Reliability of medical audit in quality assessment of medical care

Confiabilidade da auditoria médica na avaliação de qualidade da atenção médica

Luiz Antonio Bastos Camacho [1]
Haya Rahel Rubin [2]

[1] *Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz. Rua Leopoldo Bulhões 1480, Rio de Janeiro, RJ 21041-210, Brasil.*
[2] *Johns Hopkins School of Medicine, Johns Hopkins University. 1830 E. Monument St. Baltimore, MD. 21205, U.S.A.*

Abstract *Medical audit of hospital records has been a major component of quality of care assessment, although physician judgment is known to have low reliability. We estimated interrater agreement of quality assessment in a sample of patients with cardiac conditions admitted to an American teaching hospital. Physician-reviewers used structured review methods designed to improve quality assessment based on judgment. Chance-corrected agreement for the items considered more relevant to process and outcome of care ranged from low to moderate (0.2 to 0.6), depending on the review item and the principal diagnoses and procedures the patients underwent. Results from several studies seem to converge on this point. Comparisons among different settings should be made with caution, given the sensitivity of agreement measurements to prevalence rates. Reliability of review methods in their current stage could be improved by combining the assessment of two or more reviewers, and by emphasizing outcome-oriented events.*

Key words *Reproducibility; Health Care Quality Assurance; Medical Audit; Quality of Health Care; Medical Records*

Resumo *Auditoria médica de prontuários hospitalares tem sido um componente importante da avaliação da atenção à saúde, embora se saiba que o julgamento médico tem baixa confiabilidade. Nós estimamos a concordância interobservador da avaliação médica da qualidade da atenção hospitalar em uma amostra de pacientes com problemas cardíacos admitidos em um hospital universitário americano. Os médicos revisores aplicaram métodos estruturados de revisão desenvolvidos para melhorar a avaliação subjetiva de qualidade. A concordância corrigida para o acaso (Kapa) dos itens considerados mais relevantes do processo e do resultado da atenção médica variaram de baixo a moderado (0,2 a 0,6), dependendo do diagnóstico principal e dos procedimentos a que os pacientes tinham sido submetidos. Nesse aspecto, os resultados obtidos por outros autores parecem convergir. No entanto, as comparações entre diferentes estudos são limitadas, pois que as medidas de concordância corrigida para o acaso são influenciadas pela prevalência do evento de interesse, sobre a qual os trabalhos publicados raramente informam. No estágio atual, a confiabilidade dos métodos de avaliação de prontuários médicos pode ser melhorada pela combinação da avaliação de dois ou mais médicos, e por uma maior ênfase nos eventos relacionados aos resultados da atenção médica.*

Palavras-chave *Reprodutibilidade dos Resultados; Garantia de Qualidade dos Cuidados de Saúde; Auditoria Médica; Qualidade dos Cuidados de Saúde; Registros Médicos*

Introduction

Quality of care evaluation has been typically based on retrospective analysis of clinical records using either clearly stated criteria or judgment based on individual experience (medical audit). Although medical audit dates back to the 1920s, it was during the 1950's that more structured formats replaced collection of opinions based on individual subjective assessment (Lembcke, 1967; Butler & Quinlan, 1958). Initially, medical audits were applied mainly on charts selected because of deaths or complications. In the 1970's systematic medical audits were started in the United States by the Professional Standards Review Organization program, and risk management programs. More recently, new approaches such as continuous quality improvement have deemphasized individual review, which is thought to be counterproductive since it fuels adverse relationships among peers. Nevertheless, peer review is still recognized as a major tool for quality assessment, be it to detect substandard care or to provide clues about opportunities for process improvement.

