Anesthesiology
1998; 89:507–15

The Effect of Group Discussion on Interrater Reliability of Structured Peer Review

Rhoda D. Levine, M.D., Michael Sugarman, M.D.,† Wilma Schiller, M.D.,† Sarah Weinshel, M.D.,† Ellen J. Lehning, Ph.D.,† Robert S. Lagasse, M.D.‡*

PEER review, the evaluation of the appropriateness of clinical decisions made by physicians with a similar spectrum of expertise, has become an important aspect of medical quality management. Peer review of adverse outcomes has been suggested as a means of evaluating clinical competence[1] and recommended for the purpose of physician relicensure.[2] The use of an expert witness, as in malpractice litigation, also represents a unique form of peer review, wherein a single reviewer evaluates the quality of care provided by a physician involved in an adverse outcome.

For decisions of such magnitude, one would hope that physicians could agree about what constitutes appropriate care on a case-by-case basis, but this is often not true. Agreement among physicians (interrater reliability) regarding quality of care or clinical performance is often only slightly better than that expected by chance.[3] Suggestions for improving the interrater reliability of peer review include the use of multiple reviewers,[4] the availability of outcome data,§[5,6] and the application of formalized guidelines or structure to the evaluation process. A structured peer review (SPR) system focuses the reviewers' efforts on particular aspects of care and increases agreement among multiple reviewers.[7,8] Discussion of cases among multiple reviewers, rather than independent reviews, also may improve concordance through a similar mechanism.[9]‖

Systems of SPR have been used throughout medicine to evaluate quality of care.[8,10,11] We are aware of two well-described systems of SPR in anesthesiology: one developed by Vitez and endorsed by the American Society of Anesthesiologists# to judge clinical competence, and the other described by Lagasse *et al.*,[12] which evaluates system and human errors. In the current study, we evaluate these two SPR models and examined the effect of group discussion on interrater reliability.

Methods

Case Selection

Twenty-five cases previously judged to involve a perioperative indicator (an event or action that leads to an adverse outcome) were selected randomly from the peer review database files of the University Hospital at Stonybrook (UHSB), and another 25 cases were selected from the database files of the Jacobi Medical Center (JMC). Both institutions routinely reviewed cases involving indicators reported by practitioners, chart reviewers, and other health-care personnel that met the indicator criteria as judged by a group of anesthesiologists at their respective institutions. The UHSB, a 650-bed suburban hospital affiliated with the State University of New York Medical School at Stony Brook, used the Lagasse *et al.* model of SPR[12] to review cases involving adverse patient outcomes occurring between 1992 and 1994. The JMC, a 750-bed municipal hospital affiliated with the Albert Einstein College of Medicine and Montefiore Medical Center in New York City, used the Vitez

* Associate Professor of Anesthesiology.

† Assistant Professor of Anesthesiology.

‡ Associate Professor of Clinical Anesthesiology.

Received from the Department of Anesthesiology, Albert Einstein College of Medicine/Montefiore Medical Center, Bronx, New York. Submitted for publication July 28, 1997. Accepted for publication March 19, 1998.

Address reprint requests to Dr. Levine: Department of Anesthesiology, Albert Einstein College of Medicine, 1825 Eastchester Road, Bronx, New York 10461-2373. Address electronic mail to: rdlevine@ix.netcom.com

Key words: Error classification; indicators; outcome measurement; quality management.

§ Brook RH: Quality of care assessment: A comparison of five methods of peer review. Washington, DC, United States Department of Health, Education, and Welfare, 1973.

‖ Ludke RL, Wakefield DS, Booth BM, Kern DC: Pilot study of nonacute utilization of VAMC inpatient service: Final report. SDR 87-003, Washington, DC, United States Department of Veterans Affairs, 1990.

Vitez T: Judging clinical competence. Park Ridge, IL, American Society of Anesthesiologists, 1989.

model of SPR# to review adverse patient outcomes from 1986 to 1994. The UHSB and JMC databases, from which the 50 study cases were selected randomly, comprised 869 cases and 231 cases, respectively, and each case included a description of the circumstances surrounding the event. Results of the original peer review were not included in the database information available for this study.

