

Anesthesiology
1996; 85:1049-54
© 1996 American Society of Anesthesiologists, Inc.
Lippincott-Raven Publishers

Variation in Expert Opinion in Medical Malpractice Review

Karen L. Posner, Ph.D.,* Robert A. Caplan, M.D.,† Frederick W. Cheney, M.D.‡

Background: Expert opinion in medical malpractice is a form of implicit assessment, based on unstated individual opinion. This contrasts with explicit assessment processes, which are characterized by criteria specified and stated before the assessment. Although sources of bias that might hinder the objectivity of expert witnesses have been identified, the effect of the implicit nature of expert review has not been firmly established.

Methods: Pairs of anesthesiologist-reviewers independently assessed the appropriateness of care in anesthesia malpractice claims. With potential sources of bias eliminated or held constant, the level of agreement was measured.

Results: Thirty anesthesiologists reviewed 103 claims. Reviewers agreed on 62% of claims and disagreed on 38%. They agreed that care was appropriate in 27% and less than appropriate in 32%. Chance-corrected levels of agreement were in the poor-good range ($\kappa = 0.37$; 95% CI = 0.23 to 0.51).

Conclusions: Divergent opinion stemming from the implicit nature of expert review may be common among objective medical experts reviewing malpractice claims. (Key words: Anesthesiology: liability; peer review. Insurance: claim review; liability. Malpractice, medical: expert testimony.)

THE role of the medical expert in malpractice litigation is to assist the court in determining the cause of injury and whether the applicable standard of care was met. Disagreement among medical experts not only plays a prominent role in court proceedings but also influences pretrial decisions on whether a claim of injury is pur-

sued and the manner in which it is pursued.^{1,2} An expert opinion that the standard of care was breached may lead a plaintiff to aggressively pursue a case or may encourage a defendant to settle the suit. Conversely, an expert opinion that the standard of care was met may lead to an aggressive defense by the defendant or withdrawal of the suit by the plaintiff.

The existence of conflicting expert opinion is often attributed to bias arising from monetary compensation by the plaintiff or defendant. Less obvious bias may occur when the expert develops a personal affinity for the plaintiff or defendant or for one or more members of the litigation team, introducing advocacy rather than objectivity into the opinions rendered.³

Another important source of conflicting expert opinion may be the nature of expert review. Expert opinion is a form of implicit judgment. Each expert may use his or her own unstated criteria to assess quality of care.⁴ This contrasts with an explicit process based on criteria specified before the assessment.⁵

Few studies of the reliability of quality of care judgments based on implicit peer review criteria have incorporated sound statistical methods.⁶ Those that have been conducted indicate that agreement is generally poor.⁶ However, most of these studies addressed the issue of peer review of quality of care in some general sense or in the context of quality assurance review,⁷⁻¹⁶ which differs from a judgment of whether the care met current standards in medical-legal review.^{17,18} Of the studies investigating the reliability of medical-legal judgments of appropriateness of care, few^{17,19,20} provided reviewers access to original medical records. In the remaining studies, assessments relied on case abstracts rather than complete and original documents in the peer review process.^{21,22}

This study measures the level of agreement among objective medical expert reviewers of actual malpractice claim files. The review was sponsored by an outside party with no role in the litigation, and the medical experts received no compensation for their reviews. All expert reviewers were members of the same specialty

* Research Associate Professor, Anesthesiology.

† Clinical Professor, Anesthesiology; and Chair, Committee on Professional Liability, American Society of Anesthesiologists.

‡ Professor and Chair, Anesthesiology.

Received from the Departments of Anesthesiology, University of Washington School of Medicine, and the Virginia Mason Medical Center, Seattle, Washington. Submitted for publication May 1, 1996. Accepted for publication July 19, 1996. Supported by the American Society of Anesthesiologists. Preliminary results presented at the 1991 Annual Meeting of the American Society of Anesthesiologists, San Francisco, California, October 29, 1991. The opinions expressed herein are those of the authors and do not represent the policy of the American Society of Anesthesiologists.

Address reprint requests to Dr. Posner: Department of Anesthesiology, University of Washington School of Medicine, Box 356540, 1959 NE Pacific Ave., Seattle, Washington 98195-6540. Address electronic mail to: posner@u.washington.edu.

at issue in the claim. They had access to complete claim files, as would a medical expert participating in litigation. With other sources of bias eliminated or held constant, the level of agreement among expert implicit judgments of the appropriateness of care was measured.

