INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

MEASURING THE INTERRATER RELIABILITY OF A DATA COLLECTION INSTRUMENT DEVELOPED TO EVALUATE ANESTHETIC OUTCOMES

by

KAREN CRAWFORTH

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2001

MAJOR: EVALUATION AND RESEARCH
(Education)

Approved by:

Thomas S. Jaworsky 5/7/01
Advisor Date

Prudencia Worth 5/7/01

5/7/01

5/7/01

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

DEDICATION

The completion of this dissertation would not have been possible without the support and encouragement of my family. I am grateful to my father who spent countless hours computerizing the tables and charts for this manuscript. I am thankful to my mother for her continued inquiries regarding my time frame for bringing this project to closure. I am deeply indebted to my husband and children for their patience and understanding over the course of the past seven years.

ACKNOWLEDGMENTS

Completion of this educational endeavor was accomplished with the support and encouragement of many people. I would like to acknowledge those individuals who enabled me to achieve this goal. I would like to thank the members of my dissertation committee, including Dr. Sawilowsky as chairman, Dr. Marcotte, and Dr. Sikonolfee. I would like to thank Dr. Worth for her participation as a committee member and her understanding as my colleague and director. I am grateful for her confidence in my ability to complete this endeavor. I am indebted to my colleagues at Detroit Receiving and Hutzel Hospitals for their patience and numerous schedule changes that allowed me the time to accomplish this project. I am particularly thankful to Kathy Cook for her friendship and encouragement. She selflessly donated countless hours to the review and editing of information related to this document. Her constructive criticism improved the clarity and scope of this dissertation. I wish to express my sincere thanks to Dr. Hockman for her assistance in organizing and analyzing the data for this research. The quality of this project was enhanced by the contributions of all of these individuals.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Introduction

Reliability is the degree to which repeated measures or measuring procedures lead to the same results. Studies that rely on humans to obtain measurements involve estimates of agreement known as interrater reliability. This measure is used when opinions are required of multiple raters evaluating or interpreting the same behavior. The amount of error that may be attributed to the rater can be determined by various statistical techniques. The traditional psychometric methods employed to measure interrater reliability include correlational techniques, contingency tables, percentage of agreement, and the comparison of means. The application of an analysis of variance (ANOVA) within the context of generalizability theory is also used to estimate the dependability of behavioral measurements. The information provided by these techniques allows an evaluation of rater error that occurs during the extraction and interpretation of data.

The need to establish interrater reliability is important for instruments developed to evaluate the quality of medical care provided to patients. The rating scales used in these processes require the rater to make a judgment about the patient's care based on a review of their medical records. The reviewer is then required to assign a score to the characteristics being measured based on their interpretation of the event. There are many potential sources of measurement error that may occur during this type of scoring process. Inconsistency in scoring, bias, and idiosyncratic use of the instrument by raters will affect the reliability of the scores. Establishing the degree of interrater reliability of a data collection instrument is a necessary condition before validity can be ascribed to the obtained set of ratings.

Historically, traditional psychometric methods have been inadequate to address the reliability of ratings regarding clinical decisions. Significant information regarding these complex issues is concealed when the data is summarized by simple correlation coefficients. The impetus behind the evolution of the theory of generalizability (GT) was the inability of the classical model to recognize the multiple sources of error in situations when several raters measure individuals on various attributes and on different occasions. Determining the degree of reliability in a medical peer review process requires statistical analyses that can identify the multifaceted error inherent in this complex situation. Generalizability theory may supplement traditional methodology by analyzing error in terms of multiple factors and interactions.

Measuring interrater reliability is particularly significant in anesthesia morbidity and mortality because individual interpretation of events is an integral part of the process. Previous studies have demonstrated the reliability of instruments developed to measure the quality of anesthesia care only fair to good. Goldman (1992) found that few reliability studies based on implicit criteria use sound statistical techniques. Previous interrater reliability studies in anesthesia by Caplan, Posner, Ward and Cheney (1988), and Caplan, Posner, and Cheney (1991) analyzed data extracted from summaries of closed claim files. There are several potential problems associated with abstracting data from the original documents. In negligence cases crucial details that are provided in the closed claim files may not be included in the summary. The individuals responsible for removing the data from the original source may inadvertently instill their own bias into the condensed version. The admission that the summarization process itself represents a possible source of measurement error was identified as a limitation in both these studies.

Caplan et al. (1991) recognized that if the abstracted data contains insufficient information the reviewer is unable to recreate the entire process that was involved in the adverse outcome. The studies of Nobrega, Morrow, Somolt and Offord (1973), Fessel and Brunt (1972), and Brook (1977) corroborate that judgments based on outcomes do not correlate well with judgments based on performance information. The reliability studies by Caplan et al. (1988) and Caplan et al. (1991), which were based on summaries cannot be generalized to judgments regarding closed claim files without a demonstration that these two sources are interchangeable.

In a subsequent study by Posner, Caplan and Cheney (1996) the original documents were analyzed, but agreement was measured only on a single question; "Was the anesthesia care rendered during the adverse event appropriate?" The overall concordance on the appropriateness of care in this study was 62%. This percentage included agreement on the impossible to judge category an inclusion that may spuriously increase the level of concordance. The selection of this category by raters does not necessarily mean they are in agreement regarding the quality of anesthesia care, it simply means they could not decide. The kappa value on this item was 0.37, a level that exceeds the agreement that would be expected on the basis of chance only in the fair to moderate range. The previous study by Caplan et al. (1988) revealed the presence of disparate scoring patterns by some raters. The analysis of only one item by Posner et al. (1996) precluded the identification of incongruous scoring patterns when original closed claim documents are utilized.

The study by Posner et al. (1996) was completed over a period of time spanning from December of 1988 to October of 1994. This six- year period of time was marked by

some of the most significant technologic advances in anesthesia safety to date. The pulse oximeter and capnograph were introduced in the mid-1980s and were adopted as standards of practice in early 1990's. Because of the time lag associated with legal processes the events represented in the claims analyzed in this study may have actually transpired as far back as 1978. The incongruity between events that occurred nearly 20 years ago and current anesthesia practice may have contributed to the disparate opinions rendered by the reviewers.

The Purpose of this Study

The purpose of this study is to evaluate the interrater reliability of an instrument (Appendix A), developed to evaluate the quality of anesthesia care provided during adverse events. The instrument was devised to extract demographic, nominal, and ordinal scale data from closed claim files. Permission to conduct this study and use materials related to the closed claim file analysis was provided by the Director of Research of the American Association of Nurse Anesthetists Foundation (AANAF) (Appendix B). The impetus behind the development of this tool was the identification of closed claim research as a means to improve the quality of anesthesia care. The study of anesthesia related sentinel events was initiated by the AANAF in 1995 after approval was obtained from the St. Paul Insurance Companies to review their closed claim files.

The instrument was initially evaluated for reviewer agreement using three randomly selected claims. The three files represented different outcomes and settlements. Interrater reliability was determined using percentage of agreement no other statistical analyses were applied to the data. Agreement on the subjective items ranged from 62-79%, and the overall concordance based on this descriptive measure was 72%. Additional

revisions in the data collection process were made based on these measurements. The instrument has not been re-evaluated for interrater reliability since these modifications were made.

The rater agreement on the AANF closed claim items was similar to the level of concordance on data derived from the American Association of Anesthesiologists (ASA) Closed Claims Project. This comparable project is a structured evaluation of anesthetic mishaps that was initiated in 1985 by the ASA. This study of adverse anesthesia outcomes is based on information extracted from the closed claim files provided by 34 professional liability insurance companies throughout the United States. A study conducted by Caplan et al. (1988), measured the interrater reliability of the judgments rendered in the ASA Closed Claims Project. This study found a 54-80% agreement on the subjective data regarding the quality of anesthesia care.

The conclusions derived from the studies of interrater reliability using data from the ASA project support the incorporation of multiple items, current practices, and the actual documents in the study. The application of multiple statistical techniques to the interrater reliability data derived from the closed claim files will be integrated into the methodology of this research project based on the recommendations of related studies. The measurement techniques selected for this study are founded on a review of literature pertaining to the measurement of agreement using dichotomous data. Generalizability theory will also be used to analyze interrater reliability by determining the effect of outcome and provider type on judgments regarding the quality of care. The information gleaned from this research will be utilized to improve the reliability and efficiency of the data collection instrument.

Research Questions

The research questions for the study are as follows:

Research Question I.

What is the reliability of an instrument developed to extract data from anesthesia related closed claims using kappa, correlational techniques, percentage of agreement and generalizability theory?

1. What is the overall interrater reliability of the group of raters?
2. Is there a relationship between the reviewer's rating regarding appropriateness of care and the severity of the patient's injury?
3. Is there a relationship between the subjective ratings and the provider type?
4. What is the reliability of the research team of raters when compared to a designated gold standard?
5. Do different measures of reliability produce consistent results when applied to the data?
6. Based on the characteristics of the interrater reliability data would other statistical techniques be more appropriate?
7. Is there an advantage of applying multiple statistical measures of agreement to the same data?
8. Do all the items demonstrate comparable levels of agreement?
9. Are there individual raters who demonstrate disparate scoring patterns?
10. Do all the claims demonstrate comparable levels of agreement?
11. If certain claims demonstrate different levels of agreement can the reason for this difference be explained?

Research Question II.

Does generalizability theory provide an understanding of the data derived from anesthesia related closed claims that is not apparent using traditional psychometric approaches to reliability?

Research Question III.

If sources of rater variance are identified in this analysis, will they be factors that can be controlled in a manner that will impact on the future reliability of the data collection process?

The Significance of the Research

Preventable adverse events were identified as a leading cause of death in the United States by the Department of Health and Human Services. The American Hospital Association (1998) estimated that as many as 98,000 Americans die in hospitals each year as a result of medical errors. The total national costs associated with these preventable adverse events are estimated to be between 37.6 and 50 billion dollars. The primary purpose of the American Association of Nurse Anesthetists (AANA) study of closed claim data is to improve patient safety. The outcome data derived from the files is to be used to:

1. Identify major areas of anesthesia related patient injury.
2. Identify negative trends in patient outcome.
3. Provide data that can be used to design strategies to improve patient safety.
4. Identify data that should be incorporated into educational programs and curriculums.
5. Provide data that can be used to recommend changes in the AANA Anesthesia Practice Standards and Guidelines.

The anticipated benefits of this study are that it will yield recommendations for anesthesia training and continuing education programs, and will advance the body of knowledge regarding anesthesia patient safety. The reliability of the process used to extract data and generate conclusions regarding the adverse event needs to be demonstrated before confidence can be placed in the data. The purpose of measuring interrater reliability in this situation is to demonstrate that the judgments rendered were not the idiosyncratic opinion of the raters, but an opinion that could be generalized to the population of anesthesia providers.

The credibility of results obtained by a peer review process is significantly diminished if only fair agreement occurs between reviewers. The application of generalizability theory and traditional psychometric methods of analysis to the reviewers' judgments will facilitate the identification of error sources that diminish interrater reliability in this situation. The results of these findings will be utilized to develop a more reliable and cost efficient instrument for data collection.

Assumptions and Limitations of the Study

The Assumptions of this Study Are:

1. The raters represent a sample of Certified Registered Nurse Anesthetists (CRNAs) who possess similar characteristics to the population of nurse anesthetists.
2. The claims represent a sample that includes a variety of incidents and severity levels.
3. Each closed claim file is independent of other claims.
4. The raters operate independently of one another.

5. The categories on the data collection instrument are mutually exclusive and exhaustive.
6. The information contained in the closed claim files is adequate to allow the reviewer to complete the data collection instrument.

The Limitations of this Study Are:

1. Closed claims represent only events in which legal action was initiated. Therefore, they represent a small fraction of all adverse outcomes and they may not reflect the population of anesthesia related adverse outcomes.
2. This study is dependent on retrospective data collection. The accuracy and completeness of the records from which the data are derived is unknown. Critical information may be missing, notes illegible and data needed to clarify problems may not be available. Information from the participant's reviews and expert testimony may be contradictory.
3. The reviewers were not randomly selected. This same group was responsible for the development of the data collection instrument and instruction guide. The external validity of their judgments has not been demonstrated.
4. The geographic origins of the claims are dependent on the territory covered by the insurance carrier, and may not represent a cross sectional representation.
5. The total number of anesthetics administered by the providers included in the closed claim collection is unknown as well as the total number of adverse events that occur.

6. The anesthesia providers named in the claim represent only those CRNAs who are covered by the insurance company.
7. No denominator is available for calculating the risk associated with anesthesia utilizing the closed claim database.
8. Causality for the anesthetic mishaps represented in the claims can only be inferred not proven.
9. The information in the files was compiled for the purpose of resolving the claim and not for patient safety research.
10. There may be a lag in time as much as 10 years between the event and the closing of a claim. The relationship of these past events to current practice may be a limitation.

Definition of Terms

The following definitions have been developed for the purpose of this study:

1. A medical malpractice claim is a demand for financial compensation by an individual who has sustained an injury as the result of medial care.
2. Human error is a mistake in judgment, a lack in vigilance or an inappropriate response to a change in the patient's status.
3. Standards of care are the level of care that a reasonable and prudent practitioner would provide while the incident took place.
4. Appropriate care refers to anesthesia care that met the standards of practice, when the adverse event occurred.
5. The damaging event is the specific incident that led to the adverse outcome.

6. Inappropriate care refers to anesthesia care that did not meet the current standards of practice when the adverse outcome occurred.
7. Malpractice is a deviation or noncompliance with accepted standards of care that results in mental or physical damages to a patient.
8. A sentinel event is an unusual or unexpected outcome that occurs despite clinical knowledge and expertise that should have prevented the incident.
9. A closed claim is a file that has been resolved either by an out of court process or by litigation.
10. A typical closed claim includes medical records, expert review, depositions, outcome and follow-up reports, and the cost of the settlement or jury award.
11. An outcome is the patient's response to the care provided.
12. The adverse outcome is the injury (physical or mental) that is sustained by the patient as a result of the medical management.
13. The medical process includes what a health care provider does on behalf of a patient including diagnostic and therapeutic interventions.
14. An MDA is a medical doctor of anesthesiology a term synonymous with an anesthesiologist.
15. ASA status is a system of classifying the physical status of surgical patients. Patients are classified into one of five categories based on the presence and severity of systemic disease processes.

Chapter 2

Review of the Literature

The literature review is divided into the following sections: (a) background of the study, (b) theoretical framework (c) the reliability of scales utilized in the evaluation of anesthesia related adverse events, (d) previous studies of interrater reliability in anesthesia related adverse events, (e) rater error, (f) studies of interrater reliability in medicine, (g) The Rochester Region Perinatal Study, (h) the rating scale and level of measurement, (i) classical statistical techniques, (j) the use of contingency tables in analyzing interrater reliability, (k) the kappa statistic, (l) percentage of agreement, and (m) generalizability theory.

Background of the Study

Classical scientific methods are difficult to apply to the reliability of ratings concerning anesthesia safety. A study by Revicke, Klaucke, Brown, & Caplan (1990) substantiated this finding in a study of the ratings of several reviewers regarding anesthesia's contribution to adverse surgical outcomes. Although Revicke et al. (1990) found reliability coefficients acceptable for group comparisons, this study found that critical information was concealed when traditional correlation techniques were used to evaluate peer agreement. This research noted the failure of these methods to identify sources of error originating from different measuring conditions such as; raters, subjects, and rating scales limited their usefulness in the analysis of adverse outcomes. Revicke et al. (1990) recommended that subsequent studies regarding the contribution of anesthesia or surgical factors to adverse events make use of generalizability theory. This research concluded that generalizability theory offers significant advantages over conventional

techniques in reliability studies using multiple raters under varying measurement conditions.

A study by Goodwin and Prescott (1981) supported the use of generalizability theory in nursing research. This study evaluated four different techniques for estimating the interrater reliability of graduate students using a manometer to measure breast engorgement in new mothers. Rater agreement was compared utilizing; correlational approaches, comparison of means, percentage of agreement, and generalizability theory. The reliability measurements differed based on the technique used and ranged from a low of 0.03 using the percentage of agreement approach, to a high of 0.68 using a correlational technique. The study found that when a multifactorial ANOVA was applied to the data in the context of generalizability theory, the total variance in the ratings due to the raters, subjects, items, and their interactions could be evaluated. The data was not available with any of the other type of statistical analyses utilized in this research. The study concluded that the generalizability theory approach was the most informative technique available for interrater reliability estimation in nursing research dependent on the judgments of multiple raters, on several occasions.

Soeken and Prescott (1986) also encouraged generalizability studies (G-study) in the medical domain. They concluded that although kappa and weighted kappa are valuable with rating data, they do not examine the multiple sources of variability relevant to the conditions under which an index is used. Soeken and Prescott demonstrated that the flexible and comprehensive design of a generalizability study improved the understanding and interpretation of the rating index. They concluded that reliability is not a measure to be considered in isolation, rather it is a combination of the instrument, the

subjects and the conditions under which it is used. Based on this definition of reliability, their study found that generalizability theory offers a more comprehensive approach to patient classification than traditional psychometric methods.

Theoretical Framework

In the theoretical monograph, The Dependability of Behavioral Measurements, Cronbach, Gleser, and Nanda (1972) describe the "tidy theory of error developed by Spearman and Brown to lack the capacity to describe the perverse behavior of real data" (p. v). Traditional reliability measures consider the observed score to be the sum of a true score and random error, as represented in the equation:

$$O = T + E$$

These methods assume that objects have true scores on the attribute being measured but differ as to how this value is derived. The error term represents an undifferentiated conglomeration of elements that may include random or systemic error, unrecognized variables and interaction effects. The expanded interpretation of the error term as defined in generalizability theory may facilitate the diagnosis of systemic and random error in measurements concerning the reliability of medical judgments.

Landis (1975) noted that due to the uncertainties involved in most medical interventions the availability of a perfect model for evaluating iatrogenic injury is not available. The use of multiple observers in the measurement process was motivated by the difficulties encountered in estimating true scores in studies judging clinical decisions. Attempts to overcome this problem have also included using a panel of experts or a specified individual to serve as a gold standard. The ratings of others have been evaluated by comparing their classifications with the expert's decisions. The basic premise of this

technique is that the expert's judgment represents the true score. The interpretation of consensus scores, multiple reviewers and the complicated processes involved in medical management, require analyses that assume error is not a monolithic construct.

The difficulty in deriving true values in anesthetic adverse events has prompted the use of accepted standards of practice for points of comparison. These standards are developed and specified in advance by a panel of experts and are referred to as explicit criteria. Despite the fact these criteria and standards are based on a scientifically validated field of knowledge, they are often broadly defined and may be susceptible to considerable variation in interpretation. Standards of care for anesthesia practice have been adopted but they are limited in their scope of coverage. Therefore, explicit criteria are not always available for determining the appropriateness of clinical decisions in anesthesia related adverse events.

In situations where explicit criteria are not available the use of implicit criteria becomes necessary. Implicit criteria are assessed when an expert practitioner is given information about a case and is asked to use personal knowledge and experience to judge the care or its' outcome. Implicit criteria are imprecise and are limited by the individual qualifications and background of the rater. The nature of the data extracted from closed claims files necessitates the use of both implicit and explicit criteria, thereby increasing the potential for error in the ratings.

Although human nature makes perfection in practice unattainable, the philosophy that to err is human is one which society does not deem an acceptable attribute of health professionals. The belief that doctors and nurses are infallible according to Leape (1994) has led to reluctance to report medical errors. Historically, the approach to error

prevention in medicine has been reactive, and directed toward reprimanding the individual who committed the mistake. Leape (1994) found that seldom are the underlying causes involved in a medical error explored. To further confound the problem adverse events do not necessarily indicate poor quality care, nor does their absence necessarily signal good quality care. These factors become very significant in a medical peer review process. The belief that the standard of medical practice is error-free patient care, interferes with the ability to judge an adverse event without bias.

The most extensive study of adverse events was the 1984 Harvard Practice Study. The data for this study was derived from 30,121 randomly selected records from 51 randomly selected hospitals in New York State. The research estimated the incidence of injuries caused by medical management and then subdivided those injuries that resulted from negligent or substandard care. They found that the proportion of adverse events attributable to error was 58% and the proportion of adverse events due to negligence was 27.6%. This study also demonstrated the difficulty in judging the quality of medical care. They found the reliability of physicians judgments about the presence of adverse events was good, kappa = 0.61. However, the more difficult judgments regarding negligence had a lower degree of reliability, kappa = 0 .24. The researchers concluded that in incidents concerning negligence the physician reviewers found it difficult to judge whether a standard of care had been met.

A second problem with quality of care judgments is the interpretation of what quality patient care actually represents. The term quality is not well defined, and is dependent on a multitude of nebulous factors including the reviewer's own definition of quality of care. The low kappa value derived from judgments regarding malpractice

illustrates the problem of using a single numerical score in these situations. Important data regarding the reviewers, the adverse events and possible interactions between variables are not apparent. This finding supports the position of Cronbach, Glaser, Nanda, and Rafaratnam (1972) who advocate that formal reliability theory is not appropriate for use in interrater reliability because homogeneity of raters is not comparable with the homogeneity of numerous items comprising a psychometric test.

Brennan, Localio, and Laird (1989) concluded that the complexity and sophistication of clinical decision- making has long confounded efforts to judge the quality of medical practice. They found this to be especially true of judgments concerning injuries that resulted from care provided in a hospital, in contrast to injuries that stem from the patient's disease condition. Deciding whether an injury is an adverse event requires consideration of vaguely defined standards of care, and the ability to discriminate between events caused by the disease processes and those that result from medical management. These findings support the need to consider multiple factors when determining causality for adverse events in medicine.

Determining the standard of care that was provided is made more difficult when medical records serve as the primary source of data. Aaronson and Burmon (1994) identified errors that may occur when data is extracted from health care records. They found that measurement errors were possible at several points, including during the original collection of the data, during documentation, during extraction of the data, and during interpretation of the data. A problem inherent in retrospective chart reviews is that the data is removed from the actual event and the extraction process removes it one step further. The data may not be legible or in chronological order, and the records lack the

capacity to provide insight into the specific causes of adverse events including equipment failure, lack of adequate support, or the provider's rationale for making clinical decisions. The correlation between the records and the event may be low, due to the dependence on direct participants for documenting the incident. Therefore, the entries in the medical record may be a poor reflection of the care that was actually provided. The ambiguities resulting from vague practice standards, the presence of bias in issues of malpractice, and the reliance on medical records for data all represent possible sources of nonrandom error. These potential problems in medical studies emphasize the need to partition error when measuring peer agreement regarding the clinical decision making processes.

The study of outcome data using a peer review process has become the mechanism for evaluating and improving patient care management. Evidence based criteria have become the ultimate validators of the effectiveness and quality of medical care. Previous experience has shown that the detailed investigation of sentinel events can lead to effective strategies for improving health care. These findings have created an intense interest in the evaluation of this data during the last decade. Hospitals are required by accrediting organizations to have processes in place to evaluate adverse outcomes, and utilize this information to facilitate change. This data is the means by which the quality of care provided by health care organizations i.e. managed care vs. traditional care, and professional providers i.e. nurse practitioners vs. physicians are compared.

The gravity of the decisions based on these rating scales emphasizes the need to develop reliable instruments that can effectively separate causative from associated events. Soeken and Prescott (1986) state that "unreliable measures could introduce serious error into the reimbursement system, thereby placing hospitals and health care funding at risk"

p. 733. The need to demonstrate the reliability of instruments developed for this purpose cannot be over emphasized. Reliability in this instance refers to the ability of various raters to reach the same conclusions regarding the medical management, across rater types, institutions and disease conditions.