Peer Review

Five board-certified anesthesiologists who were members of the Department of Anesthesiology Quality Management Program of the Albert Einstein College of Medicine/Montefiore Medical Center served as reviewers for this study. Each reviewer received identical instructions and training in the use of the two models. In addition, an "expert" with considerable experience in the use of the Vitez and Lagasse *et al.* models was available for consultation by the reviewers during the training period, although none of the reviewers had any recollection of the previously presented cases. The same five reviewers were used throughout the study.

In the first phase of the study, the 25 UHSB cases were reviewed by the Vitez model as used at JMC. Written definitions of indicators, outcome score criteria, and error classification terminology were available at the time of the review. Each case was presented verbally in abstract form as it existed in the database. Additional materials, such as anesthesia records, were available as requested by the reviewers. The review committee members were not given any information about indicators, error classifications, or outcome judgments made by the original review at the previous institution.

After a single presentation of each case, the reviewers, individually and without discussion, selected the indicator(s), outcome score(s), and error(s) according to the structures imposed by this model. The reviewers then participated in a group discussion of the case. First, each reviewer, in no particular order, briefly presented his or her evaluation of the case. Available chart material was reexamined as appropriate to clarify factual issues. Frequently, a reviewer would critique another's evaluation briefly, to support an observation of particular importance or to offer an alternative point of view. No attempt was made to achieve consensus. There was no time limit on discussion, which ended when all of the reviewers had expressed their views to their own satisfaction. Each reviewer then reevaluated the case after discussion using the same structures imposed by the model. Reviewers were blinded to each others' and to the overall group responses throughout the study period.

In a similar manner, each of the 25 JMC cases was evaluated before and after group discussion using the Lagasse *et al.* model. Written definitions were available at the time of the reviews, and discussions proceeded as described in the preceding paragraphs.

Two months after the end of the initial review period (≈4 months after the start of the study), the 50 cases were reassessed in the same sequence with and without group discussion using the model of SPR that had not been used initially in a crossover design.

Peer Review Model Structures

Both models are similar in that they use tables of events that may lead to adverse outcomes (indicators), outcome scores that correspond to a continuum of increasing care or severity of injury (outcomes), and categorization of the reason or mechanism for the event (error classification). In addition, both systems use multiple reviewers who act as a committee to reach a conclusion about case management. The Vitez model was established primarily as a means of evaluating clinical competence. Based on the assumption that a poor practitioner would have a different error profile than a good one, this model emphasizes human errors. In contrast, Lagasse *et al.* took the position that ≈90% of quality problems are best categorized as difficulties in the system rather than the result of individual human error. Therefore, categories of system errors were included in the Lagasse *et al.* model of SPR to provide additional opportunities to identify problems in medical care.

Indicators. Indicators had been selected previously by each anesthesia department according to their needs. Table 1 lists indicators used at UHSB for the Lagasse *et al.* model. Table 2 lists the indicators selected for use at JMC with the Vitez model. Although the indicators are organized differently, most appear in both models. Indicators that appear in only one of the two models are noted by an asterisk. These indicators would have been classified as "other" in the second system's data set. To approximate the more detailed indicator list in the Lagasse *et al.* model for statistical purposes, we combined each Vitez "indicator" with a "management category" (table 2) to create a composite indicator. That is, each reviewer classified the case by first selecting an indicator and then a management category.

