

The Reliability of Medical Record Review for Estimating Adverse Event Rates

Eric J. Thomas, MD, MPH; David M. Studdert, LLB, ScD, MPH; and Troyen A. Brennan, MD, JD, MPH

Background: The data used by the U.S. Institute of Medicine to estimate deaths from medical errors come from a study that relied on nurse and physician reviews of medical records to detect the errors.

Objective: To measure the reliability of medical record review for detecting adverse events and negligent adverse events.

Design: Medical record review.

Setting: Hospitalizations in Utah and Colorado in 1992.

Measurements: After three independent reviews of 500 medical records, the following were measured: reliability and the effect of varying criteria for reviewer confidence in and reviewer agreement about the presence of adverse events.

Results: For agreements in judgments of adverse events among the three sets of reviews, the κ statistics ranged from 0.40 to 0.41 (95% CIs ranged from 0.30 to 0.51) for adverse events and from 0.19 to 0.23 (CIs, 0.05 to 0.37) for negligent adverse events. Rates for adverse events and for negligent adverse events varied substantially depending on the degree of agreement and the level of confidence that was required among reviewers.

Conclusion: Estimates of adverse event rates from medical record review, including those reported by the Institute of Medicine in its 2000 report on medical errors, are highly sensitive to the degree of consensus and confidence among reviewers.

Ann Intern Med. 2002;136:812-816.

www.annals.org

For author affiliations, see end of text.

Implicit judgments, based on medical record review, about adverse events caused by medical care have moderate to poor inter-rater reliability (1–7). However, the influential report on medical errors issued by the U.S. Institute of Medicine (IOM) in 2000 (8) relied on studies based on medical record review to estimate that 44 000 (9) to 98 000 (10) Americans die each year because of medical errors. These studies used one (9) or up to three (10) physician reviewers per record to detect adverse events.

The IOM estimates are important because they have prompted health care providers and administrators to reduce errors and have influenced the U.S. research budget. We report details on the reliability of physician judgments about adverse events and negligent adverse events in the study from which the IOM derived its estimate of 44 000 error-related deaths per year (9). We also report the effects of varying criteria for reviewer confidence in and reviewer agreement about the presence of adverse events.

METHODS

Data Sources

This study was conducted as part of the Utah and Colorado Medical Practice Study (UCMPS) (5). In UCMPS, trained nurses reviewed medical records to

identify 1 of 19 screening criteria that could indicate the presence of an adverse event. A trained physician then reviewed flagged records by using a structured chart abstraction form. Previously, the nurse review process was found to be a good screening tool, with a sensitivity of 84% (1).

We began the current study with all 2868 records referred for physician review by nurse screeners from the original 15 000-record sample (5000 from Utah and 10 000 from Colorado) (**Appendix Figure**). Two physician investigators confirmed all adverse events detected during this initial review. As a result of this confirmation process, we eliminated 13 false-positive cases. We eliminated cases only if they failed to meet explicit criteria for an adverse event, not if there was a concern about the reviewer's judgment regarding whether medical management caused the adverse event.

Next, we randomly selected 500 of the 2868 records referred for physician review by nurse screeners (167 from Utah and 333 from Colorado), maintaining the original 1 to 2 ratio of Utah to Colorado records (**Appendix Figure**). The general characteristics of this sample are described elsewhere (5). Of these 500 records, 400 were randomly sampled from referred records that a single physician reviewer had previously judged to show no adverse events. Fifty records were randomly sampled from records initially found to show nonnegligent ad-

verse events, and the remaining 50 were sampled from records initially judged to show negligent adverse events. This mix provided a sample similar to the one originally reviewed by the physicians.

Record Review

We completed three independent physician reviews of the original 500 medical records (not photocopies). The first was the original review for the UCMPS. The two subsequent physician reviews were conducted by physicians from the original UCMPS and by newly recruited physician reviewers who were trained in the same manner as the original UCMPS physician reviewers (5). Physicians could participate in more than one review; however, all reviewers were blinded to the purpose of these additional reviews, and none of them reviewed a record that they had previously reviewed. All physician reviewers used the same data form, which included the same definition of *adverse event*: an injury caused by medical management (rather than the disease process) that resulted in prolonged hospital stay or disability at discharge. *Negligence* was defined as care that fell below the standard expected of physicians in the community.