Measurement of quality of care requires that it be translated into more concrete representations that lend themselves to quantification. For the purposes of quality assessment, Donabedian (1980) defined these representations as the criteria and standards of structure, process, and outcome. Medical audit of hospital records is mainly based on process of care. A major limitation of process-based assessment is the difficulty in designing standards of care. Efforts have been made to develop practice guidelines and to incorporate them into *explicit* criteria to guide physicians' decision making. However, written guidelines have not been able to cover the infinite variety of situations that may arise in clinical practice (Chassin, 1990; Kassirer, 1993). Furthermore, scientific evidence of effectiveness of a specific procedure or intervention is often unavailable. Therefore, process assessment often relies solely on *implicit* criteria, that is, on the judgment of an expert. This is a convenient approach since no additional effort is needed. However, physician agreement regarding quality of care has been shown in several studies to be poor (Richardson, 1972; Goldman, 1992).

Reliability, also called precision and reproducibility, is a basic issue to be addressed in the development of any measurement and in dealing with misclassification problems. Reliability refers to the stability or consistency of measurements, that is, the extent to which repeated measurements of the same subjects produce the same results. Some variability in the results of any measurement process can be expected, and can be classified in three major sources: instability of the attribute being measured, criterion variation within and among observers, and poor calibration of the instrument. These sources of variation can be more or less controlled through standardization of measurement procedures and validation of instruments, which can be regarded as a form of calibration (Dunn, 1989). This is harder to achieve when the measurement involves some sort of judgment as is the case with medical assessment of quality of care. Reliability of medical audit can be "boosted" if (1) physicians undergo some sort of training, (2) a narrower aspect of care is defined, or (3) the findings of multiple physicians for the same case are combined by consensus or by some scoring method (Palmer, 1991).

We conducted a study to evaluate the performance of screening methods used to select cases with likely quality problems in hospital care. The quality of health care assessed through medical audit of hospital records was taken as a reference to measure accuracy of quality screens. A major credential for a reference is its own validity, which is hard to measure in health care because consensual standards are often unavailable. Physicians' judgment of quality of care is usually accepted for its face validity only. Reliability as a condition for validity is also a credential for a reference to the performance of screening methods. The subject of this article is the reliability of physician judgment of quality of care based on medical record reviews. We believe the theoretical and practical issues brought up in this investigation pertain not only to formal and systematic audit but to any sort of assessment of quality of care.

Methods

Research setting and subject selection criteria

A random sample of medical records from an American tertiary teaching hospital was selected from hospitalizations that occurred between July 1, 1989 and June 30, 1991. We selected cardiac conditions, namely, percutaneous transluminal coronary angioplasty (PTCA), coronary artery bypass graft (CABG), and myocardial infarction (MI) without a revascularization procedure, because they constituted an

important component of morbidity and mortality, and covered a wide spectrum of medical and surgical patients. Moreover, the severity of illness and the intrinsic risk of the interventions they involved made them more prone to the events we were interested in. We oversampled medical records that had failed routine screening by the hospital's Quality Assurance Department in order to obtain more cases with substandard care. Adding up all strata, we aimed for a random sample of 13% of records.

Physician review

Medical record reviews were conducted by cardiologists using a structured implicit review form and its guidelines, designed by Rubin et al. (1990) for all medical and surgical conditions. The implicit review form asked physicians to rate on a five-point scale a set of specific aspects of the process of care. The elements included in the form are those thought to be more relevant for quality assessment, and readily available and reliably recorded in medical records: admitting workup, use of tests and consultants, treatments prescribed, surgical and invasive procedures carried out, and patient follow-up. This way, physicians are led through the main parts of the medical record before assigning the overall score of quality, which "wraps up" the evaluation. The overall score is a five-point ordinal scale that ranks the quality of care from 1 (extreme, below standard) to 5 (extreme, above standard).

A second section of the form, adapted from the Adverse Event Analysis Form (Brennan et al., 1989), had a more outcome-oriented approach. Specifically, (1) the form ascertained whether there was an *injury* defined as morbidity that was not expected from the disease process; (2) determined whether the injury was an *adverse event*, that is, whether the injury had been caused by medical management; and (3) asked for a judgment of whether it was due to *negligence*, defined as failure to meet the average practitioner's standard of care. Upon completion of these, physician-reviewers were then unblinded to the results of nurse screening, and asked to comment on the quality issues raised.