Reviewers in the current study were instructed to assign one indicator for each untoward event. Multiple indicators could be selected in cases with multiple

Table 1. UHSB Indicators

General
- Mortality within 48 h*
- Unplanned admission to ICU*
- Surgery delayed/canceled
- Patient dissatisfaction†

Patient assessment
- Failure to recognize patient disease†
- Lack of medical optimization†
- Failure to obtain informed consent†
- Other

Anesthetic medications
- Overdose
- Inappropriate use
- Allergic reaction
- Inappropriate premedication*
- Ampule or syringe swap
- Toxic reactions
- Incorrect controlled substance count†
- Other

Fluid/blood products
- Fluid overload
- Inadequate fluid resuscitation
- Overtransfusion of blood products
- Inadequate use of blood products
- Transfusion reaction
- Transmission of disease†
- Other

Anesthesia equipment
- Failure to check equipment
- Failure to adhere to monitoring standards†
- Failure to detect disconnect/leak
- Other

Patient positioning
- Damage or loss of skin/hair
- Ocular injury
- Peripheral nerve injury
- Vascular injury
- Other

Cardiovascular
- Cardiac arrest during anesthesia care
- Myocardial ischemia
- Dysrhythmias requiring treatment
- MI within 48 h
- CHF
- Hypertensive/hypotensive outcome
- Other

Respiratory
- Hypoxemia ($Sp_{O_2} < 90$ with O_2)
- Hypercarbia/hypocarbia
- Aspiration pneumonitis
- Pulmonary embolism
- Pneumothorax/hemothorax
- Respiratory failure/reintubation within 24 h
- Bronchospasm
- Other

Airway problems
- Ocular injury
- Undetected esophageal intubation
- Failed tracheal intubation/ventilation
- Damage to larynx/trachea
- Severe epistaxis
- Airway combustion (laser surgery)*
- Laryngospasm
- Other

Nervous system
- Delayed emergence (>60 min)†
- Awareness under GA
- CNS injury
- Post dural puncture headache
- Inadvertent dural puncture
- Peripheral nervous system injury
- Failed regional anesthetic
- Other

* Indicator tracked in JMC "other" category.

† Indicator unique to UHSB.

events (*e.g.*, reintubation and dental injury). Individual reviewers were free to determine the number of adverse events and appropriate indicators present for each case.

Severity of Outcome. Reviewers selected an outcome score for each indicator they identified. The Lagasse *et al.* model uses a 5-step scale to classify severity of outcome, whereas the Vitez model uses an 11-step scale. Therefore, to compare interrater reliability, the Vitez model outcome scores were combined to achieve scores of the same range and resolution (table 3).

Error Classification. An error category was assigned to each indicator chosen by a reviewer. The error categories used in the two SPR models are presented in table 4. Error categories appearing in only one model are denoted by an asterisk. Both models include human and system (nonhuman) errors. The Vitez model allows greater detail for human errors than the Lagasse *et al.* model. For example, errors of inadequate knowledge in the Vitez model are divided into didactic and experiential errors, inadequate data errors are broken down into "failure to seek necessary data" or "collection of irrelevant data," and an error involving disregarded data could be classified as "failure to accept a conclusion" or "failure to recognize a problem" despite the availability of appropriate data. The human errors involving "lack of an alternative plan" and problems with general "vigilance" are present only in the Vitez model. In contrast, the only system error accepted by the Vitez model is a mechanical error, and no error category was assigned to indicators that were considered unavoidable or unrelated to anesthetic management. The Lagasse *et al.* model includes a detailed classification of system errors including communication errors, limitation of di-

Table 2. JMC Indicators and Management Categories

- Indicators
  - Cancellation of surgery
  - Respiratory/cardiac arrest
  - Soft tissue injury
  - New neurological abnormality
  - Difficult ventilation/intubation
  - Reintubation in OR/PACU
  - Regurgitation/aspiration
  - Unstable cardiac rhythm/vital signs
  - Perioperative MI/ischemia
  - Patient abandonment*
  - Dental injury
  - Equipment malfunction/misuse
  - Delayed response to emergency*
  - Any other event related to anesthetic management
- Management Categories
  - Airway
  - Staff behavior*
    - Rules
    - Unprofessional conduct
    - Substance abuse
  - Circulatory
  - CNS/PNS
  - Dental
  - Drug action
    - Inhalation agents
    - Relaxants
    - Opioids
    - Sedatives/hypnotics
    - Local anesthetics
    - Allergic reaction
    - Drug interaction
    - Drug swap
    - Cardiovascular drugs
  - Electrical*
  - Endocrine
  - GI*
  - Hematologic
    - Anemia
    - Transfusion
    - Coagulation
  - Hepatic*
  - Instrumentation
    - Machine
    - Invasive monitor
    - Noninvasive monitor
  - Intravenous infusion
  - Metabolic
    - Fluid/electrolytes
  - Pulmonary
    - Oxygenation
    - Parenchymal
    - Ventilation
  - Position injury
  - Regional technique
  - Renal*
  - Temperature*

