Commentary on Sharek: Adverse Events and Errors—Important to Differentiate and Difficult to Measure

Timothy P. Hofer

The paper by Sharek and colleagues in this issue of *Health Services Research* describes some of the characteristics of a method of measuring adverse events in hospitalized patients. The method was developed by the Institute for Healthcare Improvement (IHI) and is referred to in the paper by Sharek and colleagues as the IHI global trigger tool. This paper presents a large and well-designed reliability study and as such is an important contribution to the field of patient safety.

Reliability quantifies the reproducibility of a measurement technique, which is a key piece of information. For example, in this case, it would allow one to calculate how many observations per facility are necessary to generate an acceptably precise estimate of its adverse event rate. This paper does not provide evidence about validity, another measurement characteristic that describes how well the measurement captures the underlying construct.

An important point is that the reliability estimate presented in the Sharek and colleagues' study is not a characteristic of the IHI trigger tool itself. Reliability is a function of the entire measurement process, which includes the precision of the instrument, the measurement procedure or method of applying the instrument, and the amount of variability of the true adverse event rate in the underlying population. In order to infer that one will obtain the same level of reproducibility as is reported in the Sharek and colleagues' paper when using the instrument in another setting, one has to use the instrument with a similar measurement procedure and population. The broadly representative sample of hospitals included in the study by Sharek and colleagues is a strength that makes the reliability estimates more generalizable.

However, the measurement procedure described for the use of the IHI trigger tool is complex and involves several people and a sequence of steps. First, a nonphysician primary reviewer examines a hospital record looking for

specific trigger events and labels them as adverse events or not. Second, the primary reviewer and two physician reviewers discuss the events as a group. Third, the physicians each independently review the record and subsequently meet again and reach a consensus on any events about which they disagree. Much of the research using physician implicit review to measure adverse events and quality has used a similarly complex process, and this complexity is more traditional than evidence based. This particular measurement procedure is almost identical to that used in the Harvard Medical Practice study (Brennan, Localio, and Laird 1989).

Thus, the key reliability estimates in the paper are those that represent the end result of the measurement procedure. The results suggest a κ of between 0.3 and 0.5 for measuring the presence, number, and severity of adverse events. This is reassuringly similar to the estimates from almost all earlier studies where a κ of 0.4–0.6 is routinely reported when assessing whether an adverse event occurred (Brennan et al. 1991; Localio et al. 1996; Thomas et al. 2002).

Sharek and colleagues also report considerably higher κ s ranging from 0.4 to 0.9 (presented in Figure 2 in the article) that relate to the agreement between the two physician reviewers on the same team after they have already jointly discussed the adverse events with the primary reviewer. This is a much less relevant piece of information, given how the measurement procedure was designed. The findings of enhanced agreement within a team of reviewers after a group discussion is most likely illusory and merely a product of the lack of independence of the individual physician reviews (Gigone and Hastie 1993; Hofer et al. 2000). In fact in most cases, it is likely to be more efficient to conduct completely independent reviews by two physician members of a team and average the results if the goal is to improve the reliability of a measurement procedure for a single observation.

On a more technical level, I would argue that, while commonly used, the κ statistic is not the best measure to quantify reliability for physician judgments of the quality and safety of health care. The κ when applied to dichotomous ratings is affected by the base event rate in a way that is undesirable (Agresti, Ghosh, and Bini 1995). Furthermore, it is difficult to adjust or stratify a κ by covariates. These issues are addressed by other measures of agreement such as an intraclass correlation (Lilford et al. 2007). When, as in the paper by Sharek

Address correspondence to Timothy Hofer, M.D., VA Center for Practice Management and Outcomes Research, P.O. Box 13017, Ann Arbor, MI 48113-0170; e-mail: thofer@umich.edu. Timothy Hofer, M.D., is with the Division of General Internal Medicine, University of Michigan and the VA Center for Practice Management and Outcomes Research, Ann Arbor, MI.

and colleagues, a weighted κ is used for ordinal rating categories with Fleiss–Cohen κ weights, then the κ statistic becomes an intraclass correlation coefficient (Dunn 2004, p. 151). Of course, while this correspondence could be an argument for the continued use of the κ statistic, others have pointed out that “(1) these conditions are not always met, and (2) one could instead directly calculate the intraclass correlation” (Uebersax 2010).