Because judgments about adverse events may be complex, reviewers used a six-point confidence scale that has been used in previous studies (5, 6). We required a confidence score of at least four (>50% chance that medical management caused the adverse event) to indicate the presence of adverse events or negligent adverse events. For each of the three reviews, two investigators independently reviewed all detected adverse events to confirm that they met the study criteria.

Table 1. Number of Cases Resulting in Agreement and Disagreement among Three Sets of Reviews

Adverse Event Detected?			Frequency, n (%)
Review 1	Review 2	Review 3	
No	No	No	321 (67.6)
No	No	Yes	34 (7.2)
No	Yes	No	37 (7.8)
No	Yes	Yes	19 (4.0)
Yes	No	No	17 (3.6)
Yes	No	Yes	8 (1.7)
Yes	Yes	No	14 (3.0)
Yes	Yes	Yes	25 (5.3)

Context

The Institute of Medicine report that caused concern about adverse events in hospitals relied on studies that used medical record review to identify adverse events.

Many people question whether medical record reviews can accurately identify negligent adverse events.

Contribution

The following changes in the review process markedly reduced the rates of negligent adverse events: 1) increasing the number of reviewers from one to three and 2) requiring reviewers to be highly confident that an event was due to negligence.

Implications

Because review criteria can affect adverse event rates, the estimates cited by the Institute of Medicine could be inaccurate.

—The Editors

Statistical Analysis

We calculated the rates of adverse events and negligent adverse events detected during each review. We report κ statistics for adverse events and negligent adverse events as the measure of inter-rater reliability. For all statistical analyses, we used SAS software, release 6.12 (SAS Institute, Inc., Cary, North Carolina).

Role of the Funding Source

The funding source contributed to the design of the study but had no role in conducting the study or in reporting its results.

RESULTS

The first independent review of the study sample detected 64 adverse events; the second, 95; and the third, 86. Table 1 shows the agreement and disagreement among the three reviews. Comparisons between sets of reviews (review 1 compared with review 2, review 1 compared with review 3, and review 2 compared with review 3) demonstrated similar reliability. The κ statistics ranged from 0.40 to 0.41 (the lowest bound of the three 95% CIs was 0.30, and the highest was 0.51) for adverse events and from 0.19 to 0.24 (the lowest bound of the 95% CIs was 0.05, and the highest was 0.37) for negligent adverse events. The former is considered

Table 2. Adverse Event Rates Using Different Thresholds

Reviews in Agreement	Confidence Score of 2 or Higher		Confidence Score of 4 or Higher	
	Adverse Event Rate	Negligent Adverse Event Rate	Adverse Event Rate	Negligent Adverse Event Rate
	← % →			
1 or more	52.84	32.21	37.68	15.13
2 or more	27.16	10.95	19.16	4.00
All 3	12.21	4.00	7.58	0.82

“moderate”; the latter is considered “poor” (11). We had hypothesized that reliability would increase as the reviewer confidence score required to indicate an adverse event decreased. However, when adverse events were defined as having a confidence score of two or greater instead of four or greater, the overall κ statistic for adverse events decreased slightly, from 0.41 to 0.37 ($P = 0.19$).

Different review strategies produced different estimates of the total number of adverse events and negligent adverse events. Increasing the confidence score to indicate the presence of an adverse event and increasing the degree of consensus between independent reviewers (agreement of one, two, or three reviewers) substantially affected adverse event rates in this sample of 500 records (Table 2). For example, if all three reviewers had a confidence score of four or greater, the adverse event rate would be 7.58% and the negligent adverse event rate would be 0.82%. If we required a score of four or greater on only one of three reviews to confirm the presence of an adverse event, the adverse event rate would be 37.68% and the negligent adverse event rate would be 15.13%. When the required confidence score was decreased to two, estimates among the three reviewers varied even more.