Risks Associated with Anesthesia

The exact mortality and morbidity rate due to anesthesia is unknown. The reason for this uncertainty is that the causes of some anesthetic deaths are not clear and may not be apparent during the anesthetic and immediate post anesthetic period. In many instances the problem of discerning whether a death was caused by anesthesia may be insurmountable. The presence of confounding variables such as the patient's preexisting disease state, interactions between the patient's condition and the anesthetic or surgical intervention, and equipment failure or error by the anesthesia provider or the surgeon may make it difficult to attribute the mishap to a single cause. The inability to measure the role that anesthesia played in an adverse event is confounded by the fact that anesthesia can never be considered separate from the operation that was performed. For ethical reasons no control study of the risks of anesthesia will be done without surgery, just as the hazards of surgery risks would not be evaluated without anesthesia. Cases with outcomes less obvious than cardiac arrest, death, nerve or brain damage, are more difficult to attribute directly to the anesthetic.

The 1996 edition of the Vital and Health Statistics of the United States reported approximately 72 million operations are performed annually. The risk of death associated with anesthesia is estimated at 1/200,200 Eichorn, (1989), which means there may be 360 deaths each year directly related to anesthesia. Clearly these outcomes are so rare that any

proper statistical design would contain so many subjects it would render a study functionally impossible. In 1990 the Federal Centers for Disease Control (CDC) reviewed data regarding anesthesia related morbidity and mortality. They concluded that a multi-million dollar study of anesthesia outcomes was not warranted based on the low incidence of occurrence. Gaba, Maxwell and DeAnda (1987) extrapolated data from previous studies and suggested that half of the deaths attributed to anesthesia in the United States are preventable. The California Medical Insurance Feasibility Study and the Harvard Medical Practice Study both found that one-half of in hospital adverse events were found to result from treatment in the operating room.

The rationale for studying closed claims is that it offers the opportunity to study a collection of infrequent events that would otherwise be considered isolated incidents. This analysis allows the identification of the types of adverse events that contribute the largest amount to insurance costs. This benefit of this type of data is that it allows research to be focused on specific areas of clinical practice. Traditionally claim files were regarded as confidential and not available for evaluation. In the mid- 1970s the National Association of Insurance Commissioners elected to make the data available to medical organizations. The rationale for this decision was that the value of educating health care providers in risk management strategies based on this data, would far exceed the risks associated with disclosure. The study of these events offers the opportunity to evaluate common factors that may have contributed to the adverse outcomes.

The Reliability of Scales Utilized in the Evaluation of Anesthesia Related Adverse Events

Few studies have addressed the interrater reliability of ratings for anesthesia related adverse outcomes across cases and measurement occasions. Revicki et al. (1990)

conducted a study with the purpose of evaluating the interrater reliability of three different scales for assigning the contribution of anesthesia technique, surgical techniques and patient disease factors to an adverse outcome following surgery. The study was unique in that it allowed the reviewers to attribute the mortality or morbidity to a combination of factors. The data was analyzed using percent of agreement and the kappa statistic. The results of this study found these scales demonstrated excellent interrater reliability. The following three scales were used:

1. The Rating Scale requires the reviewer to rate the contribution of anesthesia, surgical, patient disease, and patient risk factors to the adverse outcome on a categorical scale which ranges from "totally responsible" to "no impact" (see Table 1).
2. The Edwards scale which requires the reviewer to assign the adverse outcome to one of eight categories (see Table 2).
3. The Percent Scale requires the reviewer to assign the percentage that each of the four factors contribute to the adverse event so that the sum of the percentages equals 100 (see Table 3).

Table 1. The Rating Scale

Rate each factor regarding the degree to which it contributed to the adverse event					
	Totally Responsible	Major Impact	Minor Impact	No Impact	Unsure
Anesthesia	_____	_____	_____	_____	_____
Surgery	_____	_____	_____	_____	_____
Patient's Disease	_____	_____	_____	_____	_____
Patient Risk Factors*	_____	_____	_____	_____	_____
*For example: smoking, obesity, substance abuse, extremes of age					

Note. Adapted from "Reliability of ratings of anesthesia's contribution to adverse surgical outcomes" by D. A. Revicke, D. N. Klaucke, R. E. Brown, and R. A. Caplan, 1990, Quarterly Review Book, p. 406.

Table 2. The Edwards Scale

	Assign the injury/death to one of the following eight categories.
I.	It is reasonably certain that injury/death was caused by the anesthetic agent or technique of administration, or was directly caused by the anesthesia provider's actions.
II.	There is an element of doubt whether the anesthetic agent or technique was entirely responsible for the injury/death.
III.	The patient's injury/death was caused by both the anesthetic and surgical technique. Underlying disease may also contribute to this. For example, an inadequate response of the anesthesia provider to a surgical problem.
IV.	Injury/death was entirely referable to the surgical technique.
V.	Inevitable injuries/deaths e.g. cases of severe general peritonitis, in which anesthetic and surgical techniques were apparently satisfactory.
VI.	Fortuitous (e.g. incidental, chance, or coincidental) injuries/deaths (e.g. caused by pulmonary embolism or ruptured cerebral aneurysm).
VII.	Cases which cannot be assessed despite considerable data.
VIII.	Cases on which an opinion could not be formed on account of inadequate data.

Note. Adapted from text in "Deaths associated with anesthesia: A report on 1,000 cases" by G. Edwards, H. J. Morton, E. A. Pask, and W. D. Wylie, 1956, Anesthesia, 11, (3) p. 195.

Table 3. Percentage of Agreement Table

Estimate the percentage that each of the following factors contributed to the adverse event. The numbers may range from 0 to 100 but the total number must equal 100 %

Factor	Percent of Contribution
Anesthesia	_____
Surgery	_____
Patient's Disease	_____
Patient Risk Factors*	_____

*For example: smoking, obesity, substance abuse, extremes of age.

Note. Adapted from "Reliability of ratings of anesthesia's contribution to adverse surgical outcomes" by D. A. Revicke, D. N. Klaucke, R. E Brown, and R. A. Caplan, 1990, Quarterly Review Book, p. 406.

The study by Revicke et al. (1990) reviewed 8,000 inpatient surgical cases in which within 48 hours of receiving an anesthetic the patients died or suffered neurological dysfunction, or cardiac arrest. Twenty-two cases were selected and reviewed by a panel

of three physicians. The agreement among the reviewers regarding the contribution of anesthesia to the adverse event was 89.3% for the Edwards Scale, 82.1% for the rating scale, and 92.9% for the percent scale. To date few studies have considered the contribution of surgery and anesthesia on outcome. The demonstration of superior reliability when causality can be attributed to an interrelationship of factors (anesthesia, surgical factors and patient disease) illustrates the need to avoid items in which the reviewer is required to select a single cause.

Previous Studies of Interrater Reliability in Anesthesia Related Adverse Events

Caplan, Posner, Ward, and Cheney (1988) conducted a similar study of interrater reliability regarding the anesthesia related adverse events. This study evaluated the incident in terms of the appropriateness of care, presence of human error, and the potential for mishap prevention. This study did not consider the contribution of factors such as preexisting disease, or surgical factors into the outcome. The objective of this study was to quantify the extent of agreement, not to explain or assess the reasons for the agreement. Caplan et al. (1988) did evaluate individual reviewer traits to see if there were particular traits that influenced judgment i.e. age, background, and years in practice. The medical records and case information analyzed by the reviewers were summaries of the original documents. The results demonstrated that one fourth of the participants disagreed with their peers on each issue. Permutation testing using a chi-square measure of agreement did not reveal any significant differences between reviewers judgments based upon individual differences. The reliability statistics utilized in this study included the O'Connell and Dobson agreement measures (S_i a measure of agreement on ratings of individual subjects, and S_{av} an average of S_i for overall agreement) and the percentage of agreement. The

Williams agreement index was also used to compare the judgments of individual raters with the rest of the group. This study demonstrated only fair to good agreement between the reviewers.

There were some interesting findings in this study related to the consistency of the different measures of interrater reliability. The statistical analysis of the data revealed that the lowest percentage of agreement (54%) occurred on the issue of whether better monitoring could have prevented the mishap. Within group consensus was defined in this study when 75% of the raters rendered the same opinion. On this particular question at least 75% of the reviewers agreed 54% of the time on the issue involving incident prevention through better monitoring. The S_{av} statistic on this same question and data demonstrated the greatest level of statistical agreement. Therefore, the question having the lowest percentage of agreement between raters, exhibited the highest level of statistical agreement. This discrepancy was attributed to a disparate pattern of responses that 28% of the reviewers displayed for this particular issue using the William's index of agreement. A recommendation arising from this study was that further evaluation of the factors influencing agreement and dispute in this type of peer review process is needed.

In a subsequent study Caplan et al. (1991) found a statistically significant association between the severity of adverse outcomes and the judgments of the reviewers regarding the appropriateness of anesthesia care. The study noted that non-disabling injuries were often associated with a rating of appropriate care, while disabling injuries and death were more often associated with a rating of less than appropriate care. In order to determine the magnitude of this relationship they utilizing matched sets of cases that differed only in outcome. This study utilized six page summaries of the original closed

claim documents. Caplan et al. (1991) found that changing the case outcome resulted in statistically significant effects ($p < .001$) consistent with the hypothesis that ratings of appropriateness of care would be directly related to severity of injury. The overall kappa statistic for interrater reliability in this study was 0.21, a value indicating only fair agreement among reviewers. This research demonstrated that the severity of the injury resulting from an adverse event exerts a significant effect on the harshness of the judgment rendered by the reviewer. In a review of 122 cases of anesthesia related incidents the AANA found that claims with inappropriate CRNA care had higher SIS ratings than claims with appropriate CRNA care $p < 0.05$.

The shift in judgment regarding the appropriateness of care may also be attributed to the use of abstracted data. The basis of the judgment regarding the adverse event may shift from clinical management to outcome if insufficient information is available to recreate the process involved in the incident. The results of this research demonstrated that changing the severity of outcome also affected the willingness of the rater to render judgments. Substituting an outcome of permanent injury for a temporary injury produced an increase in ratings of impossible to judge. Conversely changing the outcome from permanent to temporary produced a decrease in the use of the undecided category. The agreement on appropriate and inappropriate care were both $k = 0.25$, and for ratings of impossible to judge, $k = 0.11$. This also may be a result of using summary information in which the crucial details needed by the reviewer to make a decision were missing. Although clinically these findings suggest a low level of interrater reliability, statistical significance at a $p < .001$was demonstrated. This issue raises the concern that statistical significance doesn't necessarily have practical value. This study concluded that based on

the use of matched pairs of cases differing only in outcome, that reasons other that the details of clinical care were the basis for the raters' judgments.

Ironically a subsequent study by Posner et al. (1996) found the level of agreement was less for permanent disabling injuries, 60%, k = 0.27, than for temporary or non-disabling injury in which the agreement was 64%, k = 0.32 These kappa values were found statistically significant at a $p < 0.05$. Considering the error bias demonstrated by Caplan et al. (1991), that a bad outcome means bad care, a higher level of agreement on the permanent disabling injuries would have been expected. A factor that may have contributed to this inconsistent effect is that the study by Caplan et al. (1991) utilized summaries of the original documents, whereas the study by Posner et al. (1996) utilized the original closed claim file. The ability to interchange the reliability ratings when the data is derived from original vs. abstracted summaries has not been demonstrated. These studies indicate there is a relationship between the severity of outcome and the harshness of the judgments rendered.

Rater Error

A certain amount of error is involved in any measurement, whether it is the measurement of vital signs or intelligence. This may also be true regarding reviewer agreement concerning clinical decision- making. The complex issues involved in this process leave it vulnerable to human sources of random and systematic error. The lack of agreement on what constitutes quality may represent philosophical and individual biases that realistically cannot be eliminated. The need to improve the reliability of potentially invalid rating made by multiple reviewers cannot be overemphasized. Cicchetti (1991) asserts that some statistics are not appropriate for measuring reliability and that qualitative

biographies of manuscripts can reveal aspects of these peer review processes inaccessible to quantitative studies. Gilmore (1979) asserts it is the meaning ascribed to the statistic rather than the choice of the statistic that is important in reliability measures. In a study of implicit judgments in chart review Dadakis and Pozen (1977) demonstrated that the use of a single method to summarize judgments may distort the conclusions. Dadakis and Pozen (1977) found the data should be summarized by several methods and that measures of association should be used to supplement tests of significance in order to develop a comprehensive understanding of the data set.

Interrater reliability and agreement are functions of the subjects or material rated, the rating scale used and the judges making the ratings. Rater variability is one problem as different raters tend to rate the same material differently. Vokel and Asher (1995) identified several personal bias errors in rating behavior.

- Generosity error – some people rate everyone high.
- Severity error – some people rate everyone low.
- Central tendency error – the tendency to rate everyone about average.
- Halo effect – the tendency to rate an individual based on an overall impression of that person, not on the attributes being measured.
- Some raters develop their own idiosyncratic methods of scoring based on their own rules (p. 142)

Cronbach et al. (1972) also identified these types of bias and found they produce main and interaction effects in the ratings. Error may result from the reviewer's lack of understanding of the theoretical underpinnings of the research instrument. This error may be enhanced if the instrument does not have clear and standard instructions, or the rater is

unfamiliar with the research instrument. Training and familiarity with the instrument can decrease variability between the raters. Statistical techniques can determine if the raters are using the instrument similarly and should be evaluated in studies of interrater reliability.

Studies of Interrater Reliability in Medicine

Reliability issues in medicine attributable to reviewers are not new or unique to anesthesia. Researchers in medical studies involving measurement and evaluation have long been aware of the observer as an important source of measurement error. In his doctoral dissertation Landis (1975) conducted a historical review of rater reliability issues in epidemiology, psychiatric diagnosis and psychological testing. He sites the work of Fletcher, and Oldham (1964) which includes a bibliography of more than 70 papers on observer error and variation in the area of clinical studies, radiology, pathology and clinical chemistry and related fields. Landis also sites the research of Cochrane, Chapman and Oldham (1951) as one of the first papers to identify the reluctance to recognize observer error in medical judgment situations. Landis found that in the period of 1940 - 1960 researchers in many disciplines began reporting studies that indicated the importance of assessing the variation in measurement due to different observers.

The impact of observer disagreement in the diagnosis and interpretation of laboratory tests is well documented through out the literature. There are numerous studies reporting disagreement among radiologists on chest film interpretation including Birkelo , Chamberline, Phelps, Schools, Zacks, and Yerushalmy (1947), Flether and Oldham (1949), Yerushalmy, Harkness, Cope, and Kennedy (1950), Yershalmy, Garland, Harkness, and Zwerling (1951), and Cochrane and Garland (1951). In two related works

Landis and Koch (1977) evaluated the variability of the classification of carcinoma in situ of the uterine cervix, and measured agreement in the diagnosis of multiple sclerosis among neurologists. The importance of considering inter-observer error (the inconsistency of interpretation between observers) and intra-observer error, (the failure of an observer to be consistent with himself) were both found to be factors in this research.

The rater variability in diagnosis and classification of patients based on clinical symptoms is also replete in the literature. In a report by Smyllie, Blendis and Armitage (1965) the level of observer disagreement in twenty respiratory signs in patients with various diseases was evaluated. Butterworth and Repert (1960) reported the results of comparing 9333 physician's agreement on diagnoses dependent on auscultation interpretation. This study found the average number of correct diagnoses to be 49%. Variability in the recording of physical signs was noted in research by Cochrane et al. (1951), Reynolds (1952), Bearman, Kleinman, Glyer, and La Croix (1964), Fletcher (1964), and Lowenson, Bearman, and Resch (1972). These studies demonstrate the prevalence of interrater reliability problems across many medical disciplines. This data supports the position that statistical methodologies that evaluate multiple sources of error may be more appropriate in situations involving clinical decisions.

The Rochester Region Perinatal Study

The Rochester Region Perinatal Study in Albany New York looked at the paired judgments of independent reviewers on 1,258 obstetric cases. This report specifically dealt with peer judgments regarding the adequacy of care rendered to the women and their babies. The results indicated that opinions of reviewers measuring the quality of clinical care derived from retrospective analysis of hospital charts was considerably more varied

than had been expected. This study found the reviewers' judgments were influenced by personal clinical bias, careless record review, misinterpretation of recorded events, geographic differences, stringency in judgment and reluctance to provide critical judgments despite anonymity.

In an effort to improve the agreement, a second study to determine the quality of care was conducted. The scoring system was changed and the study included additional information about the judgment process. Despite these changes the expert clinicians produced estimates of the overall quality of care that were inconsistent with one another and poorly correlated with other measures of the medical care process. The conclusion of the Rochester Study was that peer judgments regarding the quality of clinical care derived from retrospective analyses of hospital charts was not sufficiently accurate nor homogenous to be of practical use in decision making by the government or any third party insurers.

This study demonstrates that the use of a single method to summarize the data may distort the conclusions. In addition to scoring the reviewers involved in the Rochester Study were asked to make additional comments when they felt that the satisfactory and unsatisfactory judgments did not adequately explain their opinions. An interesting finding in this study was that although the rater's scores were is disagreement, the written comments added considerably to the understanding of the review process and the problems related to defining the quality of care. In a study evaluating the reliability of decisions to accept or reject manuscripts based on a peer review system Cichetti (1991), also demonstrated the value of narrative summaries. The conclusion of this study was that although reliability studies are important, they should be complemented by additional types

of research aimed at determining the philosophical or sociological basis for the decision. The application of generalizability theory to this data may offer a methodology that can reveal aspects of the peer review process inaccessible to other quantitative studies.

The Rating Scale and Level of Measurement

The rating scale is one of the most frequently utilized measurement instruments in research. These scales require the rater to make a judgment about some characteristic of the object of study and then assign it to some point on a scale. The closed claim data collection instrument consists of over 150 data points of nominal and ordinal scale data. Nominal scales consist of a simple classification system without order. These questions instruct the rater to select one of several qualitatively different categories. Rater disagreements concerning these questions do not differ in degree, they are either in agreement or disagreement. Interrater agreement represents the extent to which the different judges tend to make exactly the same judgments about the rated subject.

The ordinal scale is a classification system based on an ordered relationship of one item to the next. The data is ordered so that each selection differs from the next in terms of greater or lesser severity. This data signifies a position on a scale, but does not include a precise interval from one level to the next. Interrater reliability can be determined from this data by analyzing the degree to which the ratings of different judges are proportional when expressed as deviations from their means. In this instance high reliability does not necessarily mean the two raters are in complete agreement, just as low reliability does not mean complete disagreement.

There are several potential problems in the design of a classification system that consists of mutually exclusive and exhaustive categories. An interpretation problem is

created by the inclusion of an undecided selection in the available choices. The dilemma with incorporating this category into the statistical analysis of concordance is that is does not necessarily signify agreement. The reviewers may actually have contradictory opinions regarding the case management but feel there is inadequate information to make a definitive judgment. Two categories included on the data collection instrument used in the AANA closed claim analysis are an impossible to judge and cannot determine category. The inclusion of this data into the measurement of interrater agreement may artificially inflate the number. A study by Revicke et al. (1990) demonstrated that rating scales allowing the reviewer to attribute causality to more than one source have better interrater reliability. The inclusion of this type of item may decrease the selection of undecided categories.

Classical Statistical Techniques

Classical reliability coefficients require investigators to obtain independent but interchangeable measurements and compare how these scores vary about the true score. Measures of association between two data sets have been widely studied. The earliest works that evaluated methods of measuring the association between two attributes are papers written by Yule in 1900 and 1912. Thurstone (1927) was one of the first investigators to address the rater as a source of error. Goodenough (1936) found that inconsistencies in observers, and subjects' responses to the same stimulus on different occasions may all be sources of error. Goodenough's work demonstrated that one comparison detects only some of the inconsistencies present. This line of criticisms led Thorndike (1947), in Research Problems and Techniques, to classify the type of variance that can lead to error into the following five categories:

1. Lasting and general. For example, level of ability, and general test- taking ability.
2. Lasting but specific. For example, knowledge or ignorance regarding a particular item that appears in one test form.
3. Temporary but general. For example buoyancy or fatigue reflected in performance on every test given at a particular time.
4. Temporary and specific. For example, a mental set that affects success in dealing with a particular set of items.
5. Other, particularly chance success in guessing.

Although, Thorndike (1947) recognized the multifaceted conception of variance he did not offer methods for estimating the amount of variation from each source.

According to Cronbach, et al. (1972) RA Fisher revolutionized statistical thinking with the introduction of the analysis of variance (ANOVA) for analyzing variance. These researchers stated; "investigators who adopt Fisher's concept of the factorial experiment in which the conditions of observation are classified into several respects, must abandon the concept of undifferentiated error. Therefore, error must be seen as attributed to multiple sources and estimates regarding how much variation arises form each component should be evaluated" (p. 1). These investigators found that the analysis of variance and covariance made better use of the data, because it allows the evaluation of effects that would be lumped together by reliability coefficients. Cronbach et al. (1972) found that these techniques allow the investigator to understand the sources of unwanted variation. The researcher can then use this information to plan a more efficient design for collecting further data. This information according to Cronbach et al. (1972) is not available using other methods.

In his doctoral dissertation Loveland (1952) computed components of variance to estimate the magnitude of variation from five sources:

1. Persons
2. Person-Occasion Interaction
3. Person-Form Interaction
4. A Form-Occasion Effect
5. A Residual Effect.

Cronbach and Lindquist (1953) wrote an extensive paper focusing on reliability coefficients and concluded that a multifaceted analysis would allow for alternative definitions of error. Although investigators such as Guiford (1954) applied the analysis of variance to ratings, discussing the effects for subjects, raters, traits, and their interactions, he did not relate the analysis to reliability measurement. Despite the application of the analysis of variance to psychology and education since the 1930's Cronbach et al. (1972) can identify no predecessors of multivariate error.

Measures of correlation are descriptive measures that indicate the degree in which two or more variables are related. A reliability coefficient is a ratio of true score variance to observed score variance. This value is derived from the following formula:

$$r_{xx} = \frac{V_t}{V_o}$$

This coefficient represents the proportion of true score variance in the obtained scores. The value of this measure can theoretically range from –1 to +1 although in practice the range is from zero to one. If the measurement error is zero the reliability coefficient will equal one. There is considerable variation in the literature regarding guidelines for the

interpretation of the strength of a correlation coefficient. Coefficients > 0.70 are generally considered satisfactory according to Polit and Hungler (1983) p. 388. Soeken and Prescott (1986) state that "one would expect levels of at least 0.80 in the total variance accounted for" p.734. According to Washington and Moss (1988) a reliability of 0.80 is considered the lowest acceptable coefficient for a well-developed instrument, and a reliability of 0.70 is considered acceptable for a newly developed instrument.

Reliability of ratings is often stated in terms of a one-way intraclass correlation coefficient (ICC). According to Burdock, Fleiss, and Hardesty (1963), the ICC is a measure of the intrinsic accuracy of the instrument inversely proportional to the total amount of uncontrolled variability in the scores. If the value is high it indicates that a large fraction of the variance in ratings is between subjects not raters, and it follows that the interrater reliability is also high. Many reliability indices available are versions of the intraclass correlation and are typically a ratio of the variance of interest to the variance of interest plus error. Shrout and Fleiss (1979) list 6 forms of ICC and discuss the criteria for selecting the appropriate version based on the experimental design. Maxwell and Pilliner (1968) developed a measure of association from an analysis of variance model using a random block design and adapted it for use with dichotomously scored data. Holley and Guilford (1964) proposed the G index for measures of association between categorical data. The G statistic also known as the E coefficient (Janson and Vegelius 1979) can vary between +1 and –1.

