Is there a ‘best measure’ of patient safety?

Ann M Borzecki,1,2,3 Amy K Rosen1,4,5

Despite consensus that preventing patient safety events is important, measurement of safety events remains challenging. This is, in part, because they occur relatively infrequently and are not always preventable. There is also no consensus on the ‘best way’ or the ‘best measure’ of patient safety. The purpose of all safety measures is to improve care and prevent safety events; this can be achieved by different means. If the overall goal of measuring patient safety is to capture the universe of safety events that occur, then broader measures encompassing large populations, such as those based on administrative data, may be preferable. Acknowledging the trade-off between comprehensiveness and accuracy, such measures may be better suited for surveillance and quality improvement (QI), rather than public reporting/reimbursement. Conversely, using measures for public reporting and pay-for-performance requires more narrowly focused measures that favour accuracy over comprehensiveness, such as those with restricted denominators or those based on medical record review.

There are at least two well-established patient safety measurement systems available for use in the inpatient setting, namely the administrative data-based Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicators (PSIs) and the medical record-based National Surgical Quality Improvement Programme (NSQIP) measures.1-3 The AHRQ PSIs, publicly released in 2003, are evidence-based measures designed to screen for potentially preventable medical and surgical complications that occur in the acute care setting. Since they use administrative data, they were originally designed as tools for use in case finding for local QI efforts and surveillance, as well as for internal hospital comparisons. They were developed using a rigorous process beginning with a thorough review of the literature for existing administrative data-based indicators, review by clinical expert panels, consultation with coding experts and empirical analyses to assess the statistical properties of the measures, such as reliability and predictive and construct validity. They were intentionally developed to favour specificity over sensitivity. As such, the indicators have various exclusion criteria designed to decrease the likelihood of including patients in whom a complication is very unlikely to be preventable, as well as having a well-specified denominator. (Prior to the inclusion of Present on Admission (POA) coding in US claims datasets in late 2007, patients with various secondary diagnoses were also excluded.)4 Additionally, to better compare hospitals and reflect the fact that some patients are at higher risk for a complication than others, risk adjustment is used to compare indicator rates.

The ‘flagship’ medical record or chart-based system, NSQIP, was established by the US Veterans Health Administration (VA) in 1994 over concerns of higher mortality rates and substandard surgical care in the VA. It was designed to promote QI of VA surgical care by providing reliable, valid and comparative information regarding 30-day surgical outcomes, such as morbidity and mortality, across all facilities performing major non-cardiac surgery.1 Trained nurse reviewers prospectively gather medical record information from a select sample of all eligible operations.1 The programme’s success led to establishment and launch of a similar programme in the non-VA setting by the American College of Surgery (ACS) in 2004, known as ACS-NSQIP.5 Use of this programme has been expanding; it is currently being implemented in at least nine countries, including Canada, for benchmarking and QI purposes.5
In the current issue, McIsaac et al examined the accuracy of a new set of administrative data-based PSIs developed using Canadian International Classification of Diseases (ICD)-10 coded data.6 The ‘new’ PSIs identify complications of care that arise after admission using a diagnosis timing variable (‘diagnosis type’) present in the administrative data which indicates whether the diagnosis is pre-existing or occurred after admission (potentially representing a complication of care). They were designed to improve on some of the recognised limitations of the AHRQ PSIs. Namely, they were designed to be more comprehensive (covering more complications) and applicable to a larger population of patients (they do not exclude populations at higher risk of a complication) than the AHRQ PSIs.6 Unlike the AHRQ PSIs, they do not specifically try to account for the potential preventability of an event; instead, they look for administrative data-based codes that may represent suboptimal quality or unsafe care. They also use a global denominator but they can be applied to a particular population of interest. Additionally, while the AHRQ PSIs are risk-adjusted, the new PSIs are not; only observed rates are calculated. Finally, the new PSIs were created through a slightly different process than the AHRQ PSIs.6 Rather than starting with the literature in the area, the developers identified all administrative codes representing conditions arising after admission.6 These individual codes were then rated by patient safety experts with respect to their likelihood of being related to a patient safety event and then grouped into categories (eg, hospital-acquired infections). These categories were not mutually exclusive such that codes could be assigned to more than one category.

McIsaac et al compared the accuracy of events identified by the new PSIs to those identified by NSQIP (considered the ‘gold standard’).7 Only 7 of 17 PSI categories mapped to a specific NSQIP complication and in some cases, a given NSQIP complication was mapped to more than one PSI. Of the overlapping complication categories, overall, they found low to moderate positive predictive values (PPVs) and sensitivities, and high specificities and negative predictive values of the new PSIs compared with NSQIP. Complications with the worst agreement included those related to fluid management and respiratory issues. With respect to the specific comparisons, among the overlapping categories, they did not report which specific events the PSIs identified that NSQIP did not. However, of the non-overlapping categories, the PSIs picked up several gastrointestinal complications while NSQIP identified events such as wound disruption in some cases, a given NSQIP complication was mapped to more than one PSI. Of the overlapping complication categories, overall, they found low to moderate positive predictive values (PPVs) and sensitivities, and high specificities and negative predictive values of the new PSIs compared with NSQIP. Complications with the worst agreement included those related to fluid management and respiratory issues. With respect to the specific comparisons, among the overlapping categories, they did not report which specific events the PSIs identified that NSQIP did not. However, of the non-overlapping categories, the PSIs picked up several gastrointestinal complications while NSQIP identified events such as wound disruption which the new PSIs did not.