* Indicator unique to JMC.

agnostic standards, limitation of therapeutic standards, technical accidents, equipment failure, limitation of resources, and errors associated with supervision of a resident or nurse anesthetist. The definitions of the errors in each model were available to the reviewers throughout the review process.

The interrater reliability of the error classification was evaluated for specific errors (*e.g.,* the system errors "limitation of therapeutic standards" and "limitation of diagnostic standards" were considered different errors) and for human and nonhuman errors in general. In the Lagasse *et al.* model, the system errors "limitation of therapeutic standards" and "limitation of diagnostic standards" could be grouped together as "system" errors. Similarly, the Vitez model mechanical errors and those indicators involving no error also could be grouped together as system errors. This grouping allowed us to examine the relation between the severity of outcome and the likelihood of assigning human error.

Statistical Analysis

Interrater reliability was measured using the S_{av} statistic of O'Connell and Dobson[13] with software written by John Reed for cases reviewed by both models. S_{av} was calculated for each unique indicator chosen by one or more reviewers along with the associated outcome score and error classification before and after discussion. The traditional method for evaluating interrater reliability is the kappa statistic described by Fleiss[14] and is used for data assessed by different sets of raters. S_{av} is a kappa-like statistic that is used when each case of noncategorical data is evaluated by the same set of raters.[15,16] Like kappa, S_{av} expresses the proportion of agreement among the same reviewers beyond that expected by chance. For the purpose of this study, an S_{av} value <0.40 was considered poor agreement, 0.40–0.75 was considered fair to good agreement, and >0.75 was considered excellent agreement.[16] Data are reported as the S_{av} value $\pm$ the 95% confidence interval. Variances used to calculate the 95% confidence interval were derived using the jackknife procedure without the constraint of marginal homogeneities.[16]

The extent to which individual reviewers agreed with other members of the group was measured using the Williams Index of Agreement.[17] This index expresses the ratio of the calculated agreement between the individual reviewer and the rest of the group to the average extent of agreement among all raters in the rest of the group. This index was calculated on postdiscussion data to see if one member of the group was consistently

Table 3. Lagasse Model Outcomes and Vitez Model Outcomes

Lagasse
1 No change in hospital course
2 Increased care without function deficit
3 Increased care with reversible function deficit
4 Increased care with irreversible function deficit
5 Death

Corresponding Lagasse outcomes	Vitez	
1	0	No change in hospital course
2	1	Additional unexpected care (*e.g.*, additional drugs, tests, consult)
	2	Prolonged hospital stay (*e.g.*, postpone surgery)
	3	Prolonged hospital stay with increased level of care (*e.g.*, ICU admit, bronchoscopy)
3	4	Reversible organ damage requiring additional drugs, tests or care (*e.g.*, 10% pneumothorax, corneal abrasion)
	5	Reversible organ damage involving prolonged hospitalization (*e.g.*, new myocardial ischemia)
	6	Reversible organ damage with prolonged hospital stay and increased care or risk (*e.g.*, CHF, aspiration pneumonia)
4	7	Irreversible organ damage with residual that does not significantly alter patient function (*e.g.*, small burn, small MI)
	8	Irreversible organ damage with residual that significantly alters patient function (*e.g.*, CVA, cauda equina syndrome)
	9	Irreversible organ damage incapacitating patient (*e.g.*, hemiplegia, hypoxic encephalopathy)
5	10	Death

influencing other members of the group. Values of the Williams index indicate the extent to which individual reviewer ratings agree with those of the other members of the group. A Williams index of 1.0 indicates individual agreement consistent with the average extent of agreement for the rest of the group, whereas an index in which the 95% confidence interval does not include 1.0 indicates a rating that is dissimilar to that of the rest of the group.