Methods

Data were collected as part of the American Society of Anesthesiologists (ASA) Closed Claims Project. This project, carried out by the ASA Committee on Professional Liability, is an ongoing study of adverse anesthetic outcomes based on information contained in the closed claim files of 34 professional liability insurance companies throughout the United States that insure approximately 14,500 anesthesiologists. The project and data collection procedures were described previously.²³⁻²⁷ Procedures specific to this investigation were carried out during ongoing data collection for the Closed Claims Project and are described below.

Between December 1988 and October 1994, teams of anesthesiologist-reviewers visited the offices of ten insurance companies to review files of closed malpractice claims against anesthesiologists. All claims except those for dental damage were eligible for inclusion in the study. At each company, sets of three files each were randomly selected by one of the reviewers from the entire set of files available for review at the start of the visit. Selection of three-file sets was made using a random-number list provided by the investigators. Each three-file set was reviewed by two anesthesiologists who were instructed to refrain from consulting with each other. When possible, paired reviews occurred on different dates, with time of day matched. Each participating anesthesiologist was eligible to review multiple sets of files with different partners on the same or different review visits, up to a maximum of two three-file sets (six files) per reviewer per visit and five three-file sets (15 files) total per reviewer over the course of the study. Reviewers with previous experience reviewing files for the ASA Closed Claims Project were eligible to participate in this study if they were active in the clinical practice of anesthesia and had been in practice at least 3 y. Completed reviews were sent to the central project office for analysis.

The claim review process consists of review of all materials in the file and completion of the ASA Closed Claims Project data collection form according to a standardized set of instructions. Typically a closed claim file

contains the hospital record, anesthesia record, narrative statements, statements of involved health care personnel, expert and peer reviews, deposition summaries, outcome reports, and cost of settlement or jury award. Reviewers completed a data collection form for each claim in which there was enough information to reconstruct the sequence of events and nature of the injury. The data collection form consists of more than 140 items covering basic demographics, anesthetic techniques and personnel, damaging events, patient injury, settlement information, assessments of preventability, judgment of appropriateness of care, and a narrative description of the sequence of events.

For this study, we measured agreement on only a single item on the data collection form: Was the anesthesia care appropriate, less than appropriate, or impossible to judge? Appropriate care was defined as care that was reasonable and prudent by the standards of anesthetic care at the time of the event. Less-than-appropriate care was defined as care that was less than that standard of a reasonable and prudent practitioner at the time of the event. Reviewers were instructed to render a judgment of "impossible to judge" if, due to inadequate or conflicting information, they could not determine if the standard of care had been met. Claims for which this question was unanswered by reviewers were excluded from analysis.

Because severity of patient injury can influence reviewer judgments of appropriateness of care,²¹ we separated the claims into two subsets for analysis: (1) temporary or nondisabling injuries and (2) permanent and disabling injuries. We also analyzed the entire set of claims as a group. Temporary or nondisabling injuries include such complications as emotional distress, sore throat, corneal abrasion, and uncomplicated pneumothorax. Permanent and disabling injuries include death, permanent brain damage, major nerve damage, and other injuries from which full recovery cannot occur or is not expected. Claims on which reviewers disagreed on this generalized level of severity of injury were excluded from the subset analysis.

Agreement between paired reviewers was measured using the kappa statistic (Appendix 1).^{28,29} Kappa values provide an index of the amount of agreement beyond that expected purely by chance. Because agreement is expected to exceed chance levels among a group of specialists, it is the amount of agreement beyond chance rather than simply statistical significance that serves as the measure of reliability. A kappa value less than 0.40 is considered poor agreement, and 0.40 to 0.75 is fair to good agreement beyond chance. A kappa

VARIATION IN EXPERT OPINION IN MEDICAL MALPRACTICE

Table 1. Paired Ratings of Appropriateness of Care

First Reviewer	Second Reviewer			Total
	Less than Appropriate	Appropriate	Impossible to Judge	
Less than appropriate	33 (32%)	13 (13%)	3 (3%)	49 (48%)
Appropriate	7 (7%)	28 (27%)	2 (2%)	37 (36%)
Impossible to judge	6 (6%)	8 (8%)	3 (3%)	17 (17%)
Total	46 (45%)	49 (48%)	8 (8%)	103 (100%)

Observed disagreement, 38%; observed agreement, 62%; expected disagreement, 60%; expected agreement, 40%. Percentages are based on the table total of 103 claims.

value greater than 0.75 is considered excellent.³⁰ Because a sample size of 25 to 30 is required for significance testing of kappa,²⁹ data collection continued until at least 30 claims in each of the analysis subsets were reviewed. Confidence intervals of kappa were calculated using jackknife calculations of standard error.²² Probability values ≤ 0.05 were considered statistically significant.