The Use of Contingency Tables in Analyzing Interrater Reliability

A variety of methods of analyzing interrater reliability in situations involving a dichotomous judgment are available. A contingency table is a two-dimensional frequency

distribution that cross tabulates the frequencies of the variable. Each cell of the table contains information about the counts obtained for each level of the two variables. The sum of all the cells equals the total number of coded observations. In reliability measurements the extent to which the observers agree can be described as functions of the proportions obtained from underlying contingency table. This data is used to derive inter-observer agreement on the classification of subjects and to detect bias in the overall usage of the measurement scale.

Item by item agreement is reflected in the main diagonal of the matrix (cells A + D). This diagonal proportion is referred to as the "index of crude agreement" by Rogot and Goldberg (1966), and reflects the subjects classified into the same category by two observers. Cells C and B reflect the number of rater disagreements. The marginal totals (p_1, p_2, q_1, q_2) reflect the consistency of the raters in using the category system. The notation of the cells as illustrated in the Table 4 will be referenced throughout this study.

Table 4. Appropriate/Inappropriate 2x2 Contingency Table

Rater Two	Rater One		Total
	Appropriate Care (+)	Inappropriate Care (-)	
Appropriate Care (+)	*(A)* (+) (+)	*(B)* (-) (+)	$A + B = p_1$
Inappropriate Care (-)	*(C)* (+) (-)	*(D)* (-) (-)	$C + D = q_1$
Total	$A + C = p_2$	$B + D = q_2$	$A + B + C + D = N$

The assumption of marginal homogeneity is the assumption that all the raters use the classification scheme in the same way, yielding identical underlying marginal distributions for all raters. The observers are said to act independently if each allocates

categories at random according only to his or her marginal distribution. A marked degree of marginal asymmetry would suggest the two observers were consistently using different criteria for assigning items to categories. In typical reliability studies the rates are a priori deemed equal in their ability to make judgments and no restrictions are placed on the distribution of ratings across categories. The advantages of contingency data are that they facilitate both the measurement of agreement and consistency in the use of the instrument.

The Kappa Statistic

The kappa statistic was first proposed by Cohen in 1960 and is the most common measure of overall group agreement for nominal data. Kappa is defined as the proportion of agreement between paired observation, corrected for chance agreement. P_0 is the proportion of agreement that was observed to occur, and P_e denotes the proportion of interrater agreement expected by chance alone. Cohen's formula for kappa is represented by the following formula:

$$k = \frac{P_o - P_e}{1 - P_e}$$

Kappa was originally formulated for situations in which two raters each evaluate one subject by selecting a single response category. This original formulation has since been adapted to the case of multiple raters. The research of Scott (1965) and Maxwell and Pilliner (1968) proposed variants to this statistical measure of agreement. Kappa can be used to provide a measure of agreement on any particular category in a coding scheme, as well as on overall agreement. Occurrence and nonoccurrence kappa can be determined using the formulas devised by Kent and Foster (1977). Kappa is useful when all disagreements may be considered equally serious. A weighted form of kappa was

developed for use in ratings in which certain disagreements were deemed more serious. Cohen (1968), Everitt (1968) and Fleiss et al. (1969) have provided formulas for differentially weighting disagreements using nominal ratings.

Kappa, pi, and lambda are all derived from the same theoretical formula (Cohen, 1960), and differ only in the interpretation of the expected proportion of agreement. All three of these measures will reach unity only when the error cell frequencies are zero, this situation is referred to as absolute association, in contrast to zero frequencies in only one error cell which is referred to as complete association (Kendall and Stuart, 1961). Phi can be interpreted as a derivation of chi-square corrected for sample size House et al. (1980). Fleiss (1973) provided values for the use of phi as a measure of association, he suggested that levels of 0.35 or less indicated no significant association.

The literature has grown prodigiously over the past two decades with information on various forms and uses of kappa. According to Fleiss (1971) kappa assumes independent raters, but the raters need not be the same across subjects. Fleiss's formula does require the number of judges per subject to be constant and is represented by the formula:

$$K_v = \frac{P_{0_v} - P_{c_v}}{1 - P_{c_v}}$$

Where P_{0v} represents the proportion of ratings in which the judges agree, and P_{cv} represents the proportion of ratings for which agreement is expected by chance. A generalized kappa statistic K_g can be used in situations in which each of a sample of subjects is rated by two or more observers. Generalized kappa reflects the reliability across all rater combinations corrected for chance. A second approach with multiple

raters is to use a consensus score as a basis for examining each rater. The premise underlying this model would be that the model chosen for comparison would represent a true score against which the others will be compared. This type of design could be used without violating the assumptions of kappa if the individual rater being compared is excluded when the consensus score is determined. The estimate of K_g will be less than kappa or the value of kappa derived from the comparison of an individual rater and a consensus group. The various forms of kappa are used to the same statistical endpoint, which answers the question of whether the level of chance-corrected agreement is significantly greater than zero.

The use of kappa in assessing reliability involves no assumptions about the nature of the underlying distribution. The levels of statistical significance for the kappa statistic is obtained by dividing the value of kappa by the standard error of kappa and referring to the resulting z value on a table. Standard error estimates may differ based on assumptions regarding rater selection (random versus fixed). "The only assumptions required by kappa are that the subjects to be rated are independent, the judges assign their judgments independently, and the categories of the nominal scale are independent and mutually exclusive and exhaustive Soeken and Prescott (1986) p. 735."

Investigators have provided guidelines for defining levels of kappa that may be regarded as clinically or substantively important. In the literature the range of kappa values is quite diverse in fact original kappa values were admitted to be clearly arbitrary. Ciccheti, Sparrow (1981) and Fleiss (1981) define kappa as follows; kappa values less than 0.40 represent poor levels, magnitudes between 0.40 and 0.59 denote fair levels of reliability and coefficients between 0.60 and 0.74 constitute good levels of reliability and

those greater than or equal to 0.75 reflect excellent levels of reliability. A review of the literature finds kappa ranges from –1 to + 1, with 0 indicating no agreement. Landis and Koch (1977) suggest the interpretation of kappa as presented in Table 5. The values represented on this table will be utilized in this research project.

Table 5. Kappa Values

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Excellent

Note. From "Agreement measures for categorical data" by J. R. Landis and G. G. Koch 1977, Biometrics, 33 p. 165.

The actual range of kappa values depends on the number of ratings per subject. Fleiss (1979) found that kappa may range from $-1/(n-1)$ to 1, where n is the number of ratings of each subject. Formulas for calculating the standard error of kappa for testing the significance of different values of kappa against the hypothesis that $k = 0$ are provided by Cohen (1960), and Fleiss et al. (1969). Kappa is approximately normally distributed but with a slight negative bias.

There are several criticisms commonly raised regarding the use of Cohen's kappa coefficient as a measure of agreement on categorical data. Researchers comment that kappa does not justly represent agreement in terms of the similarity between two observers marginal distributions. Despite the fact that the row and column totals for each category may be identical kappa can be very small. Another problem with this statistic is that the maximum value that kappa can attain is constrained by differences between the two marginal distributions. Therefore, item-by-item agreement is seen as being diluted by

the effects of marginal asymmetry. House, House and Campbell (1980) found that the "effect of correcting for chance agreement is simply to make it more difficult to obtain high levels of observer agreement at either high or low frequencies (p. 49)". This is illustrated by the following formulas:

$$\text{Observed Agreement} = \frac{A+D}{N}$$

$$\text{Expected Agreement} = \frac{p_1 p_2}{N} + \frac{q_1 q_2}{N}$$

In response to these criticisms Cohen (1960), stated that a measure of marginal symmetry should be available for routine calculation along with kappa, to facilitate an understanding of the nature of the disagreements. In typical studies of interrater reliability training alone should result in a level of agreement significantly greater than that expected by chance. The finding of a negative kappa value or a level of agreement less than expected by chance would indicate a serious problem.

There are several limitations in the definition of kappa and it's application to statistical sampling theory. Kraemer (1980) found the requirement that an observation be assigned to one and only one category to be unacceptable. She explained that the major limitation of kappa is that it is limited to situations in which one has an equal number of observations per subject. Fleiss (1971) found that the consequence of this is to require researchers to discard subjects to artificially equalize the numbers. This practice according to Fleiss results in a waste of information with attendant loss of power in testing procedures. Kraemer, Bliwise and Bliwise (1991) found that kappa could be inflated or deflated by changing the number of categories. Kraemer (1980) found that confounding population and classification problems could prevent kappa from attaining values between

0 and 1. This study concluded that different techniques could yield different estimates of reliability coefficients. This could occur because of the nature of the statistics used, differences in the error associated with rater agreement, distribution of the scores, number of subjects, and the number of raters especially when more than two raters are involved.

Percentage of Agreement

Baer (1977) argued that in most instances knowing the percentage of times two raters agreed is more informative to an investigator than the percent of variance explained by correlation like procedures. The results of percentage methods range from 0 (no association) to 100% agreement (perfect association). There are basically two types of approaches to percentage agreement; total agreement, and agreement on either occurrence or nonoccurrence. Total percentage agreement analysis includes both agreement that something happened and agreement that something did not happen. This type of measure of association was found acceptable for behaviors with moderate (40-60%), occurrence by Hawkins and Dotson (1975), and Kratochwill and Wetzel (1977). House et al. (1980) found total percentage agreement methods to be a generally unacceptable approach. A problem sited with this measure is that adding agreement on nonoccurrence to the total number of agreements will yield a high percentage agreement despite few agreements on occurrence, (Yelton, Wildman and Erickson, 1977). House et al. (1980) found that very frequent or infrequent events produce high agreement values regardless of the consensus of the raters. In percentage of agreement two raters may never agree on a behavior as occurring yet still achieve a high total percentage of agreement. An alternative approach in these types of situations is to restrict the analysis to either occurrence or nonoccurrence. The main problem with this approach is that a small n may cause this value to be too

conservative.

Another dilemma with percentage agreement is that just how much agreement is appropriate is not determined on an a priori basis. There is some consensus among behavioral scientists that an average of 70% is necessary, 80% is adequate, and 90% is good (Hartman 1977, House, House, and Campbell, 1981). Guttman (1971) concluded after a review of the literature that 65% was a minimum acceptable proportion of agreement. This number has been criticized because it would allow disagreements to exist one-third of the time. The major criticism of percentage agreement is that lower and higher frequencies in the cells shown are associated with lower and higher interrater agreements respectively. As n increases, agreement due to chance increases (Hartmann 1977, Wakefield 1980, and Yelton et al., 1977).

A related problem is that similar percentage agreements with different sample sizes do not necessarily reflect the same degree of consistency between raters. This is due to the failure of the procedure to correct for chance agreements (Yelton et al., 1977) Kappa solves the problem of inflated percentage agreement due to chance by controlling for chance agreement (Hartman, 1977). However, considerable uncertainty exists as to how the chance correction should be incorporated into the measure of agreement. Goodman and Kruskal (1954) contend that chance agreement need not cause much concern because the observed degree of agreement may be assumed to be in excess of chance. This is particularly true in situations where individuals share expertise in a field of study and are making decisions regarding their specialty area. A significant degree of concordance between these individuals is expected based on their common knowledge and the necessity for introducing chance expected agreement in this situation is controversial.

Generalizability Theory

Generalizability theory (GT) is a statistical theory about the dependability of behavioral measurements. Dependability refers to the accuracy of generalizing a judge's rating to the average rating the person would have given under all possible conditions.

> The score (on a test or other measure) on which the decision is to be based is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to a particular stimulus, objects, or questions , to the particular tester, at the particular moment of testing. Some, at least of these conditions of measurement could be altered without making the score any less acceptable to the decision maker.... The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations (Cronbach, et al. 1972, p.15).

In this respect generalizability theory is an extension of domain sampling in which the ratings of a particular judge is not the main concern but the ability to generalize to all judges is the issue under consideration. In this context the generalizablity of the ratings is the main concern.

Nunnally and Bernstein (1994) write that generalizability theory is one of the most significant extensions of classical measurement theory and should be utilized in situations where the data consists of ratings. These researchers encourage the application of GT to occasions in which raters evaluate individuals on multiple attributes. "The basic assumption in this process is that the rater's knowledge, attitude, skill or other measured attribute is in a steady state. Therefore, any differences in scores on different occasions of measurement are due to one or more sources of error, and not to systematic changes in the individual due to maturation or learning" (Shavelson & Webb, 1991, p. 1). Generalizability theory offers a more comprehensive framework for evaluating difficulties associated with measurements derived from multiple raters. The theory of generalizability

is described by Cronbach et al. (1972), in The Dependability of Behavioral Measurements. In this text Cronbach describes the following advantages of generalizability theory:

1. Explicit consideration of the several facets of a measuring operation dispels ambiguities that were present in, and concealed by, the classical model.
2. The multi-facet study can appraise interactions inaccessible to the older methods, and so can improve one's understanding of the measure.
3. One multi-facet study answers questions that formerly required several separate sets of data.
4. Multi-facet information enables one to design more efficient procedures for collecting data, either for the measurement of individuals or for the determination of group means. (p. 2)

GT also enables the decision maker to determine how many occasions, test forms and raters are needed to obtain dependable scores. This can be done applying the Spearman-Brown prophecy formula to the data and substituting raters for items. A stepped down formula can estimate the reliability if fewer raters than used in the original reliability estimate are employed.

G theory provides a summary coefficient that is analogous to classical test theory's reliability coefficient. It is equal to the intra- class correlation coefficient that would result if a one-way ANOVA were applied to the data. A generalizability coefficient symbolized by p^2 is a ratio of the universal variance of the ratings to the expected variance of the ratings. The generalizability coefficient expresses on a zero to one scale, how well the rating on the subject reflects the universe of measurements on the subject. This number

reflects measurements of reliability across all rater combinations. The interpretation of a generalizability coefficient is similar to that of traditional reliability coefficients except that it is based on a design that evaluates multiple error sources. Depending on the design the error may or may not equal the mean square residual derived using an ANOVA technique. Despite the promise that generalizability theory may provide insight into reliability measures not available with other techniques, it has not been widely accepted. The explanation for why generalizability theory has not been included in the lexicon of commonly used reliability measures is not well understood. However, the unique terminology and the complexity of the calculations associated with the methodology may be partially to blame. Increased application of this theory may be necessary for acceptance and the evolution of this reliability measure into a user friendly form of analysis.

Generalizability theory utilizes the statistical models and research designs of analysis of variance. The emphasis and interpretation is based on the estimation of variance components. It is described by Evans (1981) to be an elaboration and application of ANOVA to multifaceted experimental designs, which can be applied to both crossed and nested factors to estimate the components of variance. Revicki (1984) writes that the primary interest in generalizability theory is the determination of the magnitude of individual variance components. This procedure then generates estimates of the extent to which the measurements are confounded with error. The expansion of factorial ANOVA allows the recognition and decomposition of multiple error factors and interactions that contribute to variance in the data.

In generalizability theory an individual judge's rating of a subject is the sum of the following components:

1. a universal effect (u) describing the mean rating of all individuals by all judges in the domain.
2. the deviation of a particular individual's mean rating u_i from $u\ (u_i - u)$
3. the deviation of a particular judge's mean rating (uj) from $u\ (uj - u)$
4. *eij* reflects both classical measurement error plus systematic disagreement among judges
5. $x_{ij} = (u_i - u) + (u_j - u) + e_{ij}$, (Nunnally and Bernstein 1994, p. 281).

The sources of error in a generalizability study (GS) are called facets and the levels of the facets are called conditions. A facet is a set of conditions under which measurements are carried out i.e., in the closed claim study the items, raters and severity classification represent facets. The selection of the appropriate facets of differentiation are based on the purpose of the study.

The term population represents the object of measurement and the term universe describes all the conditions under which the measurements are made. Cronbach et al. (1972) did not consider subjects to be facets and distinguished between them and other conditions of the design. Cardinet et al. (1976) elaborated on Cronbach's theory and made the subjects a facet in order to extend generalizability theory to a wider range of measurement problems. These researchers gave equal status to the persons being measured with the conditions of measurement. They redefined a facet of generalization as any source of variation that affects the values of the measures on the material under study. Generalizability theory recognizes that any rating is subject to the influence of systematic and random error that cannot be controlled. The main advantage of this technique is that a

single multi-facet study addresses issues simultaneously, that other techniques would have to analyze using several tests.

The first stage in a generalizability study is the selection of facets and the computation of mean squares. A GS is designed to estimate variance components underlying a measurement process by defining the universe of admissible observations as extensively as possible. According to Cardinet et al. (1976) a generalizability study cannot be adequately designed unless there is a priori consideration of fundamental issues associated with each phase of the analysis. The second stage is the determination if the facets are infinite or finite or random or fixed. A random sample as defined by Cronbach et al. (1972) does not necessarily have to be randomly drawn. It is described as being smaller than the size of the universe and is considered to be equivalent with any other sample of the same size drawn from the universe. This abstraction of a random sample is similar to the Bayesian concept of exchangeability, (Novick 1976).

A facet is considered fixed when the levels associated with the facet exhaust all the possible conditions of interest. A fixed facet may be indicated when the entire universe is small or when the decision-maker is not interested in generalizing beyond the conditions represented in the study. Cardinet et al. (1976) found that fixing or not fixing a facet could have important consequences on the value of the generalizability coefficient. For example if a design were composed only of fixed facets than the generalizablity coefficient would be 1.0 since all the information concerning the variables would be available.

A facet may be nested or crossed with other facets. According to Nunnally and Bernstein (1994), the most useful G study is one that crosses judges with individuals so each judge evaluates each individual. Crossed designs allow the estimation of more

variance components than nested models. The third stage is the estimation of error and generalizability. This stage identifies which facets may limit the generalization of the measurement. In the study of closed claims the literature has demonstrated that the severity of the outcome and provider type may restrict the generalizability of the results. The fourth stage involves the recommendations for subsequent studies based on the relative contribution of each variance component to the error variance for a given design. The information provided by the analysis will identify modifications that may be used to improve the design for future research.

Generalizability distinguishes between G (generalizability) and D (decision) studies. The distinction between G and D studies is that a G study is carried out during the development of a measuring procedure. A G study enables the investigator to identify multiple sources of measurement error and identify what variables need to be controlled. This study allows an identification of factors responsible for inconsistencies in the ratings. A D study then uses this information to design a measurement process that minimizes the identified sources of error. The distinctions between relative and absolute decisions are also important to this methodology. Relative decisions are dependent on the differences between individuals and absolute decisions are scores that are interpreted without reference to the performance of others. Generalizability information offers possibilities for improving instruments that could not be attained using other methods of statistical analysis.

The two major contributions of generalizabilty theory are its emphasis on the multiple sources of measurement error and its de-emphasis on the role played by summary reliability coefficients. The ability to weigh the relative contribution of each source of error

and determine the dependability of the measurement is made possible by the estimated variance components. The main advantage of a generalizabilty study is that it promotes improvement in research design through the reduction of error. A second benefit is that it provides a framework wherein the theory of reliability and validity merge. According to Cronbach (1963) the results of a generalizability study indicate how validly a measure can be considered to represent the entire collection of all possible measures. He states that the validity of a process will be improved by the elimination or reduction of bias.

Chapter 3

Methodology

This chapter provides a description of the research methodology. The research design utilized for this study is described. The sample of individuals participating in the study and the materials used are discussed. The measurement instrument and the procedures utilized to establish the validity of the instrument as well as the data collection procedure are explained. The statistical analyses that were performed to address the research questions stated in Chapter 1 are described.

Research Design

The research design for this study includes the application of traditional psychometric and generalizability theory to interrater reliability data. The data was derived from an instrument developed to use in the analysis of anesthesia related adverse events. The reliability is described using the intra-class correlation coefficient, percentage of agreement, and the kappa statistic. The generalizability data was obtained by the application of an analysis of variance to the data. The design is completely crossed with each rater scoring each claim and item. The variability associated with the severity of injury, provider type and the interaction of both these factors on the raters' judgments is analyzed. The percentage of agreement was obtained by designating the scores of the committee chairperson and mode as the correct response. Kappa values were calculated as the measure of chance corrected agreement. These measurements were derived from pair-wise comparisons of the ratings of each reviewer to those of the gold standard. The dispersion of scores and consistency between reviewers in the use of the instrument was analyzed using 2 x 2 contingency tables.

Independent Variables

Raters – are a fixed variable, and consist of the eight closed claim committee members.

Claims – the 12 claims were not randomly selected. They represent a stratified sample of incidents, providers and severity levels. They are a fixed variable.

Severity level – the severity level represented by the outcome is a fixed variable. There are 2 levels associated with this variable. Level I represents injuries with an SIS score of 6 or less and are classified as temporary or non-disabling injuries. Level II contains all injuries with an SIS score of 7 or more and are classified as permanent and disabling injuries.

Providers – are a fixed variable with 2 levels. Level I is care provided by a CRNA, and level II represents care delivered by a CRNA and MDA when the adverse event occurred.

Dependent Variable

Item Score – The rating assigned to each of the seven items on the data collection instrument by the rater.

The Sample

The claims were obtained using a computerized data search of all the branches of St. Paul Fire and Marine Insurance Companies. This national search included all claims filed against CRNAs since 1995 naming St. Paul as the insured party. The current database is comprised of over 200 files that have been reviewed and analyzed. A total of twelve claims were selected for estimating the degree of interrater reliability of the research team. A stratified sample was chosen to facilitate the inclusion of a variety

of adverse events. These claims were analyzed at the main branch of the insurance carrier in St. Paul Minnesota between1996 and 2000. The twelve files were reviewed by the chairman and determined to meet the inclusion criteria for the study. The following characteristics deemed a claim unsuitable for review:

1. Claims in which the documentation is inadequate to allow reconstruction of the sequence of events and the nature of the injury.
2. Those files in which the medical records are missing.
3. Claims that involve dental damage, since the causation for this injury is understood.
4. Claims in which the amount paid for the CRNA was less than $1,000.00
5. Claims in which the anesthesia provider was named in the suit but bore no responsibility either directly or indirectly for the basis of the claim.
6. No claim previously evaluated for interrater reliability was included in the analysis.

Data Collection

The data was collected on a standardized form, according to the written set of instructions that are supplied to each reviewer (Appendix A & C). The claims were all analyzed at the main branch of St. Paul Companies in St. Paul Minnesota, in quiet and secure locations. The reviewers did not discuss the file's content and the ratings were made independently. The conditions and time of day were not significantly varied. Permission was obtained from the AANAF to use any and all information from the closed claim data forms (Appendix B). Anonymity regarding the identities of the patients, providers and reviewers was assured. Confidentiality regarding specific details of the claims including the geographic location was guaranteed.

Data

The instrument consists of over 150 data points including:

- Demographic data.
- The patient's history and physical condition.
- The type of surgery.
- The anesthesia providers.
- The specific anesthetic management including drugs, monitors and documentation.
- Complications that occurred and their management.
- Case management including preoperative, intra-operative and postoperative events.
- The patient's recovery course following the anesthetic.
- The legal proceedings.
- The disposition of the lawsuit, and the amount of the settlement.

The primary interest in this study is the evaluation of items that require the reviewer to make a subjective decision regarding the care provided to the patient. The reviewer is asked to evaluate the care provided based on what a reasonable and prudent anesthetist would have done at the time of the event. These questions include the following:

1. Were inadequate pre-induction anesthesia activities the basis for the lawsuit?
2. What was the ability of the records to provide an understanding of the details surrounding the event named in the lawsuit?
3. Was the anesthesia treatment by the CRNA appropriate?
4. Could the CRNA have prevented the basis of the lawsuit?
5. Did a lack of CRNA vigilance contribute to the basis for the lawsuit?