There are also some important things to be pointed out about measuring adverse events. First and foremost, adverse events are not errors, although people conflate the two concepts. Despite attempts to clarify these important distinctions (Hofer, Kerr, and Hayward 2000), the terminology used in public policy discussions of patient safety and medical errors continues to elide the distinction between adverse events and errors.

Most of our definitions in this area, like the methods of measurement, are bequeathed to us by the Harvard Medical Practice study (Brennan et al. 1991) and its offshoots, the studies that were used by the Institute of Medicine (IOM) to derive the widely quoted estimates of the number of deaths from medical errors (Kohn, Corrigan, and Donaldson 1999). As defined and measured in those studies, an adverse event is a bad outcome caused by medical care. More recently, an Office of the Inspector General (OIG) report has broadened this definition to its logical limit, including all harm that occurs in a health care setting (Levinson 2010). Adverse events are not necessarily preventable or a result of substandard care, and in many cases they are an unavoidable risk of treatment. As outcomes, adverse events are of varying severity with many of them mild or even trivial, and conversely many treatments associated with frequent adverse events can save lives or prevent suffering and are often underutilized. Therefore, too much attention to adverse events without adequate attention to their severity or whether they are preventable can be a threat to patient safety (Lilford et al. 2003; Hayward et al. 2005).

In the Harvard Medical Practice study, which was designed to study tort reform, a second evaluation was done of the identified adverse events to judge whether they were negligent or due to substandard care. In subsequent publications, once there was more interest in identifying medical errors, the same set of adverse events were reevaluated in an attempt to label them as preventable or not. A preventable adverse event was called an error, although this really represented a subset of the IOM’s much broader definition of medical error “the failure of a planned action to be completed as intended (error of execution) or the use of a wrong plan to achieve an aim (error of planning).” The IOM definition focuses exclusively on process problems without considering whether an adverse event occurred as a result.

Why all this discussion of errors when the Sharek and colleagues' article describes the reliability of an instrument to measure adverse events? Because at the end of the day, the only reason to measure adverse events is to identify the subset of adverse events that are preventable by better care. By definition, if a bad outcome occurs despite the right things being done (meaning that the treatment was appropriate and risks were appropriately minimized), there is really not very much to do about it, and very little to be gained by the expense of monitoring and examining these events. In fact, monitoring adverse events from interventions without regard to the proper denominator (the number of appropriate interventions performed) simply discourages the use of any intervention with a substantial adverse event risk without regard for the intervention's net benefit produced (Lilford et al. 2003; Hayward et al. 2005).

The rub is that it is much harder to determine whether an adverse event was due to substandard care than it is to tell if an adverse event occurred. Well-done reliability studies that provide estimates for determining whether events are negligent or preventable have κ s in the 0.2–0.4 range (Dubois and Brook 1988; Brennan, Localio, and Laird 1989; Brennan et al. 1991; Goldman 1992; Hayward and Hofer 2001; Lilford et al. 2007). The paper by Sharek and colleagues does not report the reliability of whether the adverse event was preventable using the specified measurement procedure. They do provide a κ for the within team agreement of preventability, but as mentioned above, this does not reflect the reliability of the measurement procedure and is likely to be inflated. In assessing any statements about the preventability of the adverse events identified in this study, it would be important to know the reliability of the measurement of preventability using the measurement procedure.

It is gratifying to see investigators put the effort into performing a well-designed reliability study as part of a project that attempts to measure complex and intangible constructs such as adverse events, and safety and quality of care. After all, many of the most important things to measure about health care and life in general are not something that we can measure directly with a ruler or laboratory instrument. Carefully specifying a measurement procedure and testing its ability to reproducibly detect a signal is an essential and frequently neglected step. Reliability defines the precision of measurement and tells us the extent to which measurement error will undermine an effort to distinguish true differences and relationships within a specified population. The authors are to be commended for taking the first step of quantifying the reliability of measuring adverse events with the IHI trigger tool. Unfortunately our ability to reliably identify the events of greatest importance, major adverse events resulting from poor quality, remains elusive and it is still unclear how much the

IHI trigger tool will allow us to do that. Whether the measurement challenge is due to these events being difficult to identify or because they are rare occurrences is a question of great importance.