Results

One hundred three claims, each independently reviewed by two anesthesiologists, were eligible for inclusion in the study. In all, 30 anesthesiologists reviewed 2 to 15 claims each (median of five claims). The median age of reviewers was 48 y (range, 31 to 68 y) and all were board certified. Most had previous experience in an expert witness capacity (25 or 83%). Sixteen (53%) practiced in an academic setting, five (17%) were in private practice, and nine (30%) engaged in private practice with teaching responsibilities. Reviewers had been in practice 5 to 41 y (median, 16 y).

Overall, reviewers agreed on whether the care was appropriate in 64 (62%) of the claims and disagreed in 39 (38%; table 1). This level of agreement exceeded chance levels but was in the poor-to-good range (kappa = 0.37; 95% CI = 0.22 to 0.52). Reviewers agreed that care was appropriate in 27% of claims, less than appropriate in 32%, and impossible to judge in 3% (table 1).

Forty-two (41%) of the claims reviewed were for temporary or nondisabling injuries, whereas 50 (49%) involved permanent and disabling injuries. Agreement on severity of injury was excellent (kappa = 0.80). However, in 11 claims (11%), the reviewers disagreed on

the general severity of injury. These claims were not included in the subset analysis.

On claims for temporary or nondisabling injuries, the proportion of claims on which there was agreement on appropriateness of care was similar (64%) to agreement for the overall group of claims, although the chance-corrected level of agreement was less (kappa = 0.32). The level of agreement was less for permanent disabling injuries (60%, kappa = 0.27) than agreement for the overall group. All kappa values were statistically significant.

Discussion

Although the distinction between appropriate and less-than-appropriate care may seem evident to individual physicians, the results of the present study suggest that practicing anesthesiologists exhibit only fair agreement on this issue. When presented with identical malpractice claim files containing extensive documentation and records, objective reviewers agreed on appropriateness of care in 62% of claims (kappa = 0.37). Agreement was not improved by controlling for severity of injury. Although this level of agreement was statistically significant and the upper 95% confidence limit (0.52) did fall into the "good" range, it did not approach the excellent level (0.75). Nonrandom agreement (*i.e.*, statistical significance of kappa) is expected among any group of reviewers sharing similar training and is not a particularly meaningful assessment tool. A kappa value of 0.40 is generally considered the minimally acceptable level, and kappa values in the 0.4 to 0.6 range are common in studies of medical diagnosis and tests.

The use of multiple experts has the potential to improve the reliability of reviewer assessments.³¹ Application of the Spearman-Brown formula for stepped-up reliability to the findings of this study suggests that five objective reviewers would be needed to increase reliability to the "excellent" level of kappa = 0.75.^{30,31} Although a proposal to incorporate yet more experts into medical malpractice review might at first seem a costly alternative, the long-term consequences of repeatedly consulting a single expert whose opinions deviate from the community norm may be much greater. Incorporation of explicit criteria such as clinical practice guidelines into a structured review process might reduce the number of reviewers needed for a reliable assessment.

The process of review in the present study was analogous to that which occurs when a medical expert reviews a case for a malpractice proceeding, except that

We must be cautious in making generalizations from the results of this study because of several limitations of the study design. Reviewers were not selected at random but rather represent an opportunity sample selected from a national set of volunteers. Criteria for selection were geographic proximity to the site of claim files (insurance companies) and review experience. Although reviewers were not matched with claims by subspecialty, exclusion of cardiac, pediatric, and obstetric claims from the analysis did not change the results. Similarly, claims were not selected purely at random. Although claims reviewed at each site were selected by a random process, the companies were not. Only companies that allow access to their files and had more than 25 closed claims available for review were in-

Some of the proposals to mitigate the problem of poor agreement on implicit quality-of-care judgments might be applicable to expert review in medical malpractice. These include the use of a structured review process, acknowledged experts, multiple reviewers, practice guidelines, and separation of process and outcome assessments.⁶ The review process in our study was structured in that each reviewer completed the entire data form, which focused the review on a consistent set of case elements, a process similar to other studies.^{17,19,20} The experts in the present study were experienced in this review process and were practicing in the specialty being reviewed, thus fitting some of the criteria proposed for acknowledged experts.⁶ Although clinical practice guidelines are available in anesthesiology, the role of guidelines in the claims included in this study was not specifically assessed. Clinical practice guidelines have not replaced the medical expert in malpractice proceedings and do not appear to be the solution to the expert witness problem.^{36,37} Although clinical practice guidelines introduce explicit criteria into the review process, the expert may be relied on to determine whether the clinical practice guidelines apply to the case at hand or whether other evidence or factors

VARIATION IN EXPERT OPINION IN MEDICAL MALPRACTICE

		Reviewer 2		Total	
		A	B		
Reviewer 1	A	60 Agree	40	100 (0.50)	
	B	40	60 Agree	100 (0.50)	
Total		100 (0.50)	100 (0.50)	200	

Observed agreement = $\frac{60 + 60}{200} = 0.60$

Expected agreement =

$$\left(\frac{100}{200} \times \frac{100}{200}\right) + \left(\frac{100}{200} \times \frac{100}{200}\right) = 0.50$$

$$\text{kappa} = \frac{0.60 - 0.50}{1 - 0.50} = 0.20$$

Table A1.