All reviewers underwent a training process to ensure that the items of the form were understood in the same way by all physicians, so that differences in ratings would not represent different interpretations of words. Written guidelines provided basic definitions and yardsticks for the ratings (Rubin et al., 1990). For instance, "extreme, above standard" medical care was defined as the best care one could think of in a U.S. hospital at the time the care was given. It minimized the risk of complications, maximized the likelihood of a good outcome, and maximized humane care and respect for patients' wishes. "Extremely below standard" was malpractice. It was more likely to result in harm than benefit to the patient. "Standard" care was just acceptable. The guidelines expanded on all items, providing anchor-points for ratings as well as examples.

Data collection and analysis

Each medical chart was independently reviewed by two physicians. Interrater agreement of physician review was measured for review items selected for their relevance. For each item we calculated a crude (proportion) agreement and a chance-corrected agreement (Kappa [k] statistic) (Cohen, 1960). Besides being appropriate to our data set composed mostly of categorical variables, Kappa is a well-known and widely applied statistic, easy to compute and to interpret, and mathematically equivalent to the intraclass correlation coefficient, which is a true measurement of reliability. The Kappa statistic was computed with the software PC-AGREE (McMaster University, Hamilton, Ontario, Canada). These agreement measurements express the reliability of a single measurement. If we combine results of two or more observers by taking the arithmetic mean, or by considering only results in which the observers converge, the reliability will be enhanced. In other words, a quality problem agreed upon by two reviewers is more reliable than those found by one reviewer only. Accordingly, the estimates of reliability were corrected as proposed by Kraemer (1979):

$$K' = \frac{rK}{1 + (r-1)K}$$

where K is the reliability coefficient for one observation and r is the number of replicates.

We followed the conventional guidelines proposed by Landis & Koch (1977) to interpret Kappa values :

≤ 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
≥ 0.81	Almost Perfect

We also looked at the internal consistency of the implicit review form, that is, the extent to

which its items were addressing different aspects of the same attribute, namely, quality of care. The Cronbach's alpha coefficient was the correlation coefficient used to measure the degree of homogeneity of answers to sets of questions (Streiner & Norman, 1989). The items of the form were supposed to approach complementary aspects of quality of care. Therefore, the answers were expected to be correlated with each other.

Sample selection

Our sample was stratified with respect to two variables that might be related to quality problems and to the validity of screening systems: (1) the principal diagnosis and the main procedure carried out during hospitalization, and (2) the result of previous routine screening performed by the Hospital's Quality Assurance (QA) Department. Sample size calculations were primarily directed to estimate sensitivity and specificity. A total of 405 medical charts was found to provide enough precision for agreement estimates as well.

Results

Data from 423 medical charts (82% of all charts we requested) were available for analysis. Patients were mostly white males, with a mean age of 63 years (Table 1). The rate of complications and in-hospital deaths was consistent with the high-risk profile of the patients brought to this particular hospital, which is a renowned referral center.

Contrasting the findings of a physician reviewer responsible for 41% of the reviews with his "review-mates" in selected items of the review (Table 2), chance-corrected agreement was shown to range from slight to moderate. Discordant cells in the contingency tables suggested that the reviewer was more lenient in the judgment of the standard of care, but found more injuries, adverse events, and negligence than his review-mates (data not shown). Other reviewers showed similar patterns of asymmetry in discordant cells for some review items, suggesting systematic differences among reviewers.

Interrater agreement was also assessed by pooling data from all reviewers so as to contrast reviews (Table 3). The results show a pattern similar to that of table 2 which was not unexpected given the predominance of one reviewer. Prevalence rates of events agreed upon by two reviewers are also shown, since they help explain the apparent discrepancies between crude and chance-corrected agreement. Table 3 also shows a substantial increase in reliability estimates when results agreed upon by two reviewers are considered.

