmarking to peer hospitals. Three potential sources of bias could introduce error into a measure: selection bias, measurement bias, and analytic bias.

Selection bias occurs when patients included in the measurement group vary in their risk for the event. To evaluate selection bias, the organization should use explicit and consistent criteria to define which patients should be included in each period under comparison. Also, the organization should evaluate whether the population being measured has similar risk for the outcome. This can be performed with varying degrees of rigor, from evaluating whether a different patient population is included in the comparison group(s) to case-mix adjustment based on administrative or detailed clinical data.

Measurement bias represents systematic error introduced in the data collection process, as distinct from random error, which is due to small sample sizes. Less rigorous data collection limits the ability to rigorously evaluate progress in patient safety. To minimize measurement bias, each measure should have an explicit and clear definition for the event (numerator) and for those at risk of the event (denominator). This requires a robust surveillance system, including a standardized data collection form, staff training for appropriate use of the form, and a systematic process to review the quality of data collection. In addition, a plan is needed to minimize and report missing data, which often exceed 60% of the available data and limits the ability to make accurate inferences about improvements in safety.⁷

Analytic bias is introduced when inappropriate methods are used to evaluate data. When evaluating progress in patient safety, it is important to perform appropriate statistical tests to analyze changes in performance on measures over time. Also, the analysis should consider variation over time regarding the risk for the event measured and historic trends in performance. If the safety measure reports aggregate data from multiple units in a hospital or multiple hospitals within a system, the analysis should also account for "clustering" (nonindependence of outcomes) within each unit or hospital.^{8,9} Inferences made about changes in performance, and potential biases, should be presented clearly for easy comprehension. This presentation will typically include some point estimate of the magnitude of change and a measure of precision around that estimate (eg, confidence interval).

Can the Measure Be Used to Improve Safety in the Health Care Organization? Assuming the measure is important and valid, an additional important question is whether the organization can use the measure to guide its quality improvement efforts. Collecting data for safety requires commitment of scarce resources. The organization should determine if an adequate infrastructure is in place to collect complete, accurate, and reliable data; explicitly define the resources and costs required to collect complete and accurate data for the measure with minimal bias; and decide if the benefits are worth

the costs. While formal quantitative models are available to help with this decision,¹⁰ a qualitative approach is more commonly used in which the candidate measure is discussed among relevant internal stakeholders and the relative costs and benefits are assessed. Also, organizations should discontinue use of existing measures with little benefit to concentrate resources on the most important measures.

Data should be collected only if the measure will be used to guide improvement efforts. For example, if an organization decides to measure infection rates, it should implement interventions to improve these rates. Without a plan to improve performance, the value of the measure and the dedication of scarce resources to collect data should be reconsidered.

Conclusions

The need to track progress in patient safety is increasing. Measures to evaluate progress must be important, valid, and used to guide improvements in patient safety. The framework proposed in this article can help health care organizations more effectively and efficiently develop and evaluate their safety scorecards to better address the question, "are patients safer?"

Financial Disclosures: Dr Pronovost reported receiving grants from the Michigan Health and Hospital Association, the New Jersey Hospital Association, Rhode Island Quality Institute, and MCIC Inc (a liability insurer) and receiving honoraria from various hospitals to speak about quality and safety. Dr Berenholtz reported receiving consulting fees from VHA Inc. Dr Needham reported no disclosures.

Funding/Support: Dr Needham holds a Clinician-Scientist Award from the Canadian Institutes of Health Research (CIHR). Dr Berenholtz holds a Clinician Scientist Award (K23) from the National Institutes of Health/National Heart, Lung, and Blood Institute. **Additional Contributions:** We thank the following individuals for their assistance in the preparation of the manuscript: Christine G. Holzmueller, BLA (Johns Hopkins University), editing; Christine A. Goeschel, RN, MPA, MPS (Johns Hopkins University), design and drafting; Laura Morlock, PhD, and Albert W. Wu, MD (Johns Hopkins University), conception and revision; and Marlene Miller, MD, MSc (National Association of Children's Hospitals and Related Institutions), revision. None of those acknowledged received extra compensation for their contributions.

REFERENCES

- Zelman WN, Pink GH, Matthias CB. Use of the balanced scorecard in health care. *J Health Care Finance*. 2003;29(4):1-16.
- Lindenauer PK, Remus D, Roman S, et al. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med*. 2007;356(5):486-496.
- McGlynn EA. Selecting common measures of quality and system performance. *Med Care*. 2003;41(1)(suppl):139-147.
- Pronovost PJ, Miller MR, Wachter RM. Tracking progress in patient safety: an elusive target. *JAMA*. 2006;296(6):696-699.
- McAlister FA, Straus SE, Guyatt GH, Haynes RB; Evidence-Based Working Group. Users' Guides to the Medical Literature. XX: integrating research evidence with the care of the individual patient. *JAMA*. 2000;283(21):2829-2836.
- Bradley EH, Herrin J, Elbel B, et al. Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality. *JAMA*. 2006;296(1):72-78.
- Stroup DF, Berlin JA, Morton SC, et al; Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA*. 2000;283(15):2008-2012.
- Rabe-Hesketh S, Skrondal A. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press; 2005.
- Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC; 2004.
- Detsky AS, Laupacis A. Relevance of cost-effectiveness analysis to clinicians and policy makers. *JAMA*. 2007;298(2):221-224.
- Performance measurement initiatives. The Joint Commission Web site. <http://www.jointcommission.org/PerformanceMeasurement/PerformanceMeasurement/default.htm>. Accessibility verified September 26, 2007.
- Zhan C, Battles J, Chiang YP, Hunt D. The validity of ICD-9-CM codes in identifying postoperative deep vein thrombosis and pulmonary embolism. *Jt Comm J Qual Patient Saf*. 2007;33(6):326-331.