		Reviewer 2		Total	
		A	B		
Reviewer 1	A	60 Agree	0	60 (0.30)	
	B	80	60 Agree	140 (0.70)	
Total		140 (0.70)	60 (0.30)	200	

Observed agreement = $\frac{60 + 60}{200} = 0.60$

Expected agreement =

$$\left(\frac{60}{200} \times \frac{140}{200}\right) + \left(\frac{140}{200} \times \frac{60}{200}\right) = 0.42$$

$$\text{kappa} = \frac{0.60 - 0.42}{1 - 0.42} = 0.31$$

Table A2.

are applicable,³⁶⁻³⁸ again introducing implicit judgments into the review process.

Anesthesiologists commonly disagree on the appropriateness of care when approaching the task of expert review with the intent of objectivity. This finding suggests that divergent expert opinions may be easily found by seeking opinions from multiple experts.

The authors thank the members of the American Society of Anesthesiologists who served as claims reviewers for this study and the insurance organizations that served as sources of closed claims.

Appendix: The Kappa Statistic

The kappa statistic provides a measure of agreement between different raters applicable to nominal-level ratings (categories). Some agreement may be expected purely by chance. The amount of agreement expected by chance may vary depending on several factors (the number of categories, the number of raters, the prevalence of the different categories, and how these categories are used by different raters). Kappa has been indexed to account for random agreement, usually providing a value between 0 and 1. A kappa of 0 indicates no agreement beyond chance, whereas a kappa of 1 is perfect agreement.

The general expression for kappa is

$$\frac{\text{Observed agreement} - \text{Expected agreement}}{1 - \text{Expected agreement}}$$

The following example (two raters, 200 cases, and two categories) illustrates how expected agreement (and kappa) may vary when raters agree on the same proportion of cases. In Tables A1 and A2, reviewers agree that 60 of the 200 cases are category A and 60 cases are category B, giving an observed agreement of

$$\frac{60 + 60}{200} = 0.60$$

The tables differ in the total ratings for each reviewer for each category. In table A1, each reviewer puts 100 (0.50) of the cases in category A and 100 (0.50) in category B. This gives an expected agreement of

$$(0.50)(0.50) + (0.50)(0.50) = 0.50$$

In table A2, reviewer 1 puts 60 (0.30) of cases in category A and 140 (0.70) in category B. Reviewer 2 does just the opposite, putting 140 (0.70) in category A and 60 (0.30) in category B. Expected agreement in this case is

$$(0.30)(0.70) + (0.70)(0.30) = 0.42$$

The chance that the two raters will agree (expected agreement) in table A2 (0.42) is less than in table A1 (0.50). The kappa value is affected by this difference. Table A2 has a larger kappa (0.31 vs. 0.20) because agreement by chance (expected agreement) is less than in table A1, leaving more room for agreement beyond chance.

For an introduction to the literature on kappa, see Posner and colleagues.²²

References

- Hyams AL, Brandenburg JA, Lipsitz SR, Shapiro DW, Brennan TA: Practice guidelines and malpractice litigation: A two-way street. *Ann Intern Med* 1995; 122:450-5
- Sloan FA, Hsieh CR: Injury, liability, and the decision to file a medical malpractice claim. *Law Society Review* 1995; 29:413-35
- Katz J: The fallacy of the impartial expert revisited. *Bull Am Acad Psychiatry Law* 1992; 20:141-52
- Angell M: Shattuck Lecture—Evaluating the health risks of breast implants: The interplay of medical science, the law, and public opinion. *N Engl J Med* 1996; 334:1513-18