Interrater agreement within conditions is summarized in figure 1 to better visualize trends and patterns. Chance-corrected agreement (Kappa), was highest for injuries and lowest for substandard care. For most items, the

Table 1

Selected descriptive measurements of the study sample and the hospital course.

	CABG N = 215	PTCA N = 146	MI N = 90	All Conditions N = 451
1) Mean-age (years)	65.4	60.7	63.3	63.3
2) Males (%)	64.9	58.3	67.2	62.4
3) Whites (%)	88.8	81.9	52.4	82.6
4) Average length of stay (days)	12.9	5.5	8.8	9.5
5) Principal diagnosis				
i) Coronary atherosclerosis (%)	94.4	56.2	–	69.1
ii) Myocardial infarction (%)	3.2	28.1	100.0	20.9
6) Adverse occurrences (Rate per 100 patient-days) *	5.8	5.7	5.0	5.6
i) Pulmonary edema (%)	5.5	6.5	6.2	6.0
ii) Cardiac arrest (%)	5.2	0.8	11.1	3.9
iii) Pneumonia (%)	6.1	0.3	3.9	3.5
iv) Ventricular tachycardia (%)	3.5	3.2	3.2	3.3
7) In-hospital death (%)	6.0	1.0	15.0	4.8

* Only the four most frequent types of adverse occurrences are shown.

highest agreement was obtained in PTCA cases and lowest in CABG cases.

The implicit review form as a whole showed a satisfactory internal consistency as measured by the Cronbach's coefficient alpha (Table 4), although for some modules of the form it fell below the conventional limit of 0.70. This means that scores assigned to items were coherent throughout the review.

Table 2

Agreement (crude and chance corrected) between one reviewer and his reviewmates on selected items of review.

Review item	Crude agreement %	Chance-corrected agreement (Kappa)
Below standard care	78	0.10
Injury	84	0.61
Adverse event	74	0.48
Negligence	91	0.20

Table 3

Percent agreement and chance-corrected agreement (Kappa statistic) between physician reviews, for selected items; reliability and prevalence of events agreed upon by two physicians.

Review item	Crude Agreement (%)	Kappa		Prevalence (%)
		Single review	Combined* results	
Substandard care	79	0.11	0.20	2.8
Injury	78	0.58	0.73	34.3
Adverse event	73	0.40	0.57	22.2
Negligence	91	0.39	0.56	3.6
Quality problem	68	0.29	0.45	17.8

* Reliability estimates corrected for the fact that only results agreed upon by two reviewers were counted (see Methods).

Discussion

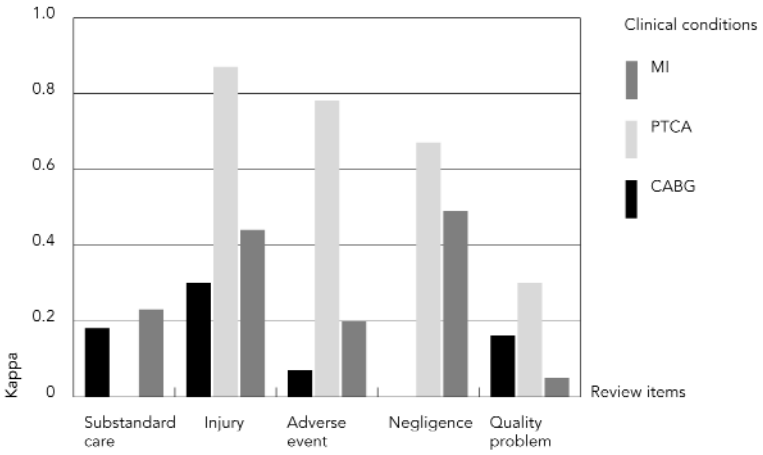
Historically, physicians' perspective of quality in health care has dominated, be it in setting standards for training and licensing practitioners, or accrediting health care organizations. Even though consumers and payers seem to be increasingly influential in setting standards of care, physicians' judgment prevails as far as quality assessment of the process of care is concerned. This is true despite the fact that physician judgment has been found to be notoriously unreliable (Goldman, 1992). A variety of approaches have been applied to improve agreement among physicians, but results obtained in different settings (ours included) where structured review was used were only able to achieve moderate reliability for some of the items (Table 5). The discrepancies in reliability measurements obtained by those studies might have resulted from differences in methods and case-mix as well as from the frequency of the events of interest. None of the studies in table 5 provide information about the internal consistency of review items, which is an important piece of evidence in favor of (or against) the quality of data collection.