REFERENCES

- Agresti, A., A. Ghosh, and M. Bini. 1995. "Raking Kappa: Describing Potential Impact of Marginal Distributions on Measures of Agreement." *Biometrical Journal* 37 (7): 811–20.
- Brennan, T., L. Leape, N. Laird, L. Hebert, A. R. Localio, A. G. Lawthers, J. P. Newhouse, P. C. Weiler, and H. H. Hiatt. 1991. "Incidence of Adverse Events and Negligence in Hospitalized Patients. Results of the Harvard Medical Practice Study I." *New England Journal of Medicine* 324 (6): 370–6.
- Brennan, T. A., R. J. Localio, and N. L. Laird. 1989. "Reliability and Validity of Judgments Concerning Adverse Events Suffered by Hospitalized Patients." *Medical Care* 27 (12): 1148–58.
- Dubois, R. W., and R. H. Brook. 1988. "Preventable Deaths: Who, How Often, and Why?" *Annals of Internal Medicine* 109 (7): 582–9.
- Dunn, G. 2004. *Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies*. 2nd Edition. London: Arnold.
- Gigone, D., and R. Hastie. 1993. "The Common Knowledge Effect: Information Sharing and Group Judgment." *Journal of Personality and Social Psychology* 65: 959–74.
- Goldman, R. L. 1992. "The Reliability of Peer Assessments of Quality of Care." *Journal of the American Medical Association* 267 (7): 958–60.
- Hayward, R. A., S. M. Asch, M. M. Hogan, T. P. Hofer, and E. A. Kerr. 2005. "Sins of Omission: Getting Too Little Medical Care May Be the Greatest Threat to Patient Safety." *Journal of General Internal Medicine* 20 (8): 686–91.
- Hayward, R. A., and T. P. Hofer. 2001. "Estimating Hospital Deaths Due to Medical Errors: Preventability Is in the Eye of the Reviewer." *Journal of the American Medical Association* 286 (4): 415–20.
- Hofer, T. P., S. J. Bernstein, S. DeMonner, and R. A. Hayward. 2000. "Discussion between Reviewers Does Not Improve Reliability of Peer Review of Hospital Quality." *Medical Care* 38 (2): 152–61.
- Hofer, T. P., E. A. Kerr, and R. A. Hayward. 2000. "What Is An Error?" *Effective Clinical Practice* 3 (6): 261–9.
- Kohn, L. T., J. Corrigan, and M. S. Donaldson. 1999. *Institute of Medicine (U.S.), Committee on Quality of Health Care in America. To Err Is Human; Building a Safer Health System*. Washington, DC: National Academy Press.
- Levinson, D. 2010. "Adverse Events in Hospitals: Methods for Identifying Events." Office of the Inspector General [accessed on July 16, 2010]. Available at <http://www.oig.hhs.gov/oei/reports/oei-06-08-00221.pdf>

- Lilford, R. J., M. A. Mohammed, D. Braunholtz, and T. P. Hofer. 2003. "The Measurement of Active Errors: Methodological Issues." *Quality and Safety in Health Care* 12 (suppl 2): ii8–12.
- Lilford, R., A. Edwards, A. Girling, T. Hofer, G. L. Di Tanna, J. Petty, and J. Nicholl. 2007. "Inter-Rater Reliability of Case-Note Audit: A Systematic Review." *Journal of Health Services Research and Policy* 12 (3): 173–80.
- Localio, A. R., S. L. Weaver, J. R. Landis, A. G. Lawthers, T. A. Brenhan, L. Hebert, and T. J. Sharp. 1996. "Identifying Adverse Events Caused by Medical Care: Degree of Physician Agreement in a Retrospective Chart Review." *Annals of Internal Medicine* 125 (6): 457–64.
- Thomas, E. J., S. R. Lipsitz, D. M. Studdert, and T. A. Brennan. 2002. "The Reliability of Medical Record Review for Estimating Adverse Event Rates." *Annals of Internal Medicine* 136 (11): 812–6.
- Uebersax, J. 2010. "Kappa Coefficients: A Critical Appraisal" [accessed on July 16, 2010]. Available at <http://www.john-uebersax.com/stat/kappa.htm>.