Results

Agreement among reviewers regarding indicators, severity of outcomes, and error classification are expressed as the S_{av} statistic (table 5). The level of agreement for error classification is shown for specific and general errors (*e.g.*, human *vs.* nonhuman), as previously described. Both models were dependent on group discussion to produce reliable agreement among reviewers regarding indicators, severity of outcomes, and error classification. Reviewer agreement on indicators was in the poor range before discussion (S_{av}, 0.23–0.28) but agreement improved to be in the excellent range (S_{av}, 0.82–0.87) after discussion. Agreement on outcome scores before discussion was in the fair to good range (S_{av}, 0.4–0.5) and improved to excellent (S_{av}, 0.89–0.93) after discussion. The worst agreement among reviewers was seen for specific error analyses (S_{av}, 0.07–0.10), but this still improved to the good range (S_{av}, 0.71–0.74) after discussion. Interrater reliability for a general classification of human *versus* system errors also showed an initial poor agreement (S_{av}, 0.29–0.30) but improved to be in the excellent range (S_{av}, 0.84–0.86) after group discussion.

Although both models of SPR demonstrated high interrater reliability after discussion, there were differences between the models regarding general error classification. Reviewers found an overall human error rate of 62% for indicators analyzed with the Vitez model and 30% for indicators analyzed with the Lagasse *et al.* model after discussion. Neither model demonstrated a tendency to assign human error to indicators involving more severe outcomes (fig. 1).

The extent to which individual reviewers agreed with other members of the group is expressed as a Williams Index of Agreement (table 6). In the Lagasse *et al.* model, one reviewer consistently demonstrated a significantly higher propensity to agree with the group after discussion. In the Vitez model, multiple reviewers showed significant differences in their ability to agree with the error analysis after discussion.

Discussion

Prior investigations, reviewed by Goldman, have suggested several modalities that might promote interrater reliability of peer review, notably the use of structure in the peer review model, multiple reviewers, case dis-

Table 4. Lagasse Model Errors and Vitez Model Errors

Lagasse model errors
Human Errors
Operator error
Improper technique
Inadequate data sought
Data disregard
Inadequate knowledge
System Errors
Equipment failure
Technical accidents
Communications
Limitation of therapeutic standards
Limitation of diagnostic standards
Limitation of resources available
Supervision of resident/CRNA
Vitez model errors
Human technical
Accidental
Improper technique
Human vigilance*
Human judgment
Inadequate knowledge
Didactic
Experiential
Inadequate data
Failure to seek
Collection of irrelevant data
Disregard data
Failure to recognize problem
Failure to accept conclusion
Lack of alternative plan*
Mechanical
None*

* Error unique to Vitez model.

Table 5. Interrater Reliability before and after Discussion

	S_{av} before Discussion	S_{av} after Discussion
Indicators		
Vitez model	0.23 ± 0.01	0.82 ± 0.004
Lagasse model	0.28 ± 0.01	0.87 ± 0.004
Severity of outcome		
Vitez model	0.40 ± 0.01	0.89 ± 0.01
Lagasse model	0.50 ± 0.01	0.93 ± 0.01
Specific error classification		
Vitez model	0.10 ± 0.004	0.71 ± 0.004
Lagasse model	0.07 ± 0.004	0.74 ± 0.005
General error classification (human *vs.* system)		
Vitez model	0.29 ± 0.005	0.84 ± 0.005
Lagasse model	0.30 ± 0.005	0.86 ± 0.004

±95% confidence interval.