6. Could anyone have prevented the events that lead to the lawsuit?
7. Would better technical monitoring probably have prevented the event named in the lawsuit?
8. Select the Severity of the Injury (SIS) based on the 10 point scoring system. (Table 6.)

Table 6. Severity of Injury Scoring (SIS) System

	Class I Injury	Examples
1	No Obvious injury	Pain
2	Emotional Injury Only	Mental Anguish, Awareness, Anxiety
3	Temporary Insignificant	Corneal abrasion, Laceration, Dental or Oral Damage
4	Temporary Minor	Burns, Prolonged Recovery Room Stay, Unplanned Admission
5	Temporary Major	Nerve Damage
6	Permanent Minor	Damage to Organs, Inadvertent Perforations, Decreased vision
	Class II Injury	Examples
7	Permanent Significant	Loss of an Eye, Deafness
8	Permanent Major	Paraplegia, Loss of Limb, Brain Damage, Blindness
9	Grave	Severe Brain Damage, Quadriplegia
10	Death	

The reviewer is also asked to complete a written summary of the events leading up to and including the adverse incident. This summary should provide enough information that it allows the reader to recreate incident.

Reliability and Validity of the Instrument

Content validity was established by a comparison of items to The Standards and Guidelines for Nurse Anesthesia Practice (1992), A + Risk Management Tool, and the

evaluation of anesthesia records and similar instruments. A panel of experts familiar with the theoretical formulations underlying the instrument examined the items, and found them to be representative of the domain of subject matter. An instrument instruction manual (Appendix B) was developed to use concurrently with the instrument. The reliability of the instrument will be discussed in chapter five of this manuscript.

Population

The closed claims research team consists of eight reviewers who were all involved in the instrument development, and have been working on this project since 1995. This group was selected by the AANAF Board of Directors in attempts to get cross sectional representation. The criteria utilized were based on clinical experience, educational background, research goals, and a committed interest in studying anesthesia-related adverse outcomes. The selection of clinically and educationally diverse team members promoted a representation of all practicing CRNAs.

The necessary sample size for the study of interrater reliability using both classic psychometric and generalizability theories varies considerably. Although Nunnally and Bernstein found that the minimum requirement for a generalizability study is two judges with each judge evaluating each individual, the stability of the reliability estimates will improve as the square root of the sample size increases. A sample size of 25 – 30 is needed for significance testing of kappa. Cichetti (1976) suggested to improve the reliability of peer review a minimum of three independent reviewers are recommended. Fleiss (1971) used six raters and five response categories with a sample size of 30. Evans, Cayten, and Green (1981) designed a two facet G study to assess patient status with an N = 15, using three raters on five occasions. Based on the recommendations of these

previous studies this research will include the analysis of a total N=12 claims, using eight raters on seven items.

The number of raters, items, and claims in this study is sufficient to provide information regarding the future use of this instrument and the confidence that may be placed in the data. The use of original documents and the evaluation of all the subjective data, constrained the sample size. The process involved in the evaluation of each claim is comparable with a case study. The average length of time for a review is 90 minutes and the claims must be evaluated at centralized areas under the supervision of the insurance carrier. The external validity of this study was the basis for selecting the actual closed claim files. The long- term goal of the AANAF Closed Claim Study is to review claims on a regular basis and expand the computerized database of adverse events. An expectation of this ongoing research is that adjustments in the instrument and membership will be made as the project evolves. Therefore, the reliability of the process will be established at each stage of the instrument's evolution.

Statistical Analysis

The statistical analysis was performed using the Statistical Package for the Social Sciences (S.P.S.S.) 10.0 for Windows (2000). The intra-class correlation was used to measure the consistency of the composite ratings of all the judges in the study. The between raters variance was included in the error term to calculate the overall reliability of the judges for each individual item. Interrater agreement was be determined by the percentage or proportion of agreement between the judges. This number represents the proportion of times the raters agreed relative to the total number of observations made. The nature of the dichotomous data necessitated that agreement be considered an

absolute. The disagreements are considered equally serious and weighting procedures were deemed unwarranted. Total percent agreement was used because agreement on occurrence and nonoccurrence are considered equally important. The proportion of agreement was evaluated using both the gold standard and the modal score. The kappa statistic was used to determine the proportion of chance corrected agreement between reviewers. This was accomplished through the evaluation of pair-wise ratings between each reviewer and the gold standard. Contingency tables were used to analyze the marginal distributions and dispersion of the paired ratings.

The variability in the dependent variable relative to each level of the independent variable was analyzed using an ANOVA randomized block design. Variance estimates were obtained for all the items. All possible two- way interactions between the scores and the severity of injury and the provider type were analyzed. The interaction effect of provider type and severity of injury on each level of the dependent variable was analyzed. The relationship between the provider type and judgments regarding the individual responsible for the adverse event were evaluated. A minimum significance level of 0.05 was used for the statistical tests. If the p value was < 0.05, the results were deemed statistically significant. The proportion of agreement will be considered adequate if 70% of the total responses were in agreement.

Chapter 4

Data Analysis and Findings

This chapter presents the results of the statistical analysis of the data and the findings of this study. Descriptive statistics pertaining to the claims and the analysis of the objective data are presented. The generalizability data will be presented in terms of the main effects and interactions. The data analysis will be summarized in relation to the research questions of this study.

Description of Claims

The events represented in these claims occurred between 1990 and 1995. In seven of the claims a nurse anesthetist working with an anesthesiologist provided the anesthesia, and in five claims CRNAs were the sole providers of the anesthetic care. The anesthetics were administered in hospital operating rooms in eleven of the claims and one adverse event occurred in an office setting. Of the eleven claims related to hospital experiences, two of these were in the labor and delivery area, one occurred in a special procedure room, and the remainder took place in surgical suites. The patients' ages ranged from 9 – 88 years of age, and the gender representation was equal. The ASA status ranged from 1 - 3, and two of the claims were classified as emergencies. There were equal numbers of inpatient and outpatient surgical procedures. In general the claims represented healthy individuals undergoing elective surgical procedures. The average patient was 46 years old without significant medical problems undergoing non-emergent surgery in a hospital setting. The anesthesia administered in the 12 study claims represented a variety of techniques. There was no relationship between the anesthetic technique and the severity of the adverse outcome.

Table 7. Description of Closed Claim Files

Types of Anesthesia Represented in Claims					
General		Regional		Other	
No.	Type	No.	Type	No.	Type
6	Intubations	1	Spinal Anesthetic	2	Monitored Anesthesia Care (MAC)
1	Mask	1	Epidural Anesthetic		
		1	Nerve Block		

The outcomes represented in the claims were as follows:

- 3 Deaths
- 2 Severe Incapacitating Brain Injuries
- Paraplegia
- Blindness
- 2 Burns
- Brachial Plexus Nerve Injury
- Prolonged Recovery Secondary to Pulmonary Complications
- Inter-operative Awareness

There was no relationship between the anesthesia provider, type of anesthesia, patient characteristics and the adverse event.

Objective Data

The intent of this study was to determine the interrater reliability on the seven subjective items. These questions represent less than 10% of the total data points on the instrument. In order to determine the degree of consensus that existed across all the items this study also included an analysis of the proportion of agreement on the objective questions. The goal of this inquiry was to identify the data points on which all eight

reviewers reached agreement across the twelve claims. This research revealed a substantial amount of missing and conflicting data in this portion of the instrument. It was not obvious if the inconsistencies resulted from variation within the closed claim files or transcription errors. The consensus items consisted of the sex of the patient, the anesthesia provider type, the site of anesthesia administration, the type of anesthesia and the date of the event.

The severity of injury score assigned by the reviewers was found to agree with that designated by the chairman in 94% of the judgments. The disparity between the outcome ratings never varied by more than two points on the injury scale. In instances when demographic information was needed and consensus was not present, the mode was used to describe the claims. This was done regarding the patient's age and ASA status as both of these items demonstrated variability between reviewers.

The large number of documents that comprise a closed claim file may be partially to blame for the lack of consensus on the objective information. The lack of agreement may reflect inaccuracies and conflicting data found in the medical records. The narrative portion of the instrument could be used to identify data problems that were noted in the original documents. The lowest agreement was noted on the items asking the reviewer to name the locations of where the adverse event occurred, and where the outcome became apparent. Agreement on whether a lawsuit was filed also had less than 80% agreement, as did the items concerning legal expenses and the disposition of the lawsuit.

A possible explanation for the inconsistent responses regarding where the injury became apparent is the options can be interpreted in a manner in which they are not mutually exclusive. If a patient's emergence occurred in the recovery room and this was

the location in which the adverse event became apparent, the selection of either the post anesthesia care unit or inter-anesthesia would be accurate. The lack of consistency between reviewers regarding if a lawsuit was filed is unclear. The randomness of the responses suggests that individuals are guessing on this item based on inferences in the claim documents. This type of variability was also present regarding the legal and settlement costs. These problems could be rectified easily by having the insurance carrier complete this part of the data collection form based on their records of the claim (Appendix D).

The documentation regarding the anesthetic agents administered indicates that some reviewers are entering all drugs and others are just entering the primary agents. Some reviewers are more meticulous in their completion of this portion of the instrument. A review of the objective portion of the data collection process has revealed that rater reliability issues are not restricted to the subjective responses. The level of consensus on the items that involve simple transcription should be in excess of 90%. Items that do not demonstrate this level of agreement require evaluation and revision by the reviewers.

Subjective Data

The purpose of this research was to determine the reliability of the composite ratings of the closed claim committee members. The reliability of the instrument was established in reference to this group of individuals using the intra-class correlation coefficient. This measure of consistency can only be generalized to these judges employing the same rating scale on a similar sample of closed claim files. The reliability for the instrument, calculated for the seven subjective items across all eight raters and twelve claims was 0.73. This value was statistically significant and the lower and upper 95%

confidence intervals equaled 0.47 and 0.91 respectively. Correlation coefficients of 0.70 according to Polit and Hungler (1983) are generally considered satisfactory, and an ICC of 0.70 on a newly developed instrument according to Washington and Moss (1988) is acceptable. The overall reliability of this instrument indicates that a relative degree of confidence may be placed in the data.

A necessary consideration in the interpretation of this coefficient is the restricted range of the data. The variance in the ratings was limited by the use of dichotomous data and the tendency of some reviewers to select the same response for certain items. This proclivity was demonstrated by the repeated selection of the unable to determine category for the majority of items by some reviewers. Item number two had fewer options than the other questions and resultantly had less variance. Item number seven demonstrated excellent agreement and this contributed to a lack of variance in the responses for that questions. The paradox of reliability coefficients is that they are not calculable if there is no variance in data. In these instances low reliability on claims that were homogenous on an item did not necessarily mean the raters were in disagreement.

The nominally scaled data was analyzed for interrater agreement using the percentage or proportion of uniform responses among the judges. Agreement regarding these judgments is absolute and disagreements are considered equally serious therefore, weighting procedures were not employed. The proportion of agreement was derived using both the gold standard and modal rating as the correct response. The kappa statistic provided a chance corrected measurement of pair-wise agreement. Each judge's ratings were compared with the gold standard to allow the evaluation of individual scoring patterns. The significance value of kappa tested the hypothesis that $k = 0$.

The advantage of analyzing data using contingency tables is that it can measure both agreement and consistency in the use of the instrument. The analysis of the distribution of data in this study using 2 x 2 contingency tables revealed some unique characteristics. The dispersion of data on several items was such that it rendered one of the agreement cells empty. This type of spread was evident in question two, which asks the reviewers to judge the ability of the records to provide an understanding of the details surrounding the event named in the lawsuit. In some paired comparisons the two raters never agreed that the records were inadequate, but agreed on the majority of cases that the records were adequate. Statistical measures such as kappa, phi, and lambda are derived from calculations in which the off diagonal cells may be empty, but assume 0 or negative values if one of the agreement cells is vacant. In these instances percent agreement is a more appropriate measure of concordance. There were also contradictory responses on some of the instruments by the reviewers. These issues will be analyzed as they pertain to each item and specific recommendations will be made.

The responses to the subjective items analyzed in this study are comparable with the results that were obtained on the 223 claims in the AANA database. This finding suggests the judgments concerning the twelve claims analyzed in this research are consistent with previous opinions. The chairman of the committee serves as the final judgment in the analysis of the closed claims, and this certainty formed the basis of designating the gold standard as the correct response in this evaluation. The failure of her response to consistently represent the majority supported the additional use of the mode. The mode exceeded the gold standard by 5 - 10 percentage points on items 1 – 6. The gold standard equaled the mode on item seven. There were no individual reviewers'

responses that were more highly correlated with those of the mode more than the designated gold standard.

Table 8. Comparisons Between Gold Standard And Mode Across Seven Items:

Item	Gold Standard	Mode
One	72%	77%
Two	70%	77%
Three	53%	63%
Four	59%	65%
Five	51%	59%
Six	59%	65%
Seven	83%	83%

Item One

Were inadequate pre-induction anesthesia activities the basis for the lawsuit?

The interrater reliability for this item was 0.37 and was not significant. The overall response to this item indicated that pre-induction activities did not contribute to the adverse event more often than they were contributory. Nineteen percent of responses deemed inadequate pre-induction activities to be the basis of the lawsuit, 68% judged preoperative activities noncontributory and 13% of the responses were unable to determine. The range of the kappa statistic for the question was 0.11 – 1.00, the mean kappa value was 0.53. This item demonstrated the second highest level of chance corrected agreement on the instrument. The average kappa value indicates a moderate amount of agreement exists on this item. The range of kappa for this item extended from poor to perfect agreement. The proportion of agreement on this item was exceeded only by item 7.

Table 9. Proportion of Agreement on Item One

Comparison	Agreement	Range
Gold Standard	72%	43 – 100%
Mode	77%	57 – 100%

The reasons for the disagreement on this item appeared to stem from a lack of uniformity in the interpretation of inadequate and adequate preoperative activities. In addition there is not a clear definition of the time span and actions encompassed by preoperative activities. This information needs to be delineated in the written instruction manual. A situation that elicited varied responses were those in which the patient was partially at fault for the adverse event by their own actions. There is not consensus whether this would fall under the auspices of pre-induction activities. The substance of the medical records may be partially responsible for the lack of agreement on this item. These documents do not provide insight into the specific causes of the adverse event. Critical information regarding the rationale for the anesthesia provider's decisions at the time of the incident are generally lacking in the claim file. Increasing the number of items concerning pre-induction activities and including specific instructions and definitions will improve the reliability of this item.

Item Two

What is the ability of the records to provide an understanding of the details surrounding the event named in the lawsuit?

This item had an interrater reliability of –0.17 and was not statistically significant. Seventy-seven percent of the responses determined the records to be adequate and 23% deemed the records inadequate in their ability to provide an understanding of the details surrounding the adverse event. The restricted range and lack of variability in responses contributed to the negative the value of the ICC. Two of the claims had perfect agreement on this item, two claims had only one reviewer that disagreed with all the other reviewers on this item, five claims had two disparate ratings, and three claims had three

judgments that varied from the mode on this item. The reliability measurement on this item is diminished by the homogeneity of the responses rather than a lack of agreement.

Table 10. Proportion of Agreement on Item Two

Comparison	Agreement	Range
Gold Standard	70%	0 – 100%
Mode	77%	57 – 100%

It is interesting to note that there was no relationship between judgments of inadequate on this item and judgments of unable to determine on the other subjective items. It was anticipated that if the reviewers judged the records to lack sufficient details to recreate the adverse event, there would be a corresponding increase in judgments of being unable to decide on the other items. An analysis of the data revealed that records deemed adequate had 16% more judgments of unable to decide than claims determined to be inadequate.

The kappa statistic for this question ranged from -0.13 to 0.44 with a mean value of 0.06. The level of agreement on this item can only be considered in the poor to fair range. A negative kappa signifies less agreement than would have been expected by chance alone. Given the common knowledge base and expertise of the reviewers the level of concordance on this item should be in excess of chance agreement. The kappa values derived from this data indicate serious problems exist with the interpretation of this question. A factor influencing the value of kappa on this item was a pattern of responses that rendered one of the agreement cells empty. Paired reviewers often never agreed that the records were inadequate in their ability to provide an understanding of events surrounding the adverse event, despite agreeing that the records were adequate in the majority of the claims. This type of dispersion was common between most pairs of

reviewers on this item, and accounts for the low kappa values. The resulting distribution is illustrated in the paired responses of rater two and the gold standard on this item (see Table 11).

Table 11. Comparison of Gold Standard to Rater Two on Item Two

Gold Standard	Rater Two		Total
	Adequate	Inadequate	
Adequate	9	2	11
Inadequate	1	0	1
Total	10	2	12

Given this data the kappa statistic will assume a value of –0.13. The G Index (Holley and Guiford 1964) was found to yield results more consistent with the actual agreement for this item. The following formula for the G index is based on the contingency table in Table 4.

$$G = \frac{(A+D)-(B+C)}{A+B+C+D}$$

Calculations using the same data result in a G value of 0.50, indicating a moderate amount of agreement. This value is more appropriate considering the reviewers were in 77% agreement with the mode on this question. Another factor that contributed to biased results on this item is that the value of kappa is greatest when the values for a given number of observer disagreements are equally divided between occurrence and nonoccurrence. Skewed error distributions occur when one of the error cells is disproportionately large. The inability of kappa to achieve values between 0 and 1 in the presence of confounding population and classification problems was described by Kraemer (1980). This distribution occurred on this item and is exemplified on the following contingency table:

Table 12. Comparison of Gold Standard to Rater Four on Item Two

Gold Standard	Rater Two		Total
	Adequate	Inadequate	
Adequate	6	1	7
Inadequate	4	1	5
Total	10	2	12

This type of dispersion may have biased the results of the kappa statistic on this item. The value of kappa derived from this data is 0.06 compared with 0.17 for the G index.

In addition to statistical concerns, several reviewer inconsistencies associated with the item were revealed. The analysis disclosed opposing opinions regarding the ability of identical documents to provide sufficient details to recreate the adverse event. This finding suggests that a common criterion for determining the adequacy of medical records is not in use and this deficiency has resulted in contradictory decisions by the reviewers. Another explanation for the disparate responses may be due to the fact that the claim documentation, specifically the anesthetic records are generated from direct participants in the adverse event. This dependency may limit the ability of the medical records to provide impartial information, and creates the need for conjecture regarding the etiology of the incident.

These records were evaluated prior to their inclusion in the study and were determined by the gold standard to have a sufficient degree of detail to allow the data instrument to be completed. The reasons why certain raters judged these claim files inadequate require clarification. A certain degree of speculation is always necessary in this type of review process. A premise of this study must be that the records will never include all the details surrounding the adverse event. There must be common expectations within the group regarding what constitutes an adequate degree of documentation.

Item Three

Was the anesthesia treatment by the CRNA appropriate?

This question had an ICC of 0.67 and was statistically significant. This reliability coefficient was equal to the ICC on item seven and represented the highest value attained. The reviewers found the anesthetic care inappropriate in 45% of the responses. The care was determined to be appropriate in 28% of the claims and impossible to judge in 27% of the responses. The kappa statistic on this question ranged from 0.14 to 1.00 with a mean value of 0.47. These values indicate a range of agreement from poor to perfect on this item. Although the percentage of agreement on this item was one of the lowest, the mean kappa value represented one of the highest on the instrument.

Table 13. Proportion of Agreement on Item Three

Comparison	Agreement	Range
Gold Standard	53%	14 – 71%
Mode	63%	43 – 71%

The reliance on implicit criteria for determining the adequacy of the anesthesia care was responsible for some of the disparate responses on this item. The reviewers are apparently basing judgments on knowledge and experiences that differ from one another. The analysis of this item emphasizes the need to use explicit criteria when possible. This may be accomplished by having the reviewers indicate if specific practice standards were violated. Currently there is excellent agreement between the reviewers regarding the sequence of events leading up to and following the adverse event. However, the rationale for judgments regarding the quality of care is lacking in the narrative summary. The reasons why the anesthesia care was determined to be inappropriate in nearly 50% of the claims is the foundation of the closed claims analysis. An evaluation of the decision

making process used by the reviewers in reference to this item may provide data that can be used to improve the reliability. The narrative summary should contain the information necessary to make this determination.

Item Four

Could the basis for the lawsuit have been prevented by the CRNA?

The ICC on this question was 0.23 and was not statistically significant. Fifty-one percent of the responses found that the CRNA could have prevented the adverse event from occurring. Twenty-five percent of the responses determined the CRNA could not have prevented the incident, and 24% were unable to determine if the CRNA could have prevented the mishap. The kappa statistic for this question ranged from -0.11 to 0.66 with a mean value of 0.31. This item was also subject to the same statistical difficulties encountered with item two. Disproportionate error cell sizes may have biased the results of the kappa statistic on this item. This item also demonstrated some unique properties in terms of the variability of responses. Some of the problems seemed to have evolved from the difficulties associated with predicting if different actions by the CRNA could have prevented the incident.

Table 14. Proportion of Agreement on Item Four

Comparison	Agreement	Range
Gold Standard	59%	14 – 86%
Mode	65%	43 – 86%

Item Five

Did a lack of CRNA vigilance contribute to the basis for the lawsuit?

This item had an ICC of -0.24 and was not statistically significant. This value signifies serious problems with this item. The distribution of responses to this question

were as follows; a lack of CRNA vigilance was determined not to be the cause of the adverse event 45% of the time, 25% of the responses to this item indicated that a lack of CRNA vigilance was a causative factor, and 30% could not determine if a lack of vigilance was the basis of the lawsuit.

Fifty-one percent of the responses to item four indicated that the CRNA could have prevented the incident. Item five demonstrated that a lack of vigilance was not considered to be the primary contributing factor. The average kappa statistic on this question was 0.16 and the range was from 0.25 to 0.40. In several instances paired reviewers failed to agree that a lack of negligence did not contribute to an adverse event. This was evident in situations where the anesthesia provider's lack of attention to the patient's condition was clearly apparent in the claim documents. This was exemplified in a claim in which the anesthesia provider's failure to closely monitor the patient's vital signs was directly responsible for the adverse outcome. This pattern of responses is exemplified by the comparison of rater three with the gold standard on this item (Table 15).

Table 15. Comparison of Gold Standard to Rater Three on Item Five

Gold Standard	Rater Three		Total
	Yes	No	
Yes	7	2	9
No	3	0	3
Total	10	2	12

The value of kappa based on the data in this example is -0.25 indicating substantially less agreement than would be due to chance alone. The G statistic for this item would be 0.17 slightly more agreement than would be expected by chance between the raters. The G statistic is a more consistent measurement of concordance than kappa

considering a 58% agreement between the two raters. Regardless of the statistic utilized the agreement on this item is poor. The wording of this item may account for some of the disagreement. Eliminating the use of negative terminology in this item may improve the consistency of the responses.

Contradictory responses on item number three with respect to this question were noted in four instances. In these situations the raters identified that a lack of CRNA vigilance contributed to the adverse event, and in response to item three these same reviewers judged the anesthesia care appropriate. A lack of vigilance is not generally associated with a standard of care that would be deemed appropriate. In the event this response pattern is justified clarification should be made in the narrative summary.

Table 16. Proportion of Agreement on Item Five

Comparison	Agreement	Range
Gold Standard	51%	28 – 86%
Mode	59%	43 – 86%

The analysis of this item revealed some serious problems that appeared to be associated with varied definitions of vigilance between reviewers. The response pattern for this item indicates that idiosyncratic criteria are being applied to this item. This question had the highest proportion of cannot determine responses, which suggests that the ability to extrapolate vigilance from written records is difficult. The reliability of this item is diminished by the ambiguity associated with the term vigilance.

Item Six

Could anyone have prevented the basis of the lawsuit?