5. Donabedian A: The quality of care. How can it be assessed? *JAMA* 1988; 260:1743-8
6. Goldman RL: The reliability of peer assessments of quality of care. *JAMA* 1992; 267:958-60
7. Bates DW, O'Neill AC, Petersen LA, Lee TH, Brennan TA: Evaluation of screening criteria for adverse events in medical patients. *Med Care* 1995; 33:452-62
8. Bigby JA, Dunn J, Adams JB, Jen P, Landefeld CS, Romaroff AL: Assessing the preventability of emergency hospital admissions. *Am J Med* 1987; 83:1031-36
9. Brook RH, Appel FA: Quality of care assessment: Choosing a method for peer review. *N Engl J Med* 1973; 288:1323-9
10. Dubois RW, Brook RH: Preventable deaths: Who, how often, and why? *Ann Int Med* 1988; 109:582-9
11. Hastings GE, Sonneborn R, Lee GH, Bick L: Peer review checklist: Reproducibility and validity of a method for evaluating the quality of ambulatory care. *Am J Public Health* 1980; 70:222-8
12. Hayward RA, McMahon LF, Bernard AM: Evaluating the care of general medicine inpatients: How good is implicit review? *Ann Intern Med* 1993; 118:550-6
13. Horn SD, Pozen MW: An interpretation of implicit judgments in chart review. *J Community Health* 1977; 2:251-8
14. Medical Review Project, Empire State Medical Scientific and Educational Foundation, Inc: Rochester region perinatal study. *N Y State J Med* 1976; 67:1205-10
15. Richardson FM: Peer review of medical care. *Med Care* 1972; 10:29-39
16. Rosenfeld LS: Quality of medical care in hospitals. *Am J Public Health* 1957; 47:856-65
17. Brennan TA, Localio RJ, Laird NL: Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. *Med Care* 1989; 27:1148-58
18. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, Hebert L, Newhouse JP, Weiler PC, Hiatt H: The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. *N Engl J Med* 1991; 324:377-84
19. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, Newhouse JP, Weiler PC, Hiatt HH: Incidence of adverse events and negligence in hospitalized patients. *N Engl J Med* 1991; 324:370-6
20. Brennan TA, Localio AR, Leape LL, Laird NM, Peterson L, Hiatt HH, Barnes BA: Identification of adverse events occurring during hospitalization. A cross-sectional study of litigation, quality assurance, and medical records at two teaching hospitals. *Ann Intern Med* 1990; 112:221-6
21. Caplan RA, Posner KL, Cheney FW: Effect of outcome on physician judgments of appropriateness of care. *JAMA* 1991; 265:1957-60
22. Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW: Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Stat Med* 1990; 9:1103-15
23. Cheney FW, Posner K, Caplan RA, Ward RJ: Standard of care and anesthesia liability. *JAMA* 1989; 261:1599-1603
24. Tinker JH, Dull DL, Caplan RA, Ward RJ, Cheney FW: Role of monitoring devices in prevention of anesthetic mishaps: A closed claims analysis. *ANESTHESIOLOGY* 1989; 71:541-6
25. Caplan RA, Posner KL, Ward RJ, Cheney FW: Adverse respiratory events in anesthesia: A closed claims analysis. *ANESTHESIOLOGY* 1990; 72:828-33
26. Kroll DA, Caplan RA, Posner K, Ward RJ, Cheney FW: Nerve injury associated with anesthesia. *ANESTHESIOLOGY* 1990; 73:202-7
27. Chadwick HS, Posner K, Caplan RA, Ward RJ, Cheney FW: Comparison of obstetric and non-obstetric anesthesia malpractice claims. *ANESTHESIOLOGY* 1991; 74: 242-9
28. Fleiss JL: Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; 76:378-83
29. Fleiss JL, Nee JCM, Landis JR: Large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 1979; 86:974-7
30. Fleiss JL: *Statistical Methods for Rates and Proportions*, Second edition. New York, John Wiley and Sons, 1981, p 218
31. Fleiss JL: *The Design and Analysis of Clinical Experiments*. New York, John Wiley and Sons, 1986, pp 14-5
32. Hirshfeld EB: Economic considerations in treatment decisions and the standard of care in medical malpractice litigation. *JAMA* 1990; 264:2004-12
33. Cheney FW, Posner KL, Caplan RA: Adverse respiratory events infrequently leading to malpractice suits. *ANESTHESIOLOGY* 1991; 75:932-9
34. Beck M: The hired gun expert witness. *Missouri Medicine* 1994; 91:179-82
35. Lundberg GD: Expert witness for whom? *JAMA* 1984; 252:251
36. Hyams AL, Shapiro DW, Brennan TA: Medical practice guidelines in malpractice litigation: An early retrospective. *J Health Polit Policy Law* 1996; 21:289-313
37. Garnick DW, Hendricks AM, Brennan TA: Can practice guidelines reduce the number and costs of malpractice claims? *JAMA* 1991; 266:2856-60
38. Hirshfeld EB: Should practice parameters be the standard of care in malpractice litigation? *JAMA* 1991; 266:2886-91