DISCUSSION

We found moderate to poor inter-rater reliability among physicians trying to identify adverse events and negligent adverse events by medical record review. Increasing the required levels of reviewer confidence to indicate the presence of an event or increasing the number of physician reviewers who detected an event resulted in markedly different event rates. Similar studies of medical record review have found almost identical reliability (κ statistics of approximately 0.4 for both studies) (6, 7).

Our study is limited because our data come from hospital records in two states that may not be generalizable to other geographic locations. Furthermore, we could not measure the validity of chart review because there is no true gold standard that avoids some type of implicit assessment.

Despite the poor to moderate reliability of medical record review, each of the adverse events identified represents a potential opportunity for quality improvement (12, 13). Each case involving an adverse event may contain valuable information for improving patient safety. However, researchers should use caution when estimating incidence and prevalence of errors solely on the basis of these data.

For example, the IOM used data from our larger study (9), which used medical record review, to estimate that 44 000 Americans die each year of preventable adverse events (or, to use the IOM term, medical errors) (8). Our results suggest that these estimates are sensitive to the reviewer consensus and confidence required to indicate the presence of adverse events. The estimate of 44 000 deaths could be approximately 50% lower if the study used to estimate that figure required independent agreement by two physician reviewers and could be 30% higher if the required reviewer confidence about the presence of an adverse event was lower (Table 2). We cannot quantify the reduction more precisely because the IOM figures are based on judgments about the number of preventable adverse events and our results are based on judgments about adverse events and negligent adverse events.

The figures for error-related death reported by the IOM are imprecise. However, it is shortsighted to focus on the exact number of deaths and thereby ignore the vast additional research on errors and adverse events cited by the IOM (8, 14). Regardless of whether the number of annual U.S. deaths due to medical error is

30 000 or 300 000, the need to test interventions to reduce errors and adverse events is clear.

Our findings suggest that persons and institutions interested in improving patient safety will need more reliable methods of measurement to evaluate safety interventions based on, for example, differences in error rates before and after the intervention or between control and intervention groups in experiments. Because low reliability of outcome measures requires larger sample sizes to detect the effects of an intervention, chart review will often prove to be logistically and financially unfeasible. Recent studies continue to document the low reliability of medical record review (15), and efforts to improve the reliability of this method have been unsuccessful (16). Given this limitation, other methods of measuring well-defined errors and adverse events, such as direct observation of patient care, should be further explored to potentially improve reliability.

From Brigham and Women's Hospital and Harvard School of Public Health, Harvard University, Boston, Massachusetts; Medical University of South Carolina, Charleston, South Carolina; and University of Texas–Houston Medical School, Houston, Texas.

Grant Support: By the Robert Wood Johnson Foundation.

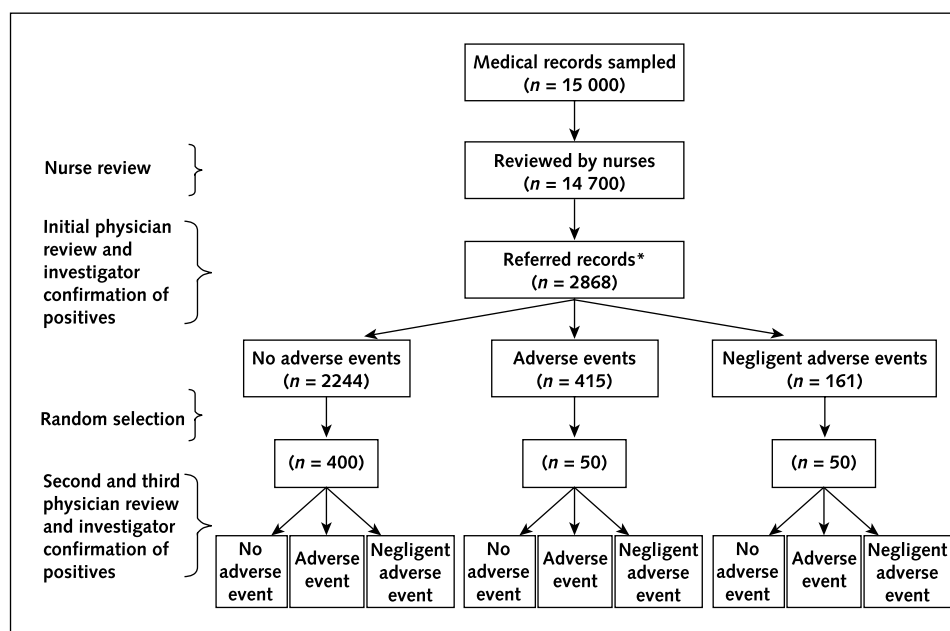
Requests for Single Reprints: Eric J. Thomas, MD, MPH, University of Texas–Houston Medical School, 6431 Fannin MSB 1.122, Houston, TX 77030.