The ICC on this question was 0.15 and was not statistically significant. Sixty-two percent of the responses to this item determined that someone could have prevented the

basis of the lawsuit. The analysis of this item found 10% of the responses found that no one could have prevented the incident and 28% could not determine if an individual could have interceded to stop the event from occurring. This question demonstrated a range of kappa values between each reviewer and the gold standard from -0.31 to 0.40. The average value of kappa across raters was 0.05 indicating only slightly more agreement on this item than would be expected based on chance alone. The ambiguity associated with the term anyone is one of the major problems with this item. The definition and scope of this item needs to be clearly defined and limited.

Table 17. Proportion of Agreement on Item Six

Comparison	Agreement	Range
Gold Standard	59%	14 – 86%
Mode	65%	43 – 86%

Reviewers appeared to have difficulty predicting if anyone could have prevented the outcome. The scores may reflect problems assigning blame to a specific individual when the anesthesia was provided by a team approach. Conflicting responses were noted between this question and item number 4. In three instances the reviewers answered yes to question four indicating that the CRNA could have prevented the incident, and then selected no to item six indicating that no one could have prevented the event. Some of the judges seem to be interpreting this question to mean anyone other than the CRNA, while others are including the CRNA in the other category. This interpretation would explain some of the disparate scoring. If this question was in closer proximity to the other item or was combined with number four it may improve consistency in scoring. The individual and the actions that may have prevented the incident should be discussed in the narrative summary.

Item Seven

Could better technical monitoring have prevented the adverse incident?

This question had an ICC of 0.63 and was significantly significant. Seventy three percent of the responses indicated that better technical monitoring would not have prevented the adverse incident. Eighteen percent of the responses found that additional monitoring techniques would have prevented the mishap and 9% were undecided. The mean kappa statistic for this question was 0.64, with a range of (0.43 – 1.00) indicating a substantial amount of agreement. The raters demonstrated consistency in their response to this item, and utilized the undecided category less than on any other item.

Table 18. Proportion of Agreement on Item Seven

Comparison	Agreement	Range
Gold Standard	83%	43 – 100%
Mode	83%	43 – 100%

This item demonstrated the highest reliability, percentage of agreement, and kappa values. The factors that separate this item from the other subjective questions is that it is clearly defined and specific. Item seven is based on explicit criteria in relation to the monitoring modalities employed by the anesthesia provider. In addition to demonstrating superior reliability this item produced specific recommendations for anesthesia practice. The advantages of adapting other items on the instrument to this format can be demonstrated by the information that was provided in the analysis of the twelve claims in this study.

In the claim involving an adverse event occurring in an office setting, a failure to monitor the patient's exhaled respiratory gases were noted. The monitoring of end tidal carbon dioxide is a standard of care for all patients undergoing general anesthesia. Every

rater recognized that regardless of the location of the administration of the anesthetic, standards of care must be consistent. The complexity and volume of surgeries performed in clinics and private offices is increasing with a concomitant increase in the number of general anesthetics performed in these locations. The closed claims database may provide a unique opportunity to study incidents that occur in office settings.

The other claim that produced specific monitoring recommendations involved a patient's complaint that she was inadequately anesthetized and consequently awake during her surgery. Lawsuits alleging inter-operative awareness have claimed injuries ranging from nightmares to severe psychological trauma. Historically, methods to monitor anesthetic depth were indirect and based on inferences from the patient's vital signs. Until the recent introduction of monitors capable of measuring brain wave activity under anesthesia the anesthesia provider had no direct method to measure a patient's level of consciousness. The advantage of these monitors is that they provide a means to insure that an adequate amount of anesthesia is administered.

The majority of the reviewers identified that the monitoring of brain waves under anesthesia using a Bispectral Index (BIS) or similar device may have prevented this claim. The benefit of the closed claims base is that it contains a number of similar cases of recall under anesthesia. The impact of the BIS monitor on the incidence and severity of intra-operative awareness can be evaluated using this data. The intent of this study was not to derive conclusions regarding anesthesia practice but to evaluate the reliability of the review process. This research found that the item determined to have the highest reliability also yielded the most valuable practice implications. The need for consistent application of standards of practice and the potential benefits of new technology represent

examples of data that can be generated from items that focus on the process and not the individual. The advantage of restricting the focus of the item to specific rather than global concepts was demonstrated by the reliability of this question.

A priori the proportion of agreement that was deemed adequate for the subjective data was 70%. This level was selected based on a review of the literature on similar instruments. The consensus of the behavioral scientists involved in these studies was that 70% was a necessary amount of agreement. The consensus reached with the mode equaled this level of agreement, and the agreement based on the gold standard was found to be below this acceptable limit. The degree of concordance found in this study is similar to the 72% agreement found in the initial analysis of interrater reliability of the AANA Closed Claim Study Data Collection Instrument. The dissimilar sample sizes and methodologies employed in these studies may explain the different results. The agreement between raters for each of the seven items is illustrated on Table 19. The overall agreement with the gold standard compared with the mode is illustrated on Table 20.

Table 19. Agreement Between Reviewers on Each of the Seven Items.

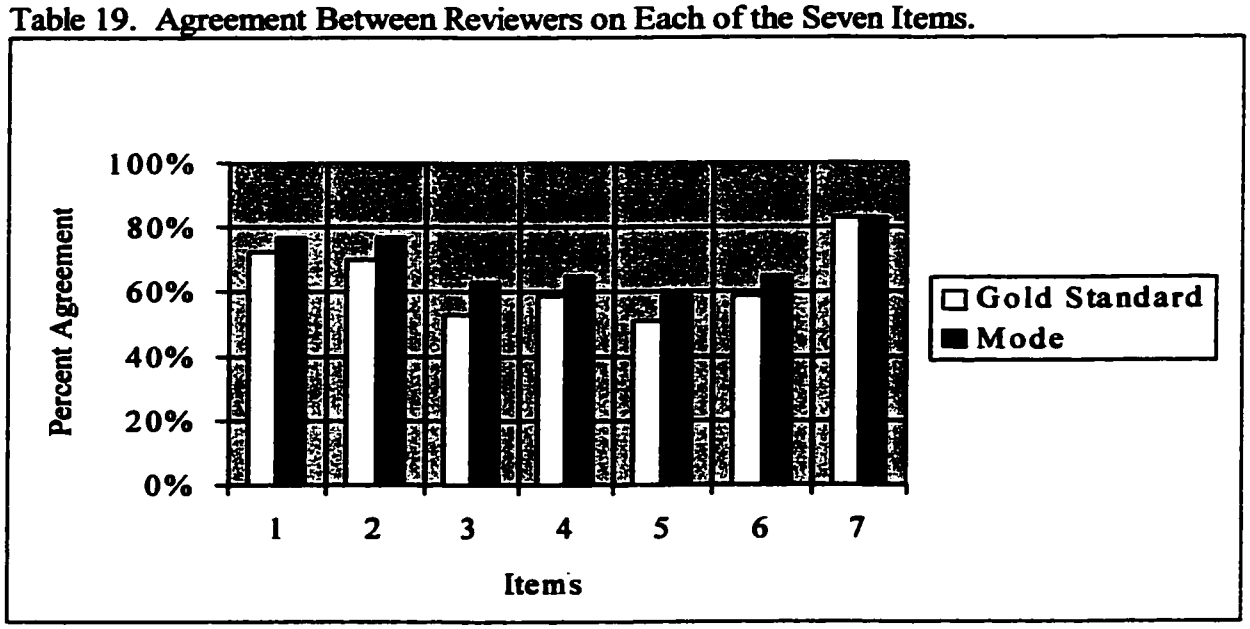
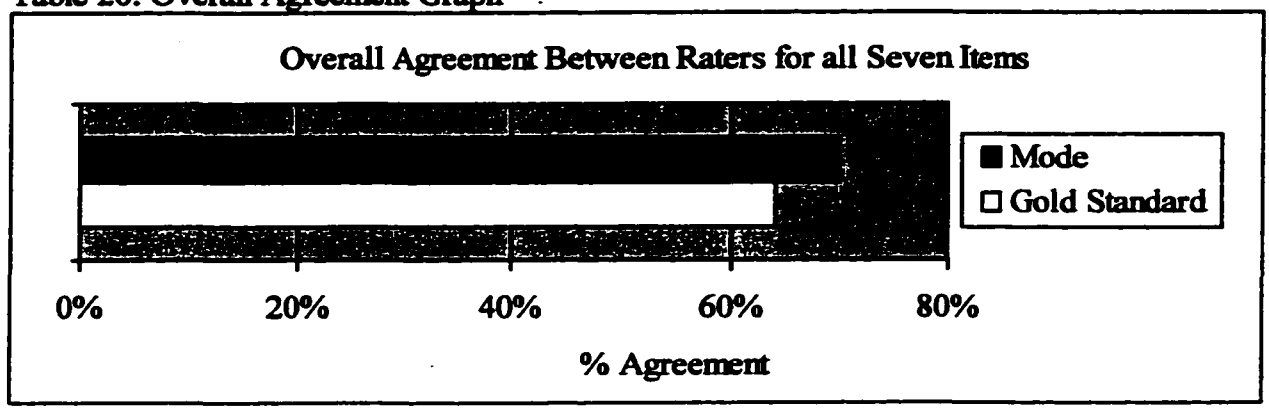


Table 20. Overall Agreement Graph



The effect of claim type on agreement was also evaluated in this study. The proportion of overall agreement for each claim was calculated, and the claims having the highest and lowest proportion of agreement were analyzed. The range of overall agreement for each claim ranged from 53 – 83%. The three claims that had the lowest level of overall agreement (53%, 59%, and 61%) shared several common factors. All the claims represented anesthesia care that was provided by both an MDA and a CRNA. The outcomes were all severe including two deaths and severe incapacitating brain damage. None of the adverse events could be attributed solely to the anesthetic technique. In each of the claims problems with ambiguous items contributed to the observed disagreement. Inconsistent interpretations of vigilance and preoperative activities lead to disparate opinions in these cases. In two of these claims preoperative activities on the part of the surgeon and the patient were subject to a wide range of interpretation. The need to clearly define the terms and restrict the focus of each item was exemplified in these three claims.

The three claims that had the highest level of overall agreement (75%, 81% and 83%) also had commonalities. These claims all represented outcomes that were clearly related to the anesthetic technique. One claim represented a brachial nerve injury

associated with mal- positioning and the other two outcomes could be attributed to the nerve block administered. Two of these claims represented solo CRNA practice and the third was a combined provider situation. The ability to identify a single individual as responsible for the incident was markedly improved in these claims. Varied definitions of vigilance were a source of disagreement on these claims but to a lesser extent than the previous three. The information provided by these claims emphasizes the impact of provider type and outcome on judgments.

Generalizability Data

A repeated measures analysis of variance was applied to the data. The influence of the severity of injury on judgments regarding the appropriateness of care was evaluated. The relationship between provider and judgments concerning responsibility for the adverse event was also analyzed. The main effects and interactions were not significant. The analysis of variance revealed the following:

- There was no difference in the severity of outcome based on the provider type.
- There was no difference in outcome based on anesthetic technique.
- The type of provider did not affect judgments regarding the appropriateness of care.
- There was no relationship between the provider type and judgments regarding the preventability of the incident by the CRNA or another health care provider.
- There was a relationship between the cannot determine selection and the provider type on items four and six.
- There was no relationship between the severity of the injury and judgments regarding the ability of the CRNA or anyone to prevent the incident.
- The judgments regarding vigilance were not related to the severity of injury score.

- There was no interaction effect regarding the severity of injury and provider type on any of the items.
- The respondents were more likely to judge the care appropriate in cases that had a SIS of I and inappropriate when the SIS was II, this finding was not statistically significant.

Research Question I

What is the reliability of an instrument developed to extract data from anesthesia related closed claims using kappa, percentage of agreement and generalizability theory?

Reliability in this instance refers to the ability of the judges in this study to reach the same conclusions regarding the medical management represented in the 12 claims without bias. The overall consistency of this instrument across all raters and items was 0.73. This coefficient indicates that the majority of the variance in the scores is due to differences in the substance of the claim, rather than between rater variability. This measure suggests the instrument has an adequate level of reliability with this group of judges.

An initial study of the interrater agreement on this instrument prior to revisions was 72%. The current instrument using the modal rating demonstrated an overall agreement of 70% with a range from 59 - 83%. The ratings using the gold standard as the correct response range from 51 – 83% with an overall agreement of 64%. These percentages represent poor to adequate levels of overall agreement across all categories. The results obtained in this study were similar to those reported by Caplan et al. (1988) in which the percentage of agreement on a similar instrument ranged from 54 to 80%.

The findings of this research project support the argument that significant information regarding data is concealed by the use of a single summarization coefficient. The overall agreement measured by the intra-class correlation coefficient and proportion of agreement indicate moderate to good overall concordance on the subjective items. However, the evaluation of the interrater reliability of the individual items demonstrated a range of values from poor to excellent. The reliability coefficients were influenced by the lack of variability in the data and therefore should be interpreted with caution. The disparity between the percent of agreement and the value of the ICC reflect the problems with deriving reliability coefficients using data with limited variance. This is demonstrated in item two which has the second highest percentage of agreement but the lowest reliability coefficient.

Table 21. Comparison of ICC with Percentage of Agreement and Kappa on Seven Items

Item	ICC	Mean Kappa	Percent Agreement
One	0.37	0.53	77%
Two	-0.17	0.06	77%
Three	0.67	0.47	63%
Four	0.23	0.31	65%
Five	0.24	-0.25	59%
Six	0.15	0.05	65%
Seven	0.63	0.64	83%

The kappa statistics derived from pair-wise comparisons identified individual raters exhibiting scoring patterns that varied appreciably from the group. The calculation of negative and zero values for kappa suggest that some of the items are not consistently interpreted between reviewers. In addition some raters selected the undecided category for a majority of their responses to certain items, and this contributed to the disparate patterns. These factors contributed to the inconsistencies between the proportion of agreement and the mean value of the kappa statistic. The application of several measures

of agreement offered different perspectives regarding interrater reliablity. This information would not have been available if the data were analyzed using only one method of analysis. The comparisons between the percentage of agreement and mean kappa values for each item are illustrated on Table 21.

Table 22. Comparison of Mean Kappa Values with Percentage of Agreement

Item	Mean Kappa Value	Percent of Agreement Gold Standard
One	0.53	72%
Two	0.06	70%
Three	0.47	53%
Four	0.31	59%
Five	-0.25	51%
Six	0.05	59%
Seven	0.64	83%

Research Question II

Does generalizability theory offer more insight into the data than traditional psychometric methods?

The analysis of the data within the framework of generalizability theory did provide details about the data that would be unavailable with traditional techniques. The application of an analysis of variance to the data allowed the evaluation of the relationship between the SIS and the judgments regarding the appropriateness of care rendered at the time of the adverse event. Although this study demonstrated that raters were more likely to judge the care appropriate if the patient suffered a temporary non-disabling injury, it was not statistically significant. This study did not find that judgments of impossible to judge increased as the severity of the injury worsened, as found by Caplan et al. (1988).

The application of generalizability theory to this data facilitated the evaluation of the relationship between the type of anesthesia provider and judgments regarding

responsibility for the incident. The analysis did reveal that there was a relationship between provider type and the number of cannot determine selections on items 4 and 6. When the care was provided by both a CRNA and MDA there was a significant increase in cannot determine responses compared to claims in which the care was provided by a CRNA only. This finding suggested that it is difficult to place blame on one individual when the anesthesia was provided using a team approach. This analysis did not find a relationship between provider type and the severity of the injury. There was no interaction effect between SIS and provider type with any of the items. The restricted range of the data and the lack random selection necessitate that the results be interpreted with caution.

Research Question III

Can the sources of variation identified in this analysis be used to improve the reliability of the instrument and data collection process?

The recommendations provided by this analysis can be used to improve the reliability of the closed claim review process. The current lay out of the data collection instrument is fragmented. Reorganizing the form to place items addressing common concepts within the same physical proximity would increase the reviewer's efficiency. The objective data points on the instrument were found to have missing, incomplete and conflicting data. This could be improved by having the chairman of the committee review all the completed forms prior to their submission. The insurance carrier should assume responsibility for the verification of data regarding legal proceedings, expenses and settlement costs. Their access to factual computerized records of these events will increase the consistency and accuracy of this data. This action would also save considerable time and expense for the committee members.

The issues associated with combined anesthesia practice must be incorporated into the design of the data collection instrument. Revising items that are poorly worded and eliminating ambiguous terminology will improve the reliability of the process. The definitions of vigilance and adequate were found to be inconsistent between the committee members. There were varied standards for determining the adequacy of records and preoperative activities. This analysis concluded that more than seven items are needed to address the complex issues involved in the determination of the quality of care provided during adverse events. The decomposition of each of the seven global concepts into specific and comprehensive items will improve the reliability of the process.

The conflicts in scoring between items four and six and three and five indicate there is variation in their interpretation. Some of these inconsistencies would be eliminated by the clarification of terms including the inclusion criteria for anyone. If conflicts are deemed appropriate these responses need to be explained in the narrative portion of the record. Increasing the length and specificity of the subjective portion of the instrument will improve the reliability of the process.

Chapter 5

Conclusion and Recommendations

Conclusions

This study measured the interrater reliability of an instrument developed to extract data from closed claims files. The primary goal of the study of closed claim files is the improvement of patient care through the identification of negative trends. The insight gained from this review will be disseminated to the anesthesia community through publications and educational programs. It is anticipated that information provided by the study of closed claim files will have a positive impact on future anesthesia practice. Because subjectivity is an inherent part of the review process the establishment of interrater agreement is a necessary condition before confidence can be placed in the data. This project utilized the recommendations of related studies in the research design. These suggestions included the use of the original documents in the analysis, and the evaluation of multiple items. Potential bias regarding the severity of the outcome and the type of anesthesia provider were evaluated. This study applied generalizability theory to the data based on the recommendations of previous research.

The results of this analysis should be interpreted cautiously due to the small sample size and the scale of the data from which it was derived. The sample of judges and claims used in this research were not randomly selected. The results of this research can only be generalized to this same group of reviewers using the study instrument with a similar sample of claims. Revisions in the instrument will require further studies of interrater reliability. The statistical analysis of the data will be discussed as it pertains to each of the research questions.

The instrument demonstrated an acceptable degree of overall agreement and interrater reliability. The proportion of agreement found in this study is comparable with instruments used for similar purposes. When this instrument was developed the nature of the data was unknown. It was assumed when this project was initiated that the claims would represent cases in which only CRNAs were responsible for the anesthetic. The files have been found to represent a multiplicity of practice situations and working relationships. The perplexities associated with assigning the responsibility to a specific individual when the anesthesia is provided within a team framework were not anticipated. The results of this research indicate that reviewers find it difficult to blame one person when the anesthesia is provided by both a CRNA and an MDA. The study of closed claims has proven to be a dynamic process requiring adjustments as the data is defined. The results of this research provide a baseline measurement of agreement and recommendations for improvement.

The application of multiple analyses to the data revealed different aspects of agreement that contributed to the total picture. The analysis of each item using the ICC and proportion of agreement revealed specific problems with each questions. An evaluation of individual responses on the subjective data using the kappa statistic demonstrated that some judges use a disparate scoring system. Istrument changes that would eliminate conflicting responses and decrease the use of the undecided category would improve the level of concordance. The application of generalizability theory to the data revealed aspects that would not be available with traditional methods. The narrative portion of the instrument was found to lack qualitative information necessary for understanding the rationale behind the reviewer's judgments.

Recommendations

The results of this study yielded several recommendations for improving the reliability and efficiency of the data collection process. The physical lay out of the instrument was found to contribute to rater error. Reorganizing the form to group similar items together would decrease incongruent responses and improve the efficiency of the process. There were several problems identified with the manner in which the items are written. This analysis found items that contain ambiguous terms that are subject to a wide range of interpretation. These terms include; appropriate, inappropriate, adequate, inadequate, vigilant and anyone. The elimination of vague terminology and responses that are not mutually exclusive are necessary adjustments. The number of items is insufficient to address the complex issues involved in this analysis. The incorporation of specific items based on explicit criteria will improve the consistency of interpretation. Items that allow causality to be attributed to multiple factors should be included on the instrument. These items have demonstrated excellent levels of agreement when used in the review of anesthesia morbidity and mortality. Increasing the number of items that address each global concept on the current instrument will improve the reliability of the process.

This analysis has revealed that objective data are not immune to rater error. Missing and conflicting data were found to be a significant problem. The need for each instrument to be meticulously checked by the chairman at the time of the review cannot be over emphasized. This process will facilitate the completion of the forms while the claim files are still readily available. The accuracy of the information regarding the filing of a law suit, the disposition of the lawsuit and claim, the expenses and the amount paid for the CRNA will be improved by delegating this responsibility to the insurance carrier.

CRNA will be improved by delegating this responsibility to the insurance carrier. Removing these items will improve the efficiency of the process by allowing the reviewers to concentrate on the anesthetic considerations.

The narrative portion of the instrument provides a brief description of events leading up to and including the adverse event. The summary also includes a detailed explanation of the injury and the patient's progress following the adverse event. Although, excellent agreement exists regarding the sequence of events, there is not a rationale for the judgments rendered. It has been suggested that a qualitative analysis of the written data may offer insight into the clinical decision making process. The instructions for the narrative summary currently do not require the reviewer to explain how they formed their judgments. The inclusion of the following points in the written summary may facilitate an understanding of how the reviewers derived the responses to the subjective data.

- Were specific standards of care breached? If yes, explain.
- If standards of practice were NOT violated explain the factors you believe contributed to the adverse event.
- What aspects of this case contributed to your overall impression regarding the quality of care the patient received?
- Provide your rationale for the subjective decisions made on the data collection instrument, items 1 – 7.
- In your opinion could different anesthesia management have prevented the adverse event? If so explain what could have been done.
- What recommendations would you make regarding the management of this case.

- If the outcome in this case were different would your judgment regarding the appropriateness of care be different.

The goal of the study of closed claims is to improve safety by the identification of negative trends and causes of patient injury. A perceived problem with this instrument is that the subjective items involve the identification of the individual responsible for the adverse event. This is demonstrated in items 3, 4,5 and 6. These questions ask the rater to decide if the CRNA, anyone, or their actions caused the adverse event. The difficulty in placing blame on an individual was evident in the number of unable to determine responses to these items.

Table 23. Unable to Determine Responses for Six Items

Item 1	*Item 3	*Item 4	*Item 5	*Item 6	Item 7
13%	27%	24%	30%	28%	9%

In a team approach to anesthesia the actions of the CRNA and MDA cannot be considered in isolation. The literature supports a peer review process that diverts attention away from the individual and toward the process. The focus of this analysis should be on practice standards and common factors contributing to adverse events. Revising the items in this manner may diminish the use of the undecided category and improve the reliability of the process.

The most recognized outcome study arising from the ASA Closed Claim Project involved the analysis of a group of patients that suffered unexpected cardiac arrests following spinal anesthesia. This collection of files enabled the analysis of a collection of similar incidents that otherwise would have been considered isolated incidents. This research by Caplan, Ward, Posner and Cheney (1988), revealed common factors involved

in these unexpected and devastating events occurring in healthy patients. Based on the findings of this study recommendations were made for the future treatment and prevention of similar incidents.

A similar study outcome study by Lee (2000) investigated postoperative visual loss in patients following non-ophthalmic surgery. This research identified common factors that contributed to decreased visual acuity and blindness postoperatively. The studies by Caplan (1988) and Lee (2000) provided valuable insight and recommendations for two events that have tragic consequences. Neither of these studies attempted to place blame on individuals or a specific individual's actions. This research demonstrates the positive results of studies that focus on the process rather than the individual. Increasing the use of explicit criteria may provide more valuable information for the AANA closed claims study and will identify strategies to improve patient care.