As the authors note, prior investigators have compared the AHRQ PSIs to NSQIP data and similarly found generally low sensitivities and low to moderate PPVs.8–10 These two previous studies used the AHRQ PSI measures based on ICD-9 codes; Romano et al lacked POA data while Cima et al incorporated POA data into their study.8,9 Although we are not aware of any similar validation studies using the ICD-10 version of the AHRQ PSIs (ICD-10 is considered to be a more specific diagnosis-based coding system than ICD-9), we know from this prior work that an important reason for the seemingly low accuracy of the PSIs has to do with fundamental differences in definitions of medical record-based versus administrative-based safety events.6–11 Despite overlap conceptually, PSIs (both the new and old) and NSQIP measures have different definitions that reflect the method of development as well as the data sources used. For example, the AHRQ PSI perioperative haemorrhage or haematoma (previously known as ‘postoperative’) gets mapped to the fairly specific NSQIP complication of postoperative bleeding requiring transfusion of greater than or equal to 4 units of blood. In the new PSI set, ‘haemorrhage’ gets mapped to this complication along with postoperative stroke which can be haemorrhagic or ischaemic.12 Despite the difficulty in mapping NSQIP complications to PSI events, even in the validation work done by both AHRQ and our VA group, in which we used medical records as the ‘gold standard’ to validate diagnosis codes, coding system limitations (especially lack of POA codes and coding specificity, eg, many codes associated with complications were not specific with respect to timing and could be used for preoperative or postoperative events) as well as coding errors, accounted for many of the false positives; similar problems contributed to a large percentage of false negatives.13–15 Although our groups considered relatively few cases to represent documentation errors, other researchers have found that documentation quality affects coding accuracy.16 Notably, these prior PSI validation studies led to modifications of the ICD-9 coding system to improve the specificity of codes relevant to several of the AHRQ PSIs. A more recent study of the AHRQ PSI Postoperative Deep Vein Thrombosis and Pulmonary Embolism found that inclusion of POA data and the presence of more specific ICD-9 codes resulted in improved PPVs, from 43%–48% to 81%–99%.17 Although no US studies have examined the validity of the ICD-10 version of the AHRQ PSIs, one study using an ICD-10 based international version of the AHRQ PSIs with POA coding examined five PSIs and found fairly high PPVs for four of them (62.5%–86%).18 How does the study by McIsaac et al contribute to the existing literature? We think an important contribution of this study is that it demonstrates that even with both the more specific ICD-10 based measurement system and the equivalent of POA coding (diagnosis timing), the new PSI measures still suffer some of the same limitations as the AHRQ PSIs that originally used an ICD-9-based system without POA (which was added later to improve criterion validity). While this new PSI system also has the potential advantage of enhancing our ability to measure the universe of safety
events (ie, the new PSIs are broader with respect to the numerator and denominator compared with AHRQs, thereby identifying more events), without knowing more about the specifics of the events, it is hard to determine whether they are true positives and/or if they represent potentially preventable events. Use of a global denominator may capture many events that are not preventable due to patient-related or procedure-related factors. Furthermore, only three countries currently have datasets with timing diagnoses, so international use of such a system is relatively limited.

We were surprised at the high rate of complications identified by both the new PSIs and NSQIP (18.7% and 22%, respectively). Although this does not negate the authors’ findings, these rates are higher than those of other studies. For example, Cima et al reported a complication rate of 7.4% using NSQIP data at one US hospital; Mull et al found a rate of 6% in a national VA sample. The current study’s findings are based on one hospital network comprising two hospitals. Presumably, the accuracy of administrative data may vary by institution (as well as condition). The authors only examined surgical patients; would the new PSIs perform better or worse in medical patients?

So is there a best measure of patient safety? All measures have their strengths and weaknesses. Although the AHRQ PSIs were originally designed for QI and surveillance, in the USA, they have been increasingly used for federal and state public reporting for QI and surveillance, in the USA, they have been used, rather than on public reporting or financial reimbursements, potentially punitive actions that may not be equitable across hospitals. Furthermore, since they are focused only on inpatient safety events, neither the new PSIs, AHRQ PSIs or NSQIP allow us to truly capture the universe of safety events. Given that most care is now delivered in the outpatient setting, the field of patient safety measurement needs to expand to capture this setting as well. Continuing to ‘reinvent’ the wheel with development of new inpatient PSIs—that in the end have some of the same limitations as the older PSI measures—is a commendable journey, but one not likely to significantly advance the patient safety field. It is time to take the road less travelled.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES