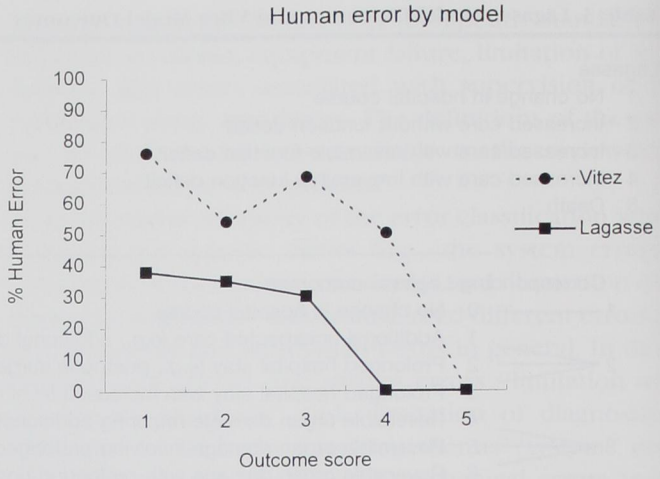

Fig. 1. Although the Vitez model of SPR# is more likely to result in the assignment of human error than the Lagasse *et al.* model,[12] neither model demonstrated a tendency to assign human error to indicators involving more severe outcomes. The outcome scores represent a continuum from no change in hospital course (1) to death (5).

cussion, and the availability of outcome data.[3] For the most part, these studies relied on kappa and kappa-like statistics to evaluate interrater reliability.

Kappa and kappa-like statistics can be difficult to interpret for several reasons. If selection of the available

Table 6. Williams Index of Agreement after Discussion

	Lagasse Model	Vitez Model
Indicators		
Reviewer 1	0.99	0.98
Reviewer 2	1.05*	1.00
Reviewer 3	1.02	0.94
Reviewer 4	0.99	1.04
Reviewer 5	0.95	1.03
Outcomes		5-step
Reviewer 1	0.96	0.99
Reviewer 2	1.08*	1.06
Reviewer 3	1.01	0.94
Reviewer 4	1.00	1.05
Reviewer 5	0.96	0.97
Errors		
Reviewer 1	0.93	0.97
Reviewer 2	1.07*	1.06*
Reviewer 3	1.04	0.90*
Reviewer 4	0.98	1.07*
Reviewer 5	0.98	1.00

* 95% confidence interval does not include 1.0. All variances are ±0.003 or smaller.

choices in each category was attributable to chance alone, the number of choices might affect the level of agreement as determined by the S_{av} statistic. We attempted to lessen this potential bias by offering similar numbers of choices for each model. Similarly, the S_{av} values for indicators and specific error classification would be expected to be lower than that seen for outcome severity. Another criticism of kappa or kappa-like statistics is that they are influenced by the prevalence of the variable being evaluated. An event that occurs rarely would tend to lower the calculated values. This potential bias would not alter the statistical effect of discussion because the frequency of events and the number of choices remained the same before and after discussion. Further, one must accept the fact that kappa and kappa-like statistics are an overall average and thus can blind the reader to the possibility that one subcategory of indicator, outcome, or error classification may be responsible for most of the disagreement among reviewers.

Despite the use of two highly structured peer review models, reviewer agreement before discussion in this study was poor for identification of indicators and classification of errors using either model. The best prediscussion agreement was observed for outcome scores, but even this was only in the fair to good range. It is particularly noteworthy that the reviewers were unable to agree independently on the identification of the indicator and the accompanying severity of outcome before discussion, even though all reviewers were exposed to the same description of the event and had a limited number of choices imposed by the models. This finding, however, is consistent with previously reported reviews of SPR systems[9,18-20] and suggests that simply providing structure to the process is insufficient to promote adequate agreement.

The use of multiple reviewers also has been recommended as a means of improving agreement.[4] Often, however, one must limit the number of reviewers because the peer review process is time consuming and costly. Although the optimum number of reviewers has not been established, most published studies have used two to five. Posner *et al.*[21] suggested that five independent reviewers were needed to bring kappa values up to the "excellent" level. Our study found that S_{av} values were low before discussion, even when using five independent reviewers. Therefore, five may not be an optimum number of reviewers if the reviewers are allowed to act independently.