Our data suggest that reliability of medical judgment of quality of care is influenced by the diagnosis and the intervention. For instance, quality of care in surgical patients is more difficult to assess from information recorded in a medical chart, as surgical reports usually do not provide sufficient details of the procedure to allow detection of breaches in quality. In this study, most quality issues raised in surgical patients concerned the postoperative period. Patients' complaints and objective clinical and laboratorial abnormalities are recorded by health professionals other than the attending physician, making it harder to conceal errors of commission and of omission. Verbatim comments by physician-reviewers reveal the difficulties in judging from incomplete evidence. In the absence of unequivocal data, some reviewers might adopt a more cautious approach when judging the quality of care of cases, whereas others might use their previous experience in similar cases to make inferences.

Awareness of the outcome is known to in-

Figure 1

Chance-corrected agreement (Kappa) on selected items of physician review within clinical conditions.



fluence the assessment of the process of care (Caplan et al., 1991). We tried to minimize this effect by using a structured review, by separating the evaluation of process and outcome of care in different forms, with the process assessment coming first, and emphasizing during physician training the need to avoid hindsight. Adequacy of process is often phrased in terms of maximizing benefits and avoiding harm and other untoward outcomes. Therefore, physicians were asked to base their reasoning on potential outcomes, while avoiding considering the actual ones. Hindsight in process of care assessment had unpredictable effects on reliability estimates.

The Kappa statistic implies a definition of chance agreement based on fixed table marginals, which has a significant impact on the magnitude and interpretation of this agreement index (Brennan & Prediger, 1981). Such an agreement attributed to chance may actually result from expert judgment rather than chance. However, raters do not get any credit for this, since chance agreement gets higher as the agreement on the marginals increases, and the agreement rate necessary to obtain the same Kappa is also higher. As the prevalence rate approaches either 0% or 100%, the population becomes more homogeneous, and the same number of diagnostic disagreements will have a greater impact on the unreliability of the diagnosis. This is known as the "base rate problem", which is the difficulty in obtaining reliable diagnoses in homogeneous (low prevalence) populations. Although indexes such as kappa give smaller values as the prevalence approaches zero for fixed levels of diagnostic errors (fixed sensitivity and specificity), some argue that this reflects the real problem of making distinctions in a homogeneous setting (Shrout et al., 1987). According to this interpretation, rather than a limitation, weighing disagreements more when the prevalence approaches 0% or 100% is a major strength. However, the problem of making comparisons of Kappa statistics obtained from different base rates still remains (Uebersax, 1987; Spitznagel & Helzer, 1985; Kraemer, 1987). It could be partly addressed by having some indication of the prevalence and the crude agreement that could enable the reader to better compare results from different settings.

Considering the low frequency of substandard care, most hospitals are characterized by homogeneous settings, according to the reasoning above. However, medical audit usually follows screening of medical records for sentinel events regarded as "markers" of potential quality problems. This selection results in a set of medical records with a higher frequency of quality problems. To the extent that the proportion of medical records with quality problems approaches 50% (higher heterogeneity) among those records failing screening, chance-corrected agreement is strengthened.