Current author addresses and author contributions are available at www.annals.org.

References

- Brennan TA, Localio RJ, Laird NL. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. *Med Care.* 1989;27:1148-58. [PMID: 2593729]
- Localio AR, Weaver SL, Landis JR, Lawthers AG, Brenhan TA, Hebert L, et al. Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. *Ann Intern Med.* 1996;125:457-64. [PMID: 8779457]
- Petersen LA, Brennan TA, O'Neil AC, Cook EF, Lee TH. Does housestaff discontinuity of care increase the risk for preventable adverse events? *Ann Intern Med.* 1994;121:866-72. [PMID: 7978700]
- O'Neil AC, Petersen LA, Cook EF, Bates DW, Lee TH, Brennan TA. Physician reporting compared with medical-record review to identify adverse medical events. *Ann Intern Med.* 1993;119:370-6. [PMID: 8338290]
- Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. *Med Care.* 2000;38:261-71. [PMID: 10718351]
- Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med.* 1991;324:370-6. [PMID: 1987460]

Appendix Figure. Review sequence.



*Forty-eight of the referred records were not reviewed by physicians.

7. Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The Quality in Australian Health Care Study. *Med J Aust.* 1995;163:458-71. [PMID: 7476634]
 8. Kohn LT, Corrigan JM, Donaldson MS, eds. *To Err Is Human: Building a Safer Health System.* U.S. Institute of Medicine. Committee on Quality of Health Care in America. Washington, DC: National Academy Pr; 2000.
 9. Thomas EJ, Studdert DM, Newhouse JP, Zbar BI, Howard KM, Williams EJ, et al. Costs of medical injuries in Utah and Colorado. *Inquiry.* 1999;36:255-64. [PMID: 10570659]
 10. Leape LL, Lawthers AG, Brennan TA, Johnson WG. Preventing medical injury. *QRB Qual Rev Bull.* 1993;19:144-9. [PMID: 8332330]
 11. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). *J Clin Epidemiol.* 1992;45:567-80. [PMID: 1607896]
 12. Blumenthal D. Making medical errors into "medical treasures" [Editorial]. *JAMA.* 1994;272:1867-8. [PMID: 7990223]
 13. Berwick DM. Continuous improvement as an ideal in health care. *N Engl J Med.* 1989;320:53-6. [PMID: 2909878]
 14. Leape LL. Institute of Medicine medical error figures are not exaggerated. *JAMA.* 2000;284:95-7. [PMID: 10872022]
 15. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. *JAMA.* 2001;286:415-20. [PMID: 11466119]
 16. Hofer TP, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Med Care.* 2000;38:152-61. [PMID: 10659689]
- © 2002 American College of Physicians—American Society of Internal Medicine

My mother said, "Of course,
it may be nothing, but your father
has a spot on his lung."
That was all that was said: My father
at fifty-one could never
speak of dreadful things without tears.
When I started home,
I kissed his cheek, which was not our habit.
In a letter, my mother
asked me not to kiss him again
because it made him sad.
In two weeks, the exploratory
revealed an inoperable
lesion.
The doctors never
told him; he never asked,
but read *The Home Medical Guidebook.*
Seven months later,
just after his fifty-second birthday
—his eyesight going, his voice reduced to a whisper,
three days before he died—he said,
"if anything should happen to me . . ."

Donald Hall
The Old Life
Boston: Houghton Mifflin; 1996:62

Submitted by:
James J. Castles, MD
University of California, Davis
Davis, CA 95616

Submissions from readers are welcomed. If the quotation is published, the sender's name will be acknowledged. Please include a complete citation (along with page number on which the quotation was found), as done for any reference.—*The Editor*