For peer review of adverse outcomes to be meaningful, reviewers must be able to agree on the key elements of the case. Our study confirms the work of Ludke *et al.*‖ and Wilson *et al.*,[9] which indicates that discussion among reviewers improves agreement in peer review. Discussion of cases increased S_{av} values for indicators, outcomes, and error classification, with agreement for all improving dramatically. Although the Lagasse *et al.* model had significantly better agreement than the Vitez model after discussion based on S_{av} and confidence intervals (table 5), the interpretation of these values is difficult. Clearly, group discussion improves interrater reliability, but the reasons for this improvement also remain unclear. Possibilities to consider include an increased knowledge of the model, improved understanding of the important aspects of a particular case, the presence of a content expert, peer pressure from the leader of the group or a member with a strong personality, or other group dynamics (*e.g.*, bargaining, lobbying).

All cases were reviewed twice by the same reviewers using the two different review systems. By re-presenting cases, reviewers had another opportunity to become more familiar with the peer review model or the case scenarios; however, we did not find that prediscussion agreement increased in the latter part of the study for either model. This is probably because the group results of the first review were not known at the time of the subsequent presentation of the case, and the reviewers were then bound by definitions of a different review system. In addition, because half of the cases were reviewed initially with the Vitez model and half initially with the Lagasse *et al.* model in a crossover design, the potential bias of familiarity should have been similar for the two models.

The authors observed no substantial improvement in interrater reliability when reviewing subsequent cases or when the model was used for the second time. Therefore, there appears to be no learning curve for the models during the study period, although the reviewers did note specific instances where the system's definition of an indicator, outcome, or error was clarified during the discussion period. It is possible that, because of the small number of cases in the study, such a learning effect could not be demonstrated.

Fine and Moorehead stated that a physician with expertise in a specific area reaches a judgment about quality of performance that agrees with 90% of others of equal experience. Only 2–3% of the remaining 10% are not resolved by discussion.[11] Our reviewers, although all board-certified anesthesiologists, had various levels

of expertise in subspecialty areas of anesthesia, such as obstetric and cardiac anesthesia. When such cases were discussed, the group member(s) with additional experience in the areas related to the case tended to dominate the discussion. We do not know to what extent, if at all, this influenced the individual reviewers after discussion. An area for further study is to see if group discussion improves the interrater reliability of experts in subspecialty areas to the same degree as it did for our reviewers.

The influence of group dynamics and peer pressure on interrater reliability after discussion is important to elucidate. Because the assessments of the individual reviewers were held in confidence in our study, pressure to agree or disagree with other reviewers was held to a minimum. The Williams index results (table 6) suggest that one of the reviewers using the Lagasse *et al.* model was more likely to agree with the overall group after discussion. This may be interpreted as an individual who is strongly influenced by the group or, in contrast, capable of making the group agree with their point of view. In our study, the number of times that this reviewer changed opinions from the prediscussion analysis suggests the former. In the Vitez model, there were significant differences in individual reviewers' propensity to agree with the group regarding error analysis. This is consistent with the findings of the S_{av} statistic that show the worst agreement for this component of the Vitez model.

Human error encompasses errors in judgment and inappropriate or untimely action. Overall, the reviewers found a 62% human error rate using the Vitez model and a 30% human error rate using the Lagasse *et al.* model. Although the Vitez model does emphasize human error, the lower rate of human error found using the Lagasse *et al.* model is in part attributable to the classification of "technical accidents" (20% in this study). Technical accidents are defined as adverse outcomes that occur despite the fact that a technique was performed correctly and are classified as human errors in the Vitez model but as system errors in the Lagasse *et al.* model. For example, post-dural puncture headache after a properly performed spinal anesthetic procedure is considered a fault of the system by the Lagasse *et al.* model. System errors that have contributed to this adverse outcome in the past include needle size and design. It has been suggested that this type of system error contributes to ≈16% of all perioperative adverse outcomes.[12]

Caplan *et al.*[19] suggested that outcome data should be withheld when determining appropriateness of care, particularly if permanent injury is involved, because knowledge of outcome may bias reviewers toward human error. A relation between severe outcome and subsequent classification as human error was not demonstrated in our study for either SPR model (fig. 1). Others have shown that outcome data is necessary to ensure adequate agreement between reviewers.[5,6,22] The Vitez and Lagasse *et al.* SPR models are outcome driven; that is, reviewers are aware of patient outcomes at the time of the review. We believe that outcome data are necessary when making decisions about quality of care.