Previous and current data suggest that regardless of the setting in which reliability was measured or the approach used, there appeared to be some intrinsic limitation to the reliability of peer assessment of quality of care. This should not be surprising considering the poor physician agreement on objective clinical measurements, such as EKG and roentgenogram, and on diagnosis and treatment (Koran, 1975a, 1975b). However, since physician judgment is generally considered the best reference available for quality assessment, several strategies have been developed to improve its reliability (Goldman, 1992). We used three of them, namely, multiple reviews, reviewers specializing in cardiology who had extensive and recent inpatient experience, and structured or guided assessment. Independent multiple reviews can control unreliability through replication of measurements and use of their mean value. It is known that the mean of several independent measurements is more reliable than a single measurement (Fleiss, 1986). Multiple reviews can also consist in the reexamination of cases on which there is initial disagreement, so as to reach a consensus (Dubois & Brook, 1988).

An improvement in interrater agreement that can be obtained by including outcomes was apparent from our data, and may also have contributed to the results of Brennan et al., (1989 and 1991). Our results showed that the assessment of negligence was more reliable than the assessment of standard of care, which was also a process measurement, but was not linked to an outcome as was negligence (according to our definition). Detection of adverse events was even more reliable and could be more useful for quality assurance since it avoids the issue of culpability implied by detection of substandard care or quality problems. In other words, the presence of an adverse event should trigger quality improvement actions whoever the culprit might be. Another strategy to enhance the reliability of peer assessment of care quality would be to raise standards for selection of peer reviewers. Brook & Lohr (1986) proposed that peer review should be conducted by acknowledged experts, not only in their specialty, but also in quality assessment techniques. The use of practice guidelines when available has also been proposed as a means to set standards for optimum care that are ideally based on scientific evidence, or at least agreed upon by most practitioners (Chassin, 1990).

Possible consequences of low reliability are attenuated correlations, lower power for statistical significance tests (or need of larger sample sizes), biased prevalence estimates (usually overestimation), and estimates of the strength of associations biased toward the null (Shrout et al., 1987; Fleiss, 1986). If the reliability of a measurement is poor, its validity will be negatively affected. Because there is no absolute criterion of truth in quality of health care, the best we could do was to approximate the optimum reference, or at least decrease the probability of using some idiosyncratic criterion as a reference, by using the cases with a quality problem agreed upon by two physicians. It is intuitive and can be mathematically demonstrated that judgments on which two or more experts converge are, on average, more reliable.

Intrarater reliability was not assessed in this study, but that does not mean that consistency of physicians' own judgments was taken for granted. Rather, it was considered that intrarater agreement was usually higher than the

Table 4

Internal consistency of the implicit review form (global and by modules).

Module	Cronbach's Coefficient Alpha
Hospital admission data collection and assessment	0.81
Tests and treatments	0.66
Components of hospital care	0.88
Surgery	0.85
Effects on outcome	0.68
All the above combined	0.71

Table 5

Reports in the literature of chance-corrected agreement of physicians' assessment of health care.

Authors	Chance-corrected agreement	Events of interest; type of review
Dubois et al., 1987	0,11 – 0,55	Preventable deaths. unstructured reviews; case-mix: CVD, MI and Pneumonia
Brennan et al., 1989	0,34 – 0,57	negligence and adverse events; case-mix: not specified.
Rubenstein et al., 1990	0,42 – 0,66	substandard care; structured review; case-mix: CVD, MI and Pneumonia
Bates et al., 1992	0,24 – 0,61	negligence and adverse events; case-mix: not specified.
Rubin et al., 1992	0,49	substandard care; structured review;
Hayward et al., 1993	0,1 – 0,5	Preventable deaths; quality of follow-up, etc.

interrater agreement (Streiner & Norman, 1989). Conceptually, interrater agreement could be considered to 'contain' intrarater reliability, that is, inconsistencies of reviewers contributed to disagreement among them. We considered that training, standardization of procedures, and use of a structured form, addressed intrarater agreement as well as interrater agreement.