In medical litigation it is not uncommon for all parties involved in a patient's care to be named in the complaint regardless of the role they played. In surgical mishaps anesthesia may be named in the lawsuit even though it clearly bore no responsibility for the occurrence. These claims should not undergo extensive time consuming reviews. Claims unrelated to anesthesia provide no educational value and inclusion of these files into the database will artificially inflate the number of adverse events. The efficiency and accuracy of the process will be improved by the elimination of inconsequential claims.

Following verification that anesthesia contributed to the adverse event, the extent of the relationship should be defined. This can be accomplished using the Edward's Scale (Table 2), or another similar rating system. The Edward's Scale has demonstrated excellent interrater reliability in analyzing the contribution of anesthesia to adverse

outcomes. Reliable data regarding the etiology of the adverse event should be included on the data collection instrument.

Records that are determined inadequate to complete the collection instrument should also be eliminated. If the raters find the records require extensive speculation regarding the etiology of the event they should be omitted. Documents that require guessing or the predominant use of the unable to determine category will likely not provide reliable data. The chairman of the committee can serve to review the documents for this purpose. The records altered category on item two was never selected and should be eliminated. The number of items addressing the adequacy of the medical records should be increased. The addition of the following items will improve the reliability of the process related to the adequacy of the closed claim documents:

1. The documents contained sufficient information to make subjective judgments regarding the care.
2. The documents contained pertinent medical records but details surrounding the adverse event are missing. The formation of judgments regarding the quality of care requires some speculation.
3. The lack of documentation necessitated considerable speculation regarding the events associated with the adverse event.
4. The documentation does not appear to reflect the incidents as they actually occurred.

Problems with legibility could also be included as an option for this item as well. If a reviewer is unsure of the adequacy of a claim file a second level review by the chairman can be completed.

The demographic information is important to collect but is of little value if it is not reliable. This is true regarding the locations of where the incident occurred and was recognized. It would be of more value to know at what point in the patient's anesthesia care the adverse event occurred. This item should be revised to read:

At what point in the anesthetic management did the adverse event occur?

1. Preoperatively
2. During Induction of Anesthesia
3. During Maintenance of Anesthesia
4. Upon Emergence from Anesthesia
5. In the Post Anesthesia Recovery Period

These options are mutually exclusive and provide the opportunity to trend data regarding the stage of anesthesia when adverse events occur. This information is more beneficial in terms of practice recommendations than knowing in what area of the hospital the adverse event took place.

The preoperative period is characterized by the interaction of many factors that will ultimately impact on the patient's outcome. The role of the preoperative period should be evaluated in a manner that will yield consistent results regarding the multiple activities performed preoperatively. Increasing the number of items that address the myriad of factors included in this period of time will improve the reliability. Some examples of potential items include the following:

1. The preoperative activities were totally unrelated to the event.
2. The preoperative evaluation was through and comprehensive and was unlikely to have contributed to the adverse event.

3. The pre-anesthesia activities were inadequate and likely contributed to the adverse event.
4. The providers decisions based on the pre-anesthesia assessment were not consistent with standards of care.
5. The patient's preoperative actions and failure to disclose information contributed to the adverse event.
6. The preoperative evaluation failed to assess the patient consistent with acceptable standards of practice, this likely contributed to the adverse event.
7. The preoperative activities are poorly documented and likely contributed to the adverse event.
8. The documentation is adequate but it is unclear the role preoperative activities played in the adverse event.
9. The preoperative activities were directly responsible for the adverse event.

These or similar questions may facilitate the selection of the category that best represents the circumstances of the claim and will decrease the use of the unable to determine categories. These items would require specific instructions for completion including the time frame and actions that would fall under the auspices of pre-induction activities.

Determining the standard of care provided to the patient is critical if negative trends regarding the anesthetic management are to be identified. The items requiring the identification of the person responsibility for the adverse event are based on implicit criteria and subject to bias. The emphasis of this evaluation should be on the anesthetic process and not on the individual. Although practice standards and guidelines may be vaguely defined their use will increase the dependence on explicit criteria and should yield

a higher degree of interrater agreement. The identification of negligent practices such a failure to be vigilance and the provision of inappropriate care are important in closed claim research. However, the mere identification that these factors contributed to the adverse events is of little value without clear definitions of their meanings. The expansion of the number of items concerned with the degree of vigilance, and the appropriateness of care may help narrow the scope of these terms. The following items will address specific aspects of these global concepts:

1. Changes in the patient's status were recognized and treated appropriately.
2. Significant changes in the patient's status were apparently not recognized.
3. Changes in the patient's condition were recognized but not treated according to acceptable standards of care.
4. The records indicate a sequence of events that seems unlikely in light of the adverse event and outcome.
5. The patient was monitored based on acceptable standards of practice but the interpretation of the values was incorrect.
6. A lack of attention on the part of the anesthesia provider was directly responsible for the adverse event.
7. Despite apparent meticulous attention to detail and appropriate treatment the patient's condition deteriorated as a result of factors outside the anesthesia providers control.

The most significant benefit of increasing the number of items would be to improve agreement on occurrences and non-occurrences and decrease the selection of the unable to decide response. The reliability of the process will be improved by the inclusion of

specific items that in conjunction address the multifaceted concepts associated with vigilance and appropriate anesthesia care. This adjustment would facilitate the formation of specific recommendations for future practice based on the care provided during the adverse event.

The reliability of the process will not improve without significant revisions in the data. A sufficient number of items for each concept are necessary to assure confidence that agreement rather than a chance result is being measured. The main problem with the current process is that it is measuring seven separate outcomes using only one item each. Consequently the decisions based on these single items are likely to be unreliable. The decomposition of global concepts into comprehensive items is needed for every outcome the instrument is attempting to measure. The concepts currently addressed using one item include:

- Vigilance on the part of the CRNA during the adverse event.
- The quality of anesthesia care provided during the adverse event.
- The contribution of preoperative activities to the adverse event.
- The preventability of the adverse event.
- The role of technical monitoring in the adverse event.
- Contributions of other health care providers to the adverse event.
- The ability of the closed claim documents to recreate the adverse event and complete the data collection form.

The variance of the data will increase by expanding the number of items and this will improve the reliability of the process. Suggestions for additional items were discussed earlier in this chapter.

Items with dichotomous responses should be evaluated in terms of inter-rater agreement rather than reliability. In these situations agreement is present or absent and the degree of agreement is not an issue. Determining the proportion of agreement is appropriate to use with this data. However, this study demonstrated the problems associated with this descriptive measure when analyzing data with a small n. The percentage of agreement was greatly affected by only slight changes because of the small number of total possible agreements. An increase in the number of items would allow higher and lower frequencies to have a less profound influence on the percentage of agreement.

Increasing the number of claims evaluated is also necessary to increase the total n. More numerous files can be reviewed by having paired raters rather than by all eight reviewers evaluate each claim. Analyses of the paired responses should be supplemented with the G statistic for items with skewed distributions. In these situations this measure provides results that are more synchronous with the actual degree of concordance than the kappa statistic. A future study may compare the G statistic with kappa for contingency table data with a vacant agreement cell and disproportionate error cell numbers. An evaluation of the effect of larger sample sizes on these values is also recommended.

This analysis has indicated that the data provided by a generalizability study provides the opportunity to analyze main and interaction effects within an analysis of variance framework. This information would not be available using other techniques. The interaction between the severity of injury and judgments regarding the quality of care require continued evaluation. A larger sample size would facilitate the calculation of generalizability coefficients.

The obstacles preventing the attainment of high levels of interrater agreement on judgments regarding the quality of care provided during anesthesia related adverse events may be insurmountable. The low degree of agreement observed in these situations can be as much the fault of the subject matter as it is due to rater error. The complex issues involved in medical decision-making may limit the ability to obtain consensus on subjective judgments. The power to accurately measure agreement regarding quality of care issues may not be possible with classical statistical techniques or generalizability theory.

The serendipitous discovery that the narrative portion of similar data collection instruments contained valuable information was noted in the literature review. A qualitative analysis of the data contained within this summary may provide insight into the peer review process. This methodology may offer a depth and understanding of the data not available with other quantitative techniques. The expansion of the written summary to include the reviewer's rationale for decisions may facilitate this type of analysis. An alternative to expanding the summary would be to conduct taped interviews with the raters following their review of the closed claim files. The tapes could then be analyzed using a qualitative methodology. This technique has not been applied to inter-rater reliability studies in anesthesia to date.

The recommendations for future studies include the use of randomly paired reviewers, a larger sample size and multiple insurance carriers. The instrument should continue to be evaluated for interrater reliability as revisions are made. The addition of new reviewers as attrition occurs will also necessitate the determination of interrater reliability. A future study may evaluate the ratings of nurse anesthetists not directly involved in the closed claim study. The determination of reliability in this situation may

allow the external validity of the study to be expanded. Continued refinement of the quantitative and qualitative analysis involved in closed claim study will enhance the value of the data derived from this research.

The number of cases in the closed claim data base that reflect anesthesia care provided by both an MDA and a CRNA suggest that a collaborative research effort may be an appropriate approach to anesthesia safety. The ultimate goal of both groups in reviewing closed claims is the improvement of patient care. Despite political differences a united approach to patient safety may provide insight not available with independent reviews. The ASA has been involved in the review of claims since 1985 and has a larger base of insurance carriers than the AANA. They may have suggestions that will improve the AANA process based on their increased experience in closed claims research. Practice recommendations arising from a collaborative effort could be communicated to a broader group of anesthesia providers.

The study of closed claims provides a mechanism to evaluate and improve patient safety. This study identified that the current instrument is able to generate reliable data regarding the anesthesia care provided in these adverse events. Suggestions for improving the items were based on the data obtained from this study. The incorporation of these changes will improve the reliability of the process. The study of closed claims is ongoing and will continue to evolve as the practice of anesthesia advances. The reviewers, items and instrument will be modified as necessitated by this dynamic process. The continued assurance that the information obtained by this study is reliable is an integral part of the analysis of closed claim files.

APPENDIX A

American Association of Nurse Anesthetists

Foundation

AANAF

Closed Claim File Review Form

Reviewer: _____
Date: _____
State Filed: _____
Case #: _____

Did the adverse outcome clearly and unequivocally have nothing to do with the CRNAs activities? ____ Yes ____No
If *No* Explain: _____

Provider Information: Employment arrangement: _____ Age: _____ Certification date: _____

PATIENT INFORMATION

Date of Anesthesia Administration ___/___/___
Date Claim Reported _____
Date Claim Closed _____

Patient's age: _____ Years _____ Months _____ Days
Weight: _____ lbs _____ kgs _____ gms
Height: _____ ft _____ in Obese: ____ yes ____ no
Sex: ____ Male ____ Female

ASA Class: On Chart		Reviewer's Assessment
I	_____	_____
II	_____	_____
III	_____	_____
IV	_____	_____
V	_____	_____
E	_____	_____
Not Recorded	_____	Unsure ____

Race:
____ American Indian
____ Alaskan Native
____ Asian/Pacific Islander
____ Black
____ Caucasian
____ Hispanic
____ Unknown
____ Other _____

Drug Allergies: _____

Anesthesia Providers:
_____ CRNA
_____ Anesthesiologist
_____ SRNA
_____ Anesthesiology Resident
_____ Multiple Providers
_____ Other

Site of Anesthesia Administration:
Hospital
_____ Surgery Suite
_____ L & D Room
_____ Birthing Room
_____ Emergency Room
_____ Radiology
_____ ICU
_____ Catheterization
_____ Clinic/Procedure Room
_____ Other in-Hospital Location

Type of Facility:
_____ University
_____ Medical Center
_____ Community
_____ Rural
Nonhospital Facilities
_____ Ambulatory Surgery Center
_____ Dr Office
_____ Other _____

Those present during the event named in the lawsuit:
_____ CRNA _____ 2nd CRNA
_____ Anesthesiologist
_____ SRNA
_____ Anesthesiology Resident
_____ Surgeon
_____ RN
_____ Other

Admission Status
_____ Inpatient
_____ Outpatient - planned admit
_____ Outpatient

Planned Anesthesia Technique:
(1) Primary (2) Secondary
_____ No anesthesia was given
_____ Epidural
_____ General-endotracheal
_____ General-LMA
_____ General-mask
_____ IV regional
_____ Local
_____ Monitored anesthesia care
_____ Nerve block
_____ Spinal

Preinduction Activities:

Were inadequate preinduction anesthesia activities the basis for the lawsuit?
_____Yes _____No _____ Unknown

List what was inadequate about file preinduction activities
_____
_____
_____
_____
_____

Principal Surgical Procedure
_____
_____
_____
_____

Regional/Pain Management/MAC/Local

Regional Type:
_____ Axillary
_____ IV Regional
_____ Intrathecal
_____ Epidura
_____ Retrobulbar block
_____ Other _____

Pain Management:
_____ PCA
_____ Pain Block
_____ Other _____

Local/MAC:
_____ Sedation
_____ Local

Block administered by:
_____ CRNA
_____ Anesthesiologist
_____ Surgeon
_____ SRNA

Agent(s) Used By Providers

INHALANTS
_____ Air
_____ Desflurane
_____ Enflurane
_____ Halothane
_____ Helium
_____ Isoflurane
_____ Nitrous oxide
_____ Oxygen
_____ Sevoflurane
_____ Other _____

INTRAVENOUS AGENTS
_____ Etomidate
_____ Ketamine
_____ Methohexital
_____ Propofol
_____ Thiopental
_____ Other _____

ANALGESICS
_____ Alfentanil
_____ Fentanyl
_____ Meperidine
_____ Morphine
_____ Sufentanil
_____ Toradol
_____ Other _____

SEDATIVES/ HYPNOTICS
_____ Diazepam
_____ Droperidol
_____ Lorazepam
_____ Midazolam
_____ Other _____

ADJUNCTS
_____ Anticholinergics
_____ Anticholinesterase
_____ Antiemetics
_____ Benzodiazepine reversal
_____ H_2/antacids
_____ Narcotic
_____ Vasoactive
_____ Other _____

MUSCLE RELAXANTS
_____ Atracurium
_____ Curare
_____ Doxacurium
_____ Mivacurium
_____ Pancuronium
_____ Pipecuronium
_____ Rocuronium
_____ Succinylcholine
_____ Vecuronium
_____ Other _____

LOCAL ANESTHETICS
_____ Bupivacaine
_____ Cocaine
_____ Chloroprocaine
_____ Epinephrine added
_____ Efidocaine
_____ Mepivacaine
_____ Procaine
_____ Tetracaine
_____ Xylocaine
_____ Other _____

Primary Patient Position During Anesthesia Administration

_____ Jackknife
_____ Kneeling
_____ Lateral
_____ Lithotoni
_____ Prone
_____ Sitting
_____ Supine
_____ Trendelenberg
_____ Other
_____ Not documented

Could positioning have been a factor in the event named in the lawsuit? _____ yes _____ no

Were the patient's arms tucked into the patient's side?
_____ yes _____ no

Was documentation of positioning adequate?
a. position ___ Yes ___ no
b. padding ___ Yes ___ no

Patient Monitoring

Used	Values Recorded	
_____	XXXX	Fluid Therap
_____	_____	EKG
_____	_____	Blood pressure
_____	_____	Pulse oxinictrN
_____	XXXX	Stethoscope
_____	_____	Temperature
_____	_____	Anesthesia Machine Check
_____	_____	02 analwer-alarnis on
_____	_____	End-tidal Analwer
_____	_____	~ __CO_2 __N_2 __N_2O
_____	_____	Agent Analyzer
_____	_____	Inspired- __CO_2 __O_2
_____	_____	Circuit Disconnect
_____	_____	Peak Inspiratory Pressure
_____	_____	Peripheral Nerve Stimulator
_____	XXXX	Urinar\ Output
_____	XXXX	Blood Loss
_____	_____	Arterial Pressure
_____	_____	Central Venous Pressure
_____	_____	PA pressures
_____	_____	PCW pressures
_____	_____	Cardiac Output
_____	_____	Cardiac Index
_____	_____	Peripheral Vascular Resistance
_____	_____	Mixed Venous O_2 Saturation
_____	XXXX	Transesopliageal Echo.
_____	XXXX	Precordial Doppler
_____	_____	Intracranial pressure
_____	XXXX	Evoked Potentials
_____	XXXX	FHT ___ External ___ Internal
_____	_____	Labor
_____	_____	Other _____

Duration of Procedure: (Military time)	Start	End	Total time	Muscle Relaxant
Anesthesia				Relaxant Reversed: ___yes ___no ___unknown
Start to anesthesia End:	____	____	____hrs. ____min	Return of Muscle Strength Confirmed:
Incision to closure:	____	____	____hrs. ____min	___yes ___no ___unknown

PREEXISTING CONDITIONS

RESPIRATORY

_____Abnormal airway anatomy
_____Airway foreign body
_____Airway tumor
_____Asthma
_____Chest tubes
_____Hypoxemia
_____Laryngeal fracture
_____Limited neck ROM
_____Obstructive disease
_____Prieumonia/atelectasis
_____Restrictive disease
_____Shortness of breath
_____Smoker
_____Tracheostomy
_____URI
_____Ventilator support
_____Other

CARDIOVASCULAR

_____Aneurysm (type
_____Angina _____Stable _____Unstable
_____Cardiac shunt
_____Cardiomyopathy
_____Dialated
_____Hypertrophic
_____CHF
_____Dvsrlivthmia
_____Ejection fraction of less than 25%
_____Einbolus (type _____)
_____Hypertension
_____MI-6 mos. or more
_____MI-less than 6 mos.
_____Pacemaker
_____Peripheral vase. insufficiency
_____Shock
_____Valvular disease
_____Mitral valve prolapse
_____Other _____

CENTRAL NERVOUS SYSTEM

_____Comatose
_____Convulsive disorder
_____Cord injury Level
_____CVA
_____Dvstropliies/dystonia
_____Epi/subdural hematoma
_____Nerve Damage
_____Parkinson's disease
_____Peripheral neuropathics
_____Tunior/ancurysm
_____Visual/auditorv disorders
_____VP/VA shunt
_____Other

GASTROINTESTINAL

_____Ascites
_____Chronicaliv Debilitated
_____Esophageal abnormalities
_____Foreign body
_____Full stomacIVNPO less than 6 lirs.
_____Hiatal hernia
_____Motion sickness(NV
_____Obstruction/distention
_____Past histon, of postoperative NV
_____Peritonitis
_____Other

OB

_____Eclampsia
_____Gestational Diabetes
_____Dvsftinctioiial placenta
_____No prenatal care
_____Coagulopathics
_____Abnormal fetal presentation
_____Multiple gestation
_____Uncontrolled deliverv
_____Previous C-Section
Uterine rupture

ENDOCRINE/RENAL

_____Adrenal disease
____Diabetes Type I _ Type 11
_____Hyperthyroidism
_____ Hyporthyroidism
_____Liver disease/hepatitis
_____Pancreatic disease
_____Pancreatic disease
_____Plicochromocvtoiiia
_____Pituitary disorder
_____Renal dsfunction
_____Other

SPECIAL

_____Abnormal lab tests
_____Acid base/elect, disorder
_____Acutc/chronic alcoholism
_____AIDS/ininruno disorders
_____Arthritis
_____Aix-pical pseudocholinesterase
_____Blood dN scrasias
_____Cancer
_____Clicniotlicrap~
_____Chronic drug therap~
_____Connect. tissue disorder
_____Enzyme deficiency
_____Fetal anamolics type
_____Glaucoma
_____Hypertlicrinia > 10IF
_____HNpotherinia < 96F
_____Malignant hyperiliernua
_____MNasthcnia gravis
_____Neuronniscular disorder (type _____)
_____Obcsity
_____Poor dentition
_____Prcgnancy __ normal __ abnormal
_____Ps\chi;itric disorder
_____Sepsis
_____Substance abuse (type _____)
_____Thernial ittjurv
Tratirni

Contributed to the Basis for the Lawsuit

What was the basis for the lawsuit? _____

Indicate with a check mark the event(s) which contributed to the lawsuit

EQUIPMENT

Breathing circuit
- Disconnection
- Obstruction
- Leak
- Excessive pressure

Inadequate ventilation
Inadequate Fio_2
Other equipment
- Anesthesia machine
- Central line
- Epi/spinal catheter
- Peripheral IV
- Ventilator
- Other

AIRWAY INCIDENTS

Airway obstruction
Aspiration
Bronchial intubation – unplanned
Dental injury (Complete Dental Form)
Difficult intubation (Complete Airway Form)
Extubation - unplanned
Obstructed endotracheal tube
Postintubation croup/stridor
Postobstructive pulmonary edema
Reintubation - unplanned
Unrecognized esophageal intubation
Unsuccessful intubation
Vocal cord paralysis
Other _____

RESPIRATORY PROBLEMS

Abnormal breath sounds
ARDS
Bronchospasm
Cyanosis (visual)
Hypen,entilation-$PaCO_2$ < 25
Hvpoventilation-$PaCO_2$> 50
Hypoxemia-SaO_2 < 90%
Larvngospasm
Obstruction
Pneurmo/hemo/chylothorax
Pulmonary aspiration
Pulmonary edema
Postoperative ventilation - unplanned
Respiratory arrest
Subcutaneous emphysema
Other _____

CARDIOVASCULAR PROBLEMS

Acute MI
Air or other eiriboli
Asvstole
Atrial fib/flutter
Cardiac Arrest
Chest pains
CHF
CoagulopathN
Dvsrhvthmias
Excessive blood loss
Hypertension - > 110 diastolic
Hypotension - < 80 systolic
Hypovolemia
Intra-art. injection – unplanned
New PVCs - > 5/minute
New ST segment changes
Sinus bradycardia < 40
Sintis tachycardia > 120 (child > 180)
Transfusion reaction
Ventricular fibrillation
Ventricular tachycardia
Other _____

CENTRAL NERVOUS SYSTEM PROBLEMS

Anticholinergic crisis
Anoxia Enccplialopathy
CVA
Delayed arousal period > 4 lirs PO
Extrapyramidal symptoms
Failure to return to preanesthesia LOC
Hyperthermia tenip > 101F CORE
Hyporthermia tenip < 96F CORE
Seizure
Unplanned patient awareness
Other _____

ENDOCRINE/RENAL PROBLEMS

Acute tubular necrosis – ATN
Adrenal suppression - CV collapse
Anuria
Hyperglycemia > 240
Hypoglycernia < 50
Hyperkaleniia > 6.0
Hypokaleinia < 3.0
Oliguria < 0.5 cc/kg/hr
Thyroid storm
Other _____

INTEGUMENTARY PROBLEMS _____
Blindness _____
Burn: 1st __ 2nd __ 3rd __ Where: ___ _____
Corneal abrasion _____
Ecchymosis _____
Infiltrated IV _____
Laceration _____
Other _____

PERIPHERAL NERVE COMPLICATIONS
Pressure/Nerve damage: _____
Unilateral _____ Bilateral _____
Ulnar _____
Radial _____
Median _____
Brachial plexus _____
Femoral _____
Sciatic _____
Femoral and sciatic _____
Lumbosacral nerve root _____

Paraplegia _____
Quadriplegia _____
Other _____
Etiology of nerve injury _____
Probably positional _____
Possibly positional _____
Block related _____
Surgen, related _____
Preexisting nerve damage _____
Unclear inechanisir./insufficient data _____
Other _____

REGIONAL _____
Epidural xvet lap _____
Failed block _____
Greater than 3 attempts at insertion _____
Hematoina _____
High sub/epidural block with CV _____
or resp. insufficiency _____
IV injection of local anesthesia _____
Persistent paresthesia _____
S/S local anesthesia toxicitv _____
Spinal headache (posdural puncture) _____
Sustained hypotension > 15 min. _____
Other _____ _____

ANESTHETIC/MEDICALPROBLEMS
Adverschinexplained anesthesia reaction _____
Allergic/anaphylactoid reaction _____
Drug/fluid error _____
Inadequate reversal of relaxants _____
Narcotic excess requiring naloxone _____
Postop delirium requiring treatment _____
Other _____ _____

MISCELLANEOUS
Alternative anesthesia plan required _____
Arterial/central line complication _____
Back pain _____
Case canceled preinduction _____
Case canceled postinduction _____
Death _____
Induction delaNed > 30 nun _____
Malignant hyperthermia _____
Patient/familv dissatisfaction _____
Protracted vorniting _____
Pruritis _____
Surgical complication _____
Unexpected return to surger.N _____
Unplanned admission to ICU _____
Unplanned outpatient surgical admission _____
Unplanned PACU over 2 hours _____
Wrong blood product _____
Other_____ _____

OB COMPLICATIONS
Apgar _____
Birth injuries _____
Cardiovascular depression _____
Coagulopathics _____
Delive" _____
C-section _____
Uncontrolled _____
Vaginal _____
Vaginal (assisted) _____
Dysfunctional placenta _____
Eclampsia _____
Fetal anomalies _____
Fetal brain damage _____
Fetal death _____
Fetal distress _____
Fetal presentation abnormal _____
Fetal respirator-N depression _____
Fetal seizure _____
Gestational diabetes _____
Multiple gestation _____
No prenatal care _____
Previous C-section _____
Uterine rupture _____
Other _____ _____

Event named in lawsuit took place in:

_____ Intra-anesthesia (induction to emergence)
_____ Emergency Room
_____ ICU
_____ In transit
_____ PACU
_____ Pre-op hold_
_____ Nursing-umt
_____ After discharge from facility
_____ Other _____

Injury became apparent:

_____ Intra-anestliesia (induction to emergence)
_____ Emergency room
_____ ICU
_____ In transit
_____ PACU
_____ Pre-op hold
_____ Nursing unit
_____ After discharge from facility
_____ Other _____

Could the basis for the lawsuit been prevented by the CRNA? _____ Yes
_____ No
_____ Cannot be determined

Did a lack of CRNA vigilance contribute to the basis for the lavvsuit?
_____ Yes
_____ No
_____ Cannot be determined

Could awne have prevented the events that led to the lawsuit?