We evaluated the interrater reliability of two different SPR models. There was a significant but not a dramatic difference between the two models regarding reviewer agreement after discussion for indicators, outcome scores, and error classification. Overall, interrater reliability before group discussion was poor. A striking increase in interrater reliability, however, occurred for all variables after group discussion with both SPR models. The authors conclude that the SPR systems studied do not alone ensure adequate agreement among reviewers. Therefore, it must be recommended that systems of peer review used for recredentialing, relicensure, or medical malpractice should not depend on the opinions of individual reviewers because of the high variability. This is not meant to suggest that improved agreement guarantees optimal quality or objectivity of each review. More research needs to be directed at refining the group-based process before it can be claimed that the goals of peer review are being met.

The authors thank Dr. Karen L. Posner, Research Associate Professor, Anesthesiology, University of Washington, for assistance in the statistical analysis of interrater reliability and Dr. Sherman Levine for advice in preparing the manuscript.

References

1. Vitez TS: A model for quality assurance in anesthesiology. J Clin Anesth 1990; 2:280–7

2. Gabel RA: Quality assurance/peer review for recredentialing/relicensure in New York State. Int Anesthesiol Clin 1992; 30:93–101

3. Goldman RL: The reliability of peer assessments of quality of care. JAMA 1992; 267:958–60

4. Dubois RW, Brook RH: Preventable deaths: Who, how often, and why? Ann Intern Med 1988; 109:582–9

5. 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: Results of the Harvard Medical Practice Study I. N Engl J Med 1991; 324:370–6

6. Brennan TA, Sox CM, Burstin HR: Relation between negligent

adverse events and the outcomes of medical-malpractice litigation. N Engl J Med 1996; 335:1963–7

7. Donabedian A: Explorations in Quality Assessment and Monitoring: The Criteria and Standards of Quality. Volume 2. Ann Arbor, Health Administration Press, 1982, pp 17–61

8. Richardson FM: Peer review of medical care. Med Care 1972; 10:29–39

9. Wilson DS, McElligott J, Fielding PL: Identification of preventable trauma deaths: Confounded inquiries? J Trauma 1992; 32: 45–51

10. Hastings GE, Sonneborn R, Lee GH, Linda V, Sasmor 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

11. Fine J, Moorehead MA: Study of peer review of inhospital patient care. N Y State J Med 1971; 71:1963–73

12. Lagasse RS, Steinberg ES, Katz RI, Saubermann AJ: Defining quality of perioperative care by statistical process control of adverse outcomes. ANESTHESIOLOGY 1995; 82:1181–8

13. O'Connell DL, Dobson AJ: General observer-agreement measures on individual subjects and groups of subjects. Biometrics 1984; 40:973–83

14. Fleiss JL: Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76:378–82

15. Davies M, Fleiss JL: Measuring agreement for multinominal data. Biometrics 1982; 38:1047–51

16. 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

17. Williams GW: Comparing the joint agreement of several raters with another rater. Biometrics 1976; 32:619–27

18. 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

19. Caplan RA, Posner KL, Cheney FW: Effect of outcome on physician judgments of appropriateness of care. JAMA 1991; 265: 1957–60

20. Brennan TA, Localio RJ, Laird NL: Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. Med Care 1989; 27:1148–58

21. Posner KL, Caplan RA, Cheney FW: Variation in expert opinion in medical malpractice review. ANESTHESIOLOGY 1996; 85:1049–54

22. Horn SD, Pozen MS: An interpretation of implicit judgments in chart review. J Commun Health 1977; 2:251–8