A major limitation of this study is its external validity. We examined medical records of patients with three clinical conditions, including two procedures performed only in tertiary hospitals. The sample examined came from a teaching hospital that is a national and international referral center. The frequency and kind of quality problems that may arise from the care of more complex cases, with more invasive techniques, and involving professionals under training, as is probably the case in most teaching hospitals, justifies some caution in generalizing our results.

An inherent limitation of most studies of quality assessment is their reliance on medical records. The problems with this source of data are well recognized: reports may conceal rather than disclose substandard care; evidences of quality problems provided are usually incomplete and inconclusive of substandard care; medical record review is labor-intensive and expensive. O'Neil and coworkers (1993) have shown that medical records missed many adverse events that were reported by physicians. Despite its drawbacks, the medical record is the best single source of data for quality assessment, and probably the most frequently used for this purpose because of its comprehensiveness and ease of access. Brennan et al. (1990) showed that most adverse events that resulted in malpractice claims could be identified in a medical record review. Moreover, very few adverse events and episodes of negligence they missed were due to deficiencies of the medical record. All the considerations above should not substantially affect interrater agreement estimation, since each pair of reviews was based on the same records.

Medical-record-based assessment of quality of health care by physicians has been accepted for its face validity, since reference criteria are usually unattainable or controversial. Reliability of medical audit of hospital charts should be pursued by (1) improving upon current structured review instruments, (2) applying explicit criteria preferentially based on practice guidelines, (3) combining reviews of several reviewers, and (4) using outcome-oriented criteria.

References

BATES, D. W.; O'NEIL, A. C.; PETERSON, L. A.; LEE, T. H. & BRENNAN, T. A., 1995. Evaluation of screening criteria for adverse events in medical patients. *Medical Care*, 33:452-462.

BRENNAN, R. L. & PREDIGER, D. L., 1981. Coefficient Kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41:687-699.

BRENNAN, T. A.; LOCALIO, R. J. & LAIRD, N. M., 1989. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. *Medical Care*, 27:1148-1158.

BRENNAN, T. A.; LOCALIO, A. R.; LEAPE, L. L.; LAIRD, N. M.; PETERSON, L. A.; HIATT, H. H. & BARNES, B. A., 1990. Identification of adverse events occurring during hospitalization. A cross-sectional study of litigation, quality assurance, and medical records at two teaching hospitals. *Annals of Internal Medicine*, 112:221-226.

BRENNAN, T. A.; LEAPE, L. L.; LAIRD, N. M.; HEBERT, L.; LOCALIO, A. R.; LAWTHERS, A. G.; NEWHOUSE, J. P.; WEILER, P. C. & HIATT, H. H., 1991. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *New England Journal of Medicine*, 324:370-376.

BROOK, R. H. & LOHR, K. N., 1986. Will we need to ration effective health care? *Issues in Science and Technology*, 3:68-77.

BUTLER, J. J. & QUINLAN, J. W., 1958. Internal audit in the department of medicine of a community hospital. Two years' experience. *Journal of the American Medical Association*, 167:567-572.

CAPLAN, R. A.; POSNER, K. L. & CHENEY, F. W., 1991. Effect of outcome on physician judgments of appropriateness of care. *Journal of the American Medical Association*, 265:1957-1960.

CHASSIN, M. R., 1990. Practice guidelines: best hope for quality improvement in the 1990's. *Journal of Occupational Medicine*, 32:1199-1206.

COHEN, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37-46.

DONABEDIAN, A., 1980. *Explorations in Quality Assessment and Monitoring*, Vol. 1, *The Definition of Quality and Approaches to its Assessment*. Ann Arbor: Health Administration Press.