_____ Yes, Explain _____
_____
_____ No
_____ Cannot be determined

Regarding the Lawsuit:
Was a lawsuit filed? _____Yes _____No

If "Yes", indicate the disposition of the lawsuit:
_____ Dismissed by court
_____ Dissmissed by settlement
_____ Judge Trial
_____ Jury Trial
_____ Arbitration/Mediation

If "no", indicate the disposition of the claim:
_____ Dismissed by settlement
_____ Arbitration/Mediation
_____ Dropped/discontinued

Amount Paid:
_____ Expenses
_____ Amount Paid for CRNA

Ability of the records to provide an understanding of the details surrounding the event named in the lawsuit:

_____ Adequate
_____ Inadequate
_____ Records altered

Was the anesthesia treatment bv the CRNA:
_____ Appropriate
_____ Inappropriate
_____ Impossible to judge

Could better technical monitoring probably have prevented the event named in the lawsuit?
_____ yes _____no _____ undecided
If yes, what kind?
_____ End tidal CO_2
_____ Pulse oxiiiiefer
_____ Othcr

Severity of the Injury:

Extremes:	___ No injury	___ Emotional only		___ Death
Temporary:	___ Insignificant	___ Minor	___ Major	
Permanent:	___ Minor	___ Significant	___ Major	___ Grave

Indicate the types of documents you reviewed when completing this form:

_____ Anesthesia record
_____ Surgeon's operative note
_____ PACU record
_____ X-rays, lab test, toxicology reports
_____ Discharge summary
_____ Follow-up evaluation by medical consultants or primary caregiver
_____ Other medical records _____
_____ Deposition transcripts or summaries
_____ Narrative from involved parties
_____ Autopsy report
_____ Photographs of patient or equipment
_____ Expert or peer reviews
_____ Claims manager evaluation. notes or summary
_____ Attornev evaluation. notes or surninan
_____ Economic analysis of damages (by economist)
_____ Other _____

NEGATIVE PATITENT OUTCOME:

Death _____ Brain Injury_____ Nerve Injury _____ Eye Injury _____ Burns Other (Describe) _____

BRIEF SUMMARY OF EVENTS: Please give a succinct narrative of the events describing the primary cause and the secondan events. Specify the sequence of events, especially details not included elsewhere on the form. and details pertaining to the quality of anesthetic care:

APPENDIX B

Permission to Conduct Research and Utilize Data Pertaining to

American Association of Nurse Anesthetists Foundation

AANAF

Closed Claim Study

FOUNDATION

February 5, 2001

Karen Crawforth, CRNA, MS
1709 West Ridge
Rochester Hills, MI 48306

Dear Ms. Crawforth,

We have reviewed your letter of request to use Closed Claim data for your doctorate study. You are more than welcome to use the data, keeping in mind that information is reported in the aggregate with no possible patient or provider identification.

We wish you the best of luck in pursuing your doctorate studies. If we can be of further help, please let us know.

Sincerely,

Lorraine Jordan

Lorraine Jordan, CRNA, PhD
Director of Research and AANA Foundation

AMERICAN ASSOCIATION OF NURSE ANESTHETISTS FOUNDATION
222 So. Prospect Ave. Park Ridge, IL 60068-4001 • Phone: (847) 692-7050 Ext. 3070 • Fax: (847) 692-7137 • http://www.aana.com

APPENDIX C

American Association of Nurse Anesthetists

Foundation

AANAF

Instructions for Completing the

Closed Claim File Review Form

American Association of Nurse Anesthetists
Instructions for Using Claim File Review Form
General Information

Page 1 of Form

Reviewer
Write your first and last name.

Date
Write the date you are reviewing the claim.

State Filed
Write the initials of the state where the claim was filed.

Case Number
Write the claim file number.
(This number can be found on the orange card on top of the file.)

Forms Completed
Indicate the forms (obstetric, airway, cover-up, dental, or behavior) that were filled out for the claim.
In obstetric cases two (2) forms need to be completed, one (1) for the mom and one (1) for the baby.

Emergency
Please indicate if this case was an emergency.
This should be separated out from the other forms if this is correct.

Provider Information
Write the type of employment arrangement between the facility and the anesthesia provider involved in the claim (i.e., provider employed by facility or anesthesia group or is an independent contractor.)

Age
Indicate the age of the provider, if known.
Indicate unknown, if not available.

Certification Date
Write the anesthesia provider's date of certification.
Indicate "unknown," if information cannot be determine.

Date of Anesthesia Administration
Write the month, day and year of anesthesia administration.

Date Claim Closed

Write the month, day and year the claim file was closed.
(This should be indicated on the orange card on the top of the file.)

Patient's Age

Write the patient's age.
If the patient is older than three (3) years, write the patient's age in the "years" blank.
If the patient is younger than three (3) years, write the patient's age in the "years", "months" and "days" blanks.
Enter the number "0" in the "months" blank if the patient experienced an event that was alleged to have caused a damaging patient outcome in utero or at birth.

Weight

Enter the patient's weight.

Height

Write the patient's height.

Obese

Indicate if the patient is obese (patient is 20% to 30% over their ideal body weight.)

Sex

Indicate the sex of the patient.

ASA Class

Indicate the ASA classification assigned by the provider as indicated in the record.
Mark "not recorded" at the bottom of the "on chart" column. if the ASA classification is not documented.

Reviewer's Assessment

Indicate what ASA status you would assign the patient.
Mark "unsure," if you are not able to make a decision.

Race

Indicate the patient's race.
Mark "unknown," if the patient's race is not documented.
Indicate "other," if the patient's race is different from those listed.
Write the patient's race.

Drug Allergies

Write the patient's drug allergies.

Anesthesia Providers

Indicate the type of anesthesia provider responsible for providing the anesthesia care.

Select "multiple providers," if two or more anesthesia providers provided anesthesia care. (You may also check the individual providers. For example, if you select multiple providers, you may also select CRNA and anesthesiologist, so there will be three (3) checks.)

Mark "other," if the anesthesia provider's practice is not anesthesiology (e.g., dentist, podiatrist and oral surgeon.)

Site of Anesthesia Administration

Indicate where in the hospital the anesthesia care was administered.

If the anesthesia care was administered in a location other than those listed, please indicate under "other".

Type of Facility

Indicate the type of facility (university, medical center, community, rural, non-hospital facility) in which the anesthesia was administered.

If the anesthesia was administered in a facility other than a hospital, indicate the type of facility. Mark "other" if the location is different from those listed.

Write the non-hospital location in which the anesthesia was administered.

Those Present During the Event Named in the Lawsuit

Indicate the type of anesthesia provider(s) who were present during the event that was alleged to have caused a damaging patient outcome.

The damaging event is the specific incident or mechanism that led to the adverse outcome. The adverse outcome is the injury sustained by the patient. If an elderly patient fell off the OR table, this would represent the damaging event and the resultant hip fracture would be the adverse outcome.

Admission Status

Identify if the patient was an inpatient or an outpatient.

Planned Anesthesia Technique

Write number one (1) in the appropriate blank to indicate the primary (first) anesthesia technique that was used.

Write number two (2) in the appropriate blank if a secondary (second) anesthesia technique was used. For example, if the case started under epidural and converted to general, it would be coded (1) epidural, (2) general.

Mark "no anesthesia was given," if anesthesia was not administered. For example, if the damaging event occurred in the emergency room .

Preinduction Activities

Indicate if inadequate preinduction anesthesia activities were related to an event that was alleged to have caused a damaging patient outcome.

Mark "unknown," if it cannot be determine if inadequate preinduction anesthesia activities were related to an event that was alleged to have caused a damaging patient outcome.

If preinduction activities were inadequate, list what was inadequate about the preinduction activities, (e.g., the fluid level in the vaporizer was not checked prior to induction which resulted in patient recall.)

Principal Surgical Procedure

Write the name of the principal surgical procedure that was performed.

Complaint

Indicate if the complaint was related to the surgical procedure (i.e., wrong foot amputated.)

Page 2 of Form

Regional / Pain Management / MAC (Monitored Anesthesia Care) / Local

Skip this section, if regional anesthesia was not used.

If this section is skipped, draw a line through the entire box to indicate this that this section does not apply to the claim being reviewed.

Regional

If regional anesthesia was used, indicate the type of anesthesia used.

Mark the "intrathecal" blank only if this was the planned site of injection.

If the type of anesthesia was different and simply acted intrathecally, *do not* mark the "intrathecal" blank. Instead, indicate the original site of injection.

Mark "other," if the type of anesthesia used is not listed.

Write the type of anesthesia used.

Pain Management

Indicate the type of pain management used.

Mark "other," if the type of pain management used is not listed.

Write the type of pain management used.

Local / MAC

Indicate whether sedation or local anesthesia was used.

Block

Indicate who administered the anesthesia block.

Primary Patient Position During Anesthesia Administration

Indicate the patient's primary position during anesthesia administered.
Mark "other," if the patient's primary position is different from those listed. Write the patient's primary position during anesthesia administration.
Mark "not documented," if the patient's primary position is not documented.
Indicate if positioning was a factor in the event that was alleged to have caused a damaging patient outcome.
Indicate if documentation in the patient care record about the padding the patent was adequate.
Indicate if adequate ventilation was confirmed after the patient's position was changed.

Agent(s) Used By Providers

Indicate the inhalants, intravenous, analgesics, sediative/hypnotics, adjuncts, muscle relaxants and local anesthetics agent(s) used by the anesthesia provider.

Patient Monitoring

Indicate the type of monitoring that was used in the "used" column.
Indicate the values that were documented for the type of monitoring that was used in the "values recorded" column, except for those values that are marked with an "x".

Page 3 of Form

Duration of Procedure

Complete this section, if anesthesia or surgical times are indicated in the claim file.
Write the time (military) that anesthesia administration was started and ended.
If a surgical incision was made, write the time (military) the surgical incision was made and when the incision was closed.
Write the total time for anesthesia administration and surgery.

Muscle Relaxant

Complete this section only, if non-polarizing muscle relaxants were used.
Indicate if a reversal agent was administered to a patient who received a muscle relaxant.
Mark "unknown," if it cannot be determined that a reversal agent was administered to a patient who received a muscle relaxant.
Indicate if it was confirmed that the patient's muscle strength had returned after receiving a reversal agent.
Mark "unknown," if it cannot be determined that the patient's muscle strength had returned after receiving a reversal agent.

Preexisting Conditions

Indicate the respiratory, cardiovascular, central nervous system, gastrointestinal, OB, endocrine/renal and special preexisting conditions of the patient.

In each of these sections mark the "other" blank, if the preexisting condition is not listed.

Write the preexisting condition.

Page 4 of Form

Contributed to the Basis for the Lawsuit

Write the event(s) that occurred during anesthesia patient care that was alleged to have caused a damaging patient outcome.

Indicate the equipment, airway incidents, respiratory problems, cardiovascular problems, central nervous system problems and endocrine/renal problems event(s) that as alleged to have caused a damaging patient outcome.

Equipment

Indicate the equipment (e.g., monitors, warming devices and catheters) that malfunctioned while being used or that was used improperly.

Mark "other" if the type of equipment that malfunctioned is different from those listed.

Write the type of equipment that malfunctioned.

Airway Incidents

If "dental injury" is marked in the "airway incidents" section, complete the Dental Claims Form. It is not necessary to complete the comprehensive Claim File Review Form.

If "difficult airway" is marked in the "airways incidents'" section, complete the Difficult Airway Data Collection Form.

Mark "other," if the type of airway incidents are different from those listed.

Write the type of airway incidents that were alleged to have caused a damaging patient outcome.

Respiratory Problems

Mark "other," if the type of respiratory problems are different from those listed.

Write the type of respiratory problems that were alleged to have caused a damaging patient outcome.

Cardiovascular Problems

Mark "other," if the type of cardiovascular problems are different from those listed.

Write the type of cardiovascular problems that were alleged to have caused a damaging patient outcome.

Central Nervous System Problems
Mark "other," if the type of central nervous system problems are different from those listed.
Write the type of central nervous system problems that were alleged to have caused a damaging patient outcome.

Endocrine / Renal Problems
Mark "other," if the type of endocrine / renal problems are different from those listed.
Write the type of endocrine / renal problems that were alleged to have caused a damaging patient outcome.

Page 5 of Form

Integumentary Problems
Mark "other," if the type of integumentary problems are different from those listed.
Write the type of integumentary problems that were alleged to have caused a damaging patient outcome.

Peripheral Nerve Complications
If "unilateral pressure / nerve damage" or "bilateral pressure / nerve damage" is marked, mark the nerves that were damaged.
Mark "other" if the type of pressure / nerve damage is different from those listed.
Write the type of pressure / nerve damage that was alleged to have caused a damaging patient outcome.
If "unilateral pressure / nerve damage" or "bilateral pressure / nerve damage" is marked, mark the etiology of the pressure / nerve damage.
Mark "other," if the etiology of the pressure / nerve damage is different from those listed.
Write the etiology of the pressure / nerve damage that was alleged to have caused a damaging patient outcome.

Regional
Mark "other," if the type of regional anesthesia problems are different from those listed.
Write the type of regional anesthesia problems that were alleged to have caused a damaging patient outcome.

Anesthetic / Medical Problems
Mark "other," if the type of anesthetic / medical problems are different from those listed.
Write the type of anesthetic / medical problems that were alleged to have caused a damaging patient outcome.

Miscellaneous

Mark "other," if the type of miscellaneous problems are different from those listed.

Write the type of miscellaneous problems that were alleged to have caused a damaging patient outcome.

OB Complications

Mark "other," if the type of obstetric complications are different from those listed.

Write the type of obstetric complications that were alleged to have caused a damaging patient outcome.

Page 6 of Form

Event Named in Lawsuit Took Place In:

Indicate the location where the event that was alleged to have caused a damaging patient outcome occurred.

Mark "other," if the event that was alleged to have caused a damaging patient outcome took place in a different place from those listed.

Write the location where the event that was alleged to have caused a damaging patient outcome occurred.

Injury Became Apparent

Indicate the location where the event that was alleged to have caused a damaging patient outcome was first noticed.

Mark "other," if the location where the event that was alleged to have caused a damaging patient outcome was first noticed is in a different place from those listed.

Write the location where the event that was alleged to have caused a damaging patient outcome was first noticed.

Could the CRNA Prevented the Basis for the Lawsuit

Indicate if the CRNA could have prevented the event that was alleged to have caused a damaging patient outcome.

Mark "cannot be determined," if it cannot be determined if the CRNA could have prevented the event that was alleged to have caused a damaging patient outcome.

Did Lack of CRNA Vigilance Contribute to the Basis for the Lawsuit

Indicate if the lack of CRNA vigilance contributed to the event that was alleged to have caused a damaging patient outcome.

Mark "cannot be determine," if it cannot be determined if the lack of CRNA vigilance contributed to the event that was alleged to have caused a damaging patient outcome.

Regarding the Lawsuit

Indicate if a lawsuit was filed.
If yes, mark the blank to indicate the disposition of the lawsuit.
If no, mark the blank to indicate the disposition of the lawsuit.

Amount Paid

Write the final dollar amount paid by all parties named in the claim or lawsuit, excluding expenses.
Write the dollar amount paid for the nurse anesthetist named in the claim or lawsuit, excluding expenses.
Write the dollar amount paid for expenses related to the claim or lawsuit.
Mark "unknown" if the expenses of the claim or lawsuit cannot be determined.

Severity of the Injury

Indicate if the severity of the injury the patient suffered. Indicate the injury as being no physical injury, emotional injury, temporary (insignificant, minor, major), permanent (minor, significant, major grave) or death. Indicate only one level of severity.

Table 1 - Severity of Injury Scoring System

Severity Scale (Score)	Example
No obvious injury	
Emotional injury only	Fright, awake during anesthetic , pain during anesthetic
Temporary Injury	
Insignificant	Lacerations, contusions, no delay in recovery
Minor	Fall in hospital, recovery delayed (extra time in recovery room or hospital)
Major	Nerve damage unable to work, prolonged hospitalization
Permanent Injury	
Minor	Damage to organs, nondisabling injuries
Significant	Loss of eye, deafness, loss of one kidney or lung
Major	Paraplegic, loss of use of limb, blindness, brain damage
Grave	Severe brain damage, quadriplegia, lifelong care or fatal prognosis
Death	

Ability of the Records to Provide an Understanding of the Details Surrounding the Event Named in the Lawsuit

Indicate if the patient care records provided an adequate description of the details surrounding the event that was alleged to have caused a damaging patient outcome.

Mark "records altered," if the patient care records were altered.

Could Anyone Have Prevented the Event that Led to the Lawsuit

Indicate if anyone could have prevented the event that was alleged to have caused a damaging patient outcome.

If "yes" is marked, explain how anyone could have prevented the event that was alleged to have caused a damaging patient outcome.

Mark "cannot be determined," if it cannot be determined if anyone could have prevented the event that was alleged to have caused a damaging patient outcome.

Was the Anesthesia Treatment Appropriate

Indicate if the anesthesia treatment was appropriate.

The term "appropriate" refers to anesthesia care that is reasonable and prudent according to the standards of anesthesia care that was provided when the event that was alleged to have caused a damaging patient outcome occurred.

The term "inappropriate" refers to anesthesia care that is not reasonable and prudent according to the standards of anesthesia care that was provided when the event that was alleged to have caused a damaging patient outcome occurred.

Mark "impossible to judge," if it cannot be determined if the anesthesia treatment was appropriate.

Would Better Technical Monitoring Probably Have Prevented the Event Named in the Lawsuit

Indicate if better technical monitoring probably could have prevented the event that was alleged to have caused a damaging patient outcome.

If "yes" is marked, indicate the type of technical monitoring that probably could have prevented the event that was alleged to have caused a damaging patient outcome.

Mark "other," if the type of technical monitoring that probably could have prevented the event that was alleged to have caused a damaging patient outcome is different from those listed.

Write the type of monitoring equipment that probably could have prevented the event that was alleged to have caused a damaging patient outcome.

Indicate the Type of Documents Reviewed When Completing this Form

Indicate the type of documents reviewed when completing this form.

It is not necessary to search through the entire claim file for each of these documents. Simply check the types of documents that were reviewed, regardless if they were helpful in reconstructing the sequence of anesthesia care.

Page 7 of Form

Brief Summary of Events

Print or write a legible and concise summary that describes the anesthesia care that was provided and damaging patient outcomes that are related to the allegations in the claim or lawsuit.

Describe the sequence of specific incidents and actions taken, or not taken, that relate to the overall quality of anesthesia care, including any special circumstances.

Include information related to professional judgments made about the quality of anesthesia care while reviewing the claim or lawsuit. Include pertinent information that is not included elsewhere on the form.

Avoid the use of abbreviations, shorthand and colloquialisms that are not common knowledge among anesthesia providers.

Complete one (1) summary for obstetric cases. Even though reviewers complete a separate Claim File Review Form, one (1) for the mother and one (1) for the newborn, only one (1) summary, written on the form related to the mother, is necessary.

BRIEF SUMMARY OF EVENTS:

Please give a succinct narrative of the events. Specify the sequence of events, details not included elsewhere on the form, and details pertaining to the quality of anesthetic care:

This patient was a 64 year old Caucasian male, who was admitted for outpatient surgery for excision and biopsy of a L neck mass on 1/6/94.

The **general surgeon,** who performed the procedure, did not obtain the appropriate diagnostics that were needed preoperatively to confirm the diagnosis (according to reviewer knowledge of this potential diagnosis, i.e., carotid ultrasound, Doppler studies, etc.)

Intraoperatively, under general anesthesia, the neck mass was noted to be a carotid artery aneurysm. The surgeon opted to repair the aneurysm. The patient had only one peripheral IV line and no other invasive monitoring devices as the planned original procedure was to be done on an outpatient basis in an outpatient facility (i.e., neckbx) with a minimal potential for intraoperative

complications. The CRNA **did not** insert invasive monitoring devices, i.e., arterial line or larger peripheral intravenous lines after the surgeon decided to proceed with dissection of the aneurysm.

Under general anesthesia the CRNA allowed the patient's blood pressure to be maintained at approximately 40% below the preoperative systolic value. For a prolonged period of time, 30 minutes according to the record, the patient's blood pressure was 80/50. (The heart rate and rhythm intraoperatively were normal.) The CRNA did not inform the surgeon of the hypotension at any time during the course of surgery. The common carotid artery was clamped for an unknown period of time during the repair of the aneurysm. The anesthesia record was suboptimal at best, i.e., missing information. The anesthesia record indicated that the CRNA did nothing to alter the anesthetic to attempt to elevate the blood pressure that would be indicated for this procedure and with this patient.

In addition to the above, it should be noted, that the CRNA had a history of alcohol abuse in the past but the record reflects that the CRNA was sober at the time of this incident.

The patient failed to emerge from anesthesia at the case end. There was no significant blood loss noted. The patient expired two days postoperatively. From the time of anesthesia completion until expiration of the patient he never recovered consciousness. The cause of death was cerebral anoxia resulting in brain death.

The CRNA involved in this case was a solo practitioner.

D:\CLOSEDCL\INSTFRM.SAM-07/24/96-08/06/96-tr

APPENDIX D

Insurance Company Data Collection Form

Insurance Company Data

1. **Case Number:**_____

2. **State Filed #:**_____

3. **Date Claim Closed:**_____

4. **Does the Claim contain a Legible Anesthetic Record?**_____ yes_____no

5. **Was a Lawsuit Filed?** _____ yes _____ no

   If "yes" , indicate the disposition of the lawsuit:

   _____Dismissed by court

   _____Dismissed by settlement

   _____Court/Jury Trial

   _____Arbitration/Mediation

   If "no", indicate the disposition of the lawsuit:

   _____Total amount Paid

   _____Amount Paid for CRNA

   _____Expenses

   _____Information not available

6. **St. Paul Reviewer:** _____

7. **Date of Review:**_____

REFERENCES

Aaronson, L. S., & Burman, M. E. (1994). Focus on psychometrics use of health records in research: Reliability and validity issues. Research in Nursing and Health. (17), 67-73.

Abedi, J. & Baker, E. L. (1995). A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. Educational and Psychological Measurement, 55(5), 701-705.

Allison, J. J., Wall T. C., Spettell, C. M., Calhoun, J., Fargason, C. A., Kobylinski, R. W., Farmer, R., & Kiefe, C. (2000). The art and science of chart review. Journal on Quality Improvement, 26(3), 115-135.

Altman, D. G. (1997). Practical Statistics for Medical Research. Washington D.C.: Chapman & Hall/CRC.

American Association of Nurse Anesthetists (1992). Guidelines & Standards for NurseAnesthesia Practice. Park Ridge, Illinois.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83(5), 762-765.

Bates, D. W., O'Neil, A., Peterson, L. A., Lee, T. H., & Brennen, T. A. (1995). Evaluation of screening criteria for adverse events in medical patients. Medical Care, 33(5), 452-462.

Baer, D. (1977). Reviewer's comment: Just because its reliable doesn't mean that you can use it. Journal of Applied Behavior Analysis, 10, 117-119.

Bearman, J. E., Kleinman, H., Glyer, V. V., & LaCroix, O. M. (1964). A study of variability in tuberculin test reading. American Review of Respiratory Diseases, 90, 913-919.

Beecher, H.K., & Todd, D.P. (1954). A study of the deaths associated with anesthesia and surgery based on a study of 599,548 anesthetics in ten institutions 1948 – 1952 inclusive. Annals of Surgery, 140, 2 –35.

Birkelo, C. C., Chamberlain, w. E., Phelps, P. S., Schools, P. E., Zacks, D., & Yerushalmy, J. (1947). Tuberculosis case finding: Comparison of effectiveness of various roentgenographic and photofluorographic methods. Journal of the American Medical Association, 133, 359-366.

Bland, M. J., & Altman, D. G. (1986, February). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 6, 307-310.

Blok, H. (1985). Estimating the reliability, validity and invalidity of essay ratings. Journal of Educational Measurement, 22(1), 41-52.

Boba, A. (1965). Death in the Operating Room. Springfield, Illinois: Charles C Thomas.

Brennan, T. A., Leape, L. L., Laird, N. L., Liesi, H., Localio, R. J., 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. The New England Journal of Medicine, 34(6), 370-376.

Brennen, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. Journal of Educational Measurement, 14(3), 277-287.

Brennan, T. A., Localio, R. J., & Laird, N. L. (1989). Reliability and validity of judgments concerning events suffered by hospitalized patients. Medical Care, 27(12), 1148-1158.

Brook, R. H. (1977). Quality: Can we measure it? New England Journal of Medicine. 296, 170-172.

Brook, R. H. & Appel, F. A. (1973). Quality-of-care assessment: Choosing a method for peer review. New England Journal of Medicine, 288, 1323-1329.

Brown, D. L. (1992). Risk and Outcome in Anesthesia. Philadelphia: J. B. Lippincott Company.

Burdock, E. I., Fleiss, J. L., & Hardesty, A. S. (1963). A new view of inter-observer agreement. Personnel Psychology, 373-384.

Burt, C. (1955). Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 8, 103.

Butterworth, J. S. & Reppert, JE. H. (1960). Auscultatory acumen in the general medical population. Journal of the American Medical Association, 174, 32 – 34.

Caplan, R. A., & Posner, K. L. (1977, June). The expert witness: Insights from the closed claims file project. American Society of Anesthesiologists

Newsletter, 61(6), 9-10.

Caplan, R. A. (1992, March). Anesthetic liability: What it is and isn't (Supplement). Anesthesia and Analgesia, 19-24.

Caplan, R. A., Posner, K., Ward, R. J., & Cheney, F. W. (1988). Peer reviewer agreement for major anesthetic mishaps. Quality Review Bulletin, 14, 363-368.

Caplan, R. A., Posner, K., & Cheney, F. W. (1991). Effect of outcome on physician judgments of appropriateness of care. Journal of the American Medical Association, 265, 1957-1960.

Caplan, R. A., Ward, R. J., Posner, K., & Cheney, F.W. (1988). Unexpected cardiac arrest during spinal anesthesia: A closed claims analysis of predisposing factors. Anesthesiology, 68, 5-11.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and it's applications in educational measurement. Journal of Educational Measurement. 18(4), 183-204.

Cardinet, J., Tournneur, Y., & Allal, L. (1976). The symmetry of generalizability theory:Application to educational measurement. Journal of Educational Measurement. 13, 119-135.

Center for Disease Control and Prevention (National Center for Health Statistics). Births and Deaths: Preliminary data for 1998. National Vital Statistics Reports 47(25):6, 1999.

Cheney, F. W. (1999). The American Society of Anesthesiologists Closed Claims Project: What have we learned, how has it affected practice, and

how will it affect practice in the future? Anesthesiology, 91, 552-556.

Cheney, F.W., Posner, K., Caplan, R. A., & Ward, R. J. (1989). Standard of care and anesthesia liability. Journal of the American Medical Association, 261, 1599-1603.

Cicchetti, D. V. (1991). Peer review reliability. Behavioral and Brain Sciences, 14(1),119-186.

Cicchetti, D. V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. British Journal of Psychiatry, 129, 452-456.

Cicchetti, K. M., & Fleiss, J. S. (1977). Comparison of the null distributions of weighted kappa and the C ordinal statistic. Applied Psychological Measurement, 1(2), 195-201.

Cicchetti, D. V., & Heavens, R. (1984). A computer program for assessing the reliability of nominal scales using varying sets of Multiple raters. Educational and Psychological Measurement, 44, 671-675.

Cicchetti, D. V. & Sparrow, S. S. (1981). Developing criteria for establishing interrater reliability of specific items. Applications to assessment of adaptive behavior. American Journal of Mental Deficiency, 86, 127-137.

Clifton, B. S., & Hotten, W. I. T. (1963). Deaths associated with anesthesia. British Journal of Anesthesia, 35, 250-259.

Cochrane, A. L., Chapman, P. J. & Oldham, P. D. (1951). Observers errors in taking medical histories. Lancet, 1, 1007-1009.

Cochrane, A. L., & Garland, L. H. (1952). Observer error in the interpretation of

chest films. Lancet, 2, 505-509.

Cohen, J. A. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.

Cohen, J.A. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Collis, G. M. (1985). Kappa, measure of marginal symmetry and intraclass correlations. Educational and Psychological Measurement. 45, 55-62.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. Psychological Bulletin, 44, 322-328.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. Journal of Applied Psychology, 80, 565-579.

Cooper, J. B., Newbower, R. S., Long, C. D., & McPeek, B. (1978). Preventable anesthesia mishaps: A study of human factors. Anesthesiology, 49(6), 399-406.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of generalizability of scores and profiles. New York: John Wiley.

Cronbach, L. F. & Lindquist, E. F. (1953). Design and Analysis of Experiments in Psychology and Education. Boston: Houghton & Mufflin.

Dadakis, S., & Pozen, M. W. (1977). An interpretation of implicit judgments in chart review. Journal of Community Health, 2(4), 251-258.

Davies, M., & Fleiss, J. (1988). Measuring agreement for multinomial data. Biometrics, 38, 1042 – 1049.

Davis, L. L., & Grant, J. (1993). Focus on psychometrics guidelines for using psychometric consultants in nursing studies. Research in Nursing and Health, 16, 151-155.

Donabedian, A. (1988). The quality of care: How can it be assessed? The Journal of the American Medial Association, 260(12), 1743-1748.

Edwards, JG., Morton, H. J. V., Pask, E. A., & Wylie, W. D. (1956). Deaths associated with anesthesia: A report on 1,000 cases. Anesthesia, 11(3), 194-220.

Eichorn, J. H. (1991). Documenting improved anesthesia outcome. Journal of Clinical Anesthesia, 3, 351.

Eichhorn, J. H. (1989). Prevention of intraoperative anesthesia accidents and related severe injury through safety monitoring. Anesthesiology, 70, 572-577.

Engelhart, M. D. (1959). A method of estimating the reliability of ratings compared with certain methods of estimating the reliability of tests. Educational and Psychological Measurement, 19,(40), 579-588.

Erlich, O., & Borich, G. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. Journal of Educational Measurement, 16(1), 11-18.

Erlich, O. S., & Shavelson, R.J. (1978). The search for correlations between Measures of teacher behavior and student achievement: Measurement

problem, conceptualization problem, or both. Journal of Educational Measurement. 15(20), 77-89.

Evans, W. J., Cayten, C. G., & Green, P. A. (1981). Determining the generalizability of rating scales in clinical settings. Medical Care, 19, 1211-1219.

Everitt, B. S. (1968). Moments of the statistics kappa and weighted kappa. Journal of Mathmatical and Statistical Psychology, 21, 97 – 103.

Fessel, W. J., & Van Brunt, E. E. ((1972). Assessing quality of care from the Medical record. The New England Journal of Medicine, 286, 134-138.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 30, 71-76.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. Biometrics, 31, 651-659.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382.

Fleiss, J. L. (1965). Estimating the accuracy of dichotomous judgments. Psychometrika, 30, 469-479.

Fleiss, J. L., & Cicchetti, D. V. (1978). Inference about weighted kappa in the non-null case. Applied Psychological Measurement, 2(1), 113-117.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the Intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33, 613-619.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of

kappa and weighted kappa. Psychological Bulletin, 72, 323-327.

Fleiss, J. L., Nee, F. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin, 86, 974-977.

Fleiss, J. L., & Shrout, P. E., (1978). Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika, 43(2), 259-262.

Fletcher, C. M. (1964). The problem of observer variation in medical diagnosis with special reference to chest diseases. Method Information in Medicine, 3, 98-103.

Fletcher, C. M., & Oldham, P. D. (1964). Diagnosis in group research. chapter 2 of Medical Surveys and Clinical Trials, 2nd edition. London: Oxford University Press.

Gaba, D. M., Maxwell, M., & DeAnda, A. (1987). Anesthetic mishaps: Breaking the chain of accident evolution. Anesthesiology, 66, 670-676.

Garnick, D. W., Hendricks, A. M. & Broyen, A. (1991). Can practice guidelines reduce the number and costs of malpractice claims? Journal of the American Medical Association, 266, 2856-2860.

Goldman, R. L. (1992). The reliability of peer assessments of quality of care. Journal of the American Medical Association. 267, 958-960.

Goodenough, F. L. (1936). A critical note on the use of the term "reliability" in mental measurement. Journal of Educational Psychology, 27, 173-178.

Goodwin, L. D., & Prescott, P. A. (1981). Issues and approaches to estimating Interrater reliability in nursing research. Research in Nursing and Health,

4, 323-337.

Greenwood, J. M., & McNamara, W. J. (1967). Interrater reliability in situational tests. Journal of Applied Psychology, 51(2), 101-106.

Guilford, J. P., (1954). Psychometric Methods. New York: McGraw – Hill.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hartmann, D. (1977). Considerations in the choice of interobserver reliability estimates. Journal of Applied Behavior Analysis, 10, 103-116.

Hastings, G. E., Sonneborn, R., Lee, G. H., Vick, L., & Sasmor, L. (1980). Peer review check list: Reproducibility and validity of a method for evaluating the quality of ambulatory care. American Journal of Public Health, 70(3), 222-228.

Hawkins, R. P., Dotson, V.A. (1975). Reliability scores that delude: An alice in wonderland trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp and G. Semb (Eds.), Behavior Analysis: Areas of Research and Application. Engle wood Cliffs, New Jersey: Prentice-Hall, 539-376.

Hiatt, H. H., Barnes, B. A., Brennan T. A. (1989). A study of medical injury and medical malpractice: An overview. New England Journal of Medicine, 321, 480-484.

Hollander, M. & Wolfe, D. A. (1999). Nonparametric Statistical Methods. (2nd

ed.). New York: John Wiley & Sons.

Holley, J. A., & Guilford, J. P. (1964). A note on the G index of agreement. Educational and Psychological Measurement, 24, 749-753.

Horn, S. D., & Pozen, M. W. (1977). An interpretation of implicit judgments in chart review. Journal of Community Health, 2(4), 251-258.

House, A., Farber, J., & Nier, L. (1983). Differences in computational accuracy and speed of calculation between three measures of interobserver agreement. Child Study, 13, 195-205.

House, A., House, B., & Campbell, M. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. Journal of Behavioral Assessment, 3(1), 37-57.

Hubert, L. J. (1978). A general formula for the variance of Cohen's weighted kappa. Psychological Bulletin, 85(1), 183-184.

Hubert, L. F. (1977). Kappa revisited. Psycological Bulletin, 84(2), 289-297.

Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. Multivariate Behavioral Research, 14, 255-269.

Keats, A. S. (1990). Anesthesia mortality in perspective. Anesthesia and Analgesia, 71, 113-119.

Keats, A. S. (1988). Anesthesia mortality – a new mechanism (Editorial views). Anesthesiology, 68, 2-4.

Kraemer, H. C. (1980). Extension of the kappa coefficient. Biometrics, 36, 207-216.

Kratochwill, T. R., and Wetzel, R. J. (1977). Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. Journal of Applied Behavior Analysis, 10, 133-139.

Landis, J. R. (1975). A general methodology for the measurement of observer Agreement when the data are categorical. UMI Dissertation Services, 76-9260. (Xerox University Microfilms).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchial kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, 33, 363-374.

Leape, L. L. (1994). Error in Medicine. Journal of the American Medical Association, 272(23), 1851-1857.

Leape, L. L., Brennan, T. A., Laird, N., Lawthers, A. G., Localio, A. R., Barnes, B. A., Liesi, H., Newhouse, J. P., Weiler, P. C., & Hiatt, H. (1991). The nature of adverse events in hospitalized patients: Results of the Harvard Medical Practice Study II. The New England Journal of Medicine, 324, 377-384.

Lee, L. A., (2000). Postoperative visual loss data gathered and analyzed. American Society of Anesthesiologists Newsletter, 64, (9), 25-27.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 76(5), 365-377.

Loewenson, R. B., Bearman, J. E., & Resch, J. A. (1972). Reliability of

measurements for studies of cardiovascular atherosclerosis. Biometrics, 28, 557-569.

Lunz, M. E. & Stahl, J. A. (1994). Interjudge reliability and decision reproducibility. Educational and Psychological Measurement, 54,(4), 913-925.

MacKenzie, E. J., Shapiro, S., & Eastham, J. N. (1985). The abbreviated injury scale and injury severity score: Levels of inter- and intrarater reliability. Medical Care, 23(6), 823-835.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. British Journal of Psychiatry, 130, 79-83.

Maxwell, A. E., & Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. British Journal of Mathematical and Statistical Psychology. 21, 105-116.

Miller, R. D. (1994). Anesthesia. (4th ed.). New York: Churchill Livingstone.

Millman, J., & Glass, G. V. (1967). Rules of thumb for writing the ANOVA table. Journal of Educational Measurement. 4(2), 41-50.

Nobrega, F. T., Morrow, G. W., Smoldt, R. K., & Offord K. P. (1977). Quality assessment in hypertension: Analysis of process and outcome methods. New England Journal of Medicine, 296, 145-148.

Nunnally, J. C., Bernstein, I. H. (1978). Psychometric Theory. New York: McGraw-Hill Book Co.

O'Connell, D. L. & Dobson, A. J. (1984). General observer-agreement measures on individual subjects and groups of subjects. Biometrics, 40, 973-983.

O'Hara, M. W., & Rehm, L. P. (1983). Hamilton rating scale for depression: Reliability and validity of judgments of novice raters. Journal of Consulting and Clinical Psychology, 51, 318-319.

Orkin, F. K. (1989). Practice standards: The midas touch or the emperor's new clothes? Anesthesiology, 70, 567-571.

Orwin, R. G., & Corday, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. Psychological Bulletin, 97, 134-147.

Polit, D., & Hungler, B. (1983). Nursing research: Principles and methods. (2nd ed.). Philadelphia: J. B. Lippincott.

Posner, K. L., Caplan, R. A., & Cheney, F. W. (1996). Variation in expert opinion in medical malpractice review. Anesthesiology, 85, 1049-1054.

Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., & Cheney, F. W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. Statistics in Medicine, 9, 1103-1115.

Rae, G. (1984). On measuring agreement among several judges on the presence or absence of a trait. Educational and Psychological Measurement, 44, 247-255.

Revicki, D. A., (1984). The dependability of medical encounter diagnostic information. Medical Care, 22(7), 661-669.

Revicki, D. A., Klaucke, D. N., Brown, JR. E., & Caplan, R. A. (1990, November). Reliability of ratings of anesthesia's contribution to adverse

surgical outcomes. Quarterly Review Book, 404-409.

Richardson, F. D. (1972). Peer review of medical care. Medical Care, 10(1), 29-39.

Richardson, F. D., Trainor, P. E., Billinson, M. R., Singer, H., & Baehm, D. (1967, May). Rochester Region Perinatal Study. The New York State Journal of Medicine, Medical review project, Empire State Medical, Scientific and Educational Foundation, Inc. 1205-1209.

Rogot, E., & Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. Journal of Chronic Disease, 19, 991-1006.

Romm, F. J., & Puttnam, S. (1981). The validity of the medical record. Medical Care, 19(3), 310-315.

Rosenfeld, L. S. (1957). Quality of medical care in hospitals. American Journal of Public Health, 47, 856-865.

Rutstein, D. D. (1976). Measuring the quality of care: A clinical method. New England Journal of Medicine, 294, 582-588.

Scheffe, Henry (1959). The Analysis of Variance. (Wiley Classics Edition). New York: John Wiley & Sons.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory-1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428.

Silber, J. H., Williams, S. V., Krakauer, H., & Schwartz, J. S. (1992). Hospital and patient characteristics associated with death after surgery. Medical Care, 30(7), 615-629.

Smith, P. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Measurement, 3, 319 – 346.

Smyllie, H. C., Blendis, L. M. & Armitage, P. (1965). The observer disagreement in physical signs of the respiratory system. Lancet, 11, 412-413.

Soeken, K. S., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. Medical Care, 24, 733 –741.

Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. (1967). Quantification of Agreement in psychiatric diagnosis. Archives of General Psychiatry, 17, 83-87.

Thorndike, R. L (1920). A constant error in psychological ratings. Journal of Applied Psychology, 4, 25-29.

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34,273-386.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 22(4), 358-376.

Topf, M. (1986). Three estimates on interrater reliability for nominal data. Nursing Research, 35(4), 253-255.

Vanderkamp, L. J. (1976). Generalizability and educational measurement. advances in Psycological and Educational Measurement. New York: Wiley.

Wakefield, J. (1980). Relationship between two expressions of reliability: Percentage agreement and phi. Educational and Psychological Measurement, 40, 593-597.

Washington, C. C., & Moss, M. (1988). Pragmatic aspects of establishing Interrater reliability in research. Nursing Research, 37(3), 190-191.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Webb, N. M., Shavelson, R. J., Shea, J., & Morello, E. (1981). Generalizability of educational development ratings of jobs in the U.S. Journal of Applied Psychology, 66, 186-192.

Williams, G. W. (1976). Comparing the joint agreement of several raters with Another rater. Biometrics, 32, 619-627.

Yelton, A., Wildman, B., & Erickson, M. (1977). A probability-based formula for calculating interobserver agreement. Journal of Applied Behavior Analysis, 10, 127-131.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques. Public Health Report (Washington), 62, 1432-1449.

Yerushalmy, J. (1950). The role of dual reading in mass radiography. American

Review of Tuberculosis, 61, 443-464.

Yerushalmy, J., Garland, L. H., Harkness, J. T., & Zwerling, H. B. (1951). Evaluation of role of serial chest roentgenograms in estimating the progress of disease in patients with pulmonary tuberculosis. American Review of Tuberculosis, 64, 225-248.

Yule, G. U. (1900). On the association of attributes in statistics. Philosophical Transactions of the Royal Society, Series A, 194-257.

Yule, G. U. (1912). On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, 75, 579-642.

Abstract

MEASURING THE INTERRATER RELIABILITY OF A DATA COLLECTION INSTRUMENT DEVELOPED TO EVALUATE ANESTHETIC OUTCOMES

By

KAREN CRAWFORTH

December 2001

Advisor: Sholomo Sawilowsky, Ph.D.

Major: Evaluation and Research (Education)

Degree: Doctor of Philosophy

The American Association of Nurse Anesthetists Foundation in collaboration with the St. Paul Insurance Companies initiated the study of closed claim files in 1995. The intent of this research is to build a database that will provide a means to evaluate and improve patient safety. This goal will be accomplished through the ongoing identification of negative trends that contribute to anesthesia related adverse events. The knowledge gained through this review will be disseminated to the anesthesia community by means of educational programs and publications.

Subjectivity is an integral part of the process involved in extracting data from claim documents. Therefore, the establishment of interrater agreement is a necessary condition before confidence may be placed in the data. The purpose of this research was to measure the interrater reliability of the instrument developed to extract data from closed claim files. This project utilized the recommendations of related studies in the research design. These suggestions included the use of the original documents in the

analysis, the evaluation of multiple items, and the application of generalizability theory to the data.

The research evaluated the reliability of the eight closed claim team members. These individuals were selected based on criteria that guaranteed a clinically and educationally diverse research team. Twelve closed claim files were selected for the analysis. These files represented a variety of outcomes related to anesthesia practice. The instrument demonstrated an overall reliability of 0.73. The analysis of each item using the intra-class correlation coefficient, proportion of agreement, the kappa statistic, and generalizability theory provided a comprehensive evaluation of scoring patterns.

The results of this analysis produced specific recommendations that will be used to improve the reliability of the process. The subjective items contain ambiguous language and represent global concepts that cannot accurately be answered with a single question. The decomposition of these questions into specific items that focus on the process and not on the individual will improve the instrument's reliability. A qualitative analysis of the narrative summary may provide information not available with quantitative methods. Involving additional insurance carriers in the project will increase the number of claims available for review. The use of paired reviewers in subsequent interrater reliability studies will also increase the sample size. A recommendation for a future study is a collaborative effort with the American Society of Anesthesiologists based on the mutual goal of improving the quality of care. The data collection process involved in this research will require adjustments as the practice of anesthesia advances. The continued assurance that the information obtained by this study is reliable will be a necessary component to the dynamic process of closed claim data research.

AUTOBIOGRAPHICAL STATEMENT

KAREN CRAWFORTH

Karen Crawforth obtained her Bachelor of Science degree in Nursing from Michigan State University. She obtained her Master of Science degree in Anesthesia from Wayne State University. Ms. Crawforth is a member of the American Association of Nurse Anesthetists (AANA), the American Statistical Association and a member of the AANA Foundation Closed Claim Research Team. She is involved in continuing research and publications regarding anesthesia related adverse events, she is the Clinical Manager of the Anesthesia Department at Detroit Receiving Hospital, and is a contingent nurse anesthetist at Hutzel Hospital. Ms. Crawforth has an adjunct faculty position in the College of Pharmacy and Allied Health at Wayne State University and is actively involved in the didactic and clinical education of nurse anesthesia students.