DUBOIS, R. W.; MOXLEY III, J. H.; DRAPER, D. & BROOK, R. H., 1987. Hospital inpatient mortality. Is it a predictor of quality? *New England Journal of Medicine*, 317:1674-1680.

DUBOIS, R. W. & BROOK, R. H., 1988. Preventable deaths: who, how often, and why ? *Annals of Internal Medicine*, 109:582-589.

DUNN, G., 1989. *Design and Analysis of Reliability Studies. The Statistical Evaluation of Measurement Errors*. London: Edward Arnold.

FLEISS, J. L., 1986. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, Inc.

GOLDMAN, R. L., 1992. The reliability of peer assessment of quality of care. *Journal of the American Medical Association*, 267:958-960.

HAYWARD, R. A.; MCMAHON JR, L. F. & BERNARD, A. M., 1993. Evaluating the care of general medicine inpatients: how good is implicit review? *Annals of Internal Medicine*, 118:552-556.

KASSIRER, J. P., 1993. The quality of care and the quality of measuring it. *New England Journal of Medicine*, 239:1263-5.

KORAN, L. M., 1975a. The reliability of clinical methods, data and judgments. Part I. *New England Journal of Medicine*, 293:642-646.

KORAN, L. M., 1975b. The reliability of clinical methods, data and judgments. Part II. *New England Journal of Medicine*, 293:695-701.

KRAEMER, H. C., 1979. Ramifications of a population model for k as a coefficient of reliability. *Psychometrika*, 44:461-472.

KRAEMER, H. C., 1987. Charlie Brown and statistics: an exchange (letter). *Archives of General Psychiatry*, 44:192-193.

LANDIS, J. R. & KOCH, G. G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:766-771.

LEMBCKE, P. A., 1967. Evolution of the medical audit. *Journal of the American Medical Association*, 199:111-118.

O'NEIL, A. C.; PETERSEN, L. A.; COOK, E. F.; BATES, D. W.; LEE, T. H. & BRENNAN, T. A., 1993. Physician reporting compared with medical-record review to identify adverse medical events. *Annals of Internal Medicine*, 119:370-376.

PALMER, R. H., 1991. Considerations in defining quality of care. In: *Striving for Quality in Health Care. An Inquiry into Policy and Practice*. Ann Arbor: Health Administration Press.

RICHARDSON, F. M., 1972. Peer review of medical care. *Medical Care*, 10:29-39.

RUBENSTEIN, L. V.; KAHN, K. L.; REINISCH, E. J.; SHERWOOD, M. J.; ROGERS, W. H.; KAMBERG, C.; DRAPER, D. & BROOK, R. H., 1990. Changes in quality of care for five diseases measured by implicit review, 1981 to 1986. *Journal of the American Medical Association*, 264:1974-1979.

RUBIN, H. R.; KAHN, K. L.; RUBENSTEIN, L. V. & SHERWOOD, M. J., 1990. *Guidelines for Structured Implicit Review of the Quality of Hospital Care for Diverse Medical and Surgical Conditions*. A RAND Note, N-3006-HCFA. Santa Monica: Health Care Financing Administration.

RUBIN, H. R.; ROGERS, W. H.; KAHN, K. L.; RUBENSTEIN, L. V. & BROOK, R., 1992. Watching the doctor-watchers. How well do peer review organization methods detect hospital care quality problems? *Journal of the American Medical Association*, 267:2349-2354.

SHROUT, P. E.; SPITZER, R. L. & FLEISS, J. L., 1987. Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry*, 44:172-177.

SPITZNAGEL, E. L. & HELZER, J. E., 1985. A proposed solution to the base rate problem in the Kappa statistic. *Archives of General Psychiatry*, 42:725-728.

STREINER, D. L. & NORMAN, G. R., 1989. *Health Measurement Scales. A Practical Guide to Their Development and Use*. New York: Oxford University Press.

UEBERSAX, J. S., 1987. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101:140-146.