McGill University

TRAUMAMETRICS

by
Ronald Lett
MD MSc. FRCSC

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE

IN

EPIDEMIOLOGY AND BIOSTATISTICS

DEPARTMENT OF EPIDEMIOLOGY AND BIOSTATISTICS
MCGILL UNIVERSITY
MONTREAL, QUEBEC

FALL 1992

ABSTRACT

In this thesis, measurement theory and the statistical evaluations of diagnostic test performance are applied to the measurement of injury severity. Relevant issues in traumatology, the rationale for measurement of injury severity and the importance of likelihood ratios (LRs) and Receiver operator characteristic (ROC) curve analyses are discussed. An assessment of the definition, scaling mechanism, reliability and validity of 22 severity instruments is organized into a reference guide. Data sources for this thesis include the literature, the Vancouver General Hospital trauma registry and the Pennsylvania State University trauma registry.

The LR and the area under ROC curves (A_{uc}) are calculated from the best published evaluations of four triage instruments. As shown in the table below instruments 1-3 which include anatomic information are superior to the RTS which contains only physiologic information.

	LR+	LR-	A_{uc}
1. CRAMS	44.4 (33.0 - 59.7)	.04 (.01 - .13)	.98 (.97 - .99)
2. PHI	9.8 (8.9 - 10.8)	0	.95 (.93 - .97)
3. RTI	7.4 (6.2 - 8.9)	.06 (.03 - .14)	.83 (.77 - .88)
4. RTS	3.8 (3.7 - 3.9)	.04 (.02 - .08)	.83 (.80 - .87)

A paired ROC curve analysis of the RTS and the RTI demonstrates that the performance of the RTI is equivalent to the RTS when mortality is the outcome; however the performance gain of the RTI over the RTS is 35% when major trauma is the outcome evaluated.

Explanations for differences in predictive validity are sought by evaluating scaling mechanisms, reliability and content validity.

ROC curve evaluation of ordinal vs interval scaling techniques, is performed by comparing the two formats of the RTS. No advantage for the interval scale is detected. The interrater reliability of the RTS is high (intraclass correlation = .95). Evaluation of content validity demonstrates that age, number of serious injuries and energy impact result in performance gains as high as 20% ($p=.0001$). Neither mechanism of injury, which only changes the cutoff, nor co-morbidity, after adjustment for age, increase A_{uc}. It is concluded that triage instruments with only physiologic content have limited predictive validity and that the widespread adoption of the RTS is premature. It is also concluded that standardized statistical techniques and the application of the principles of measurement theory are effective strategies for the refinement of injury severity instruments.

RESUMÉ

Dans cette thèse, les théories de la mesure et des méthodes d'évaluation des performances des épreuves diagnostiques sont appliquées à des indicateurs de sévérité des traumatismes. Différents aspects pertinents à la traumatologie ainsi que les raisons qui en motivent la mesure de la sévérité sont revus. L'importance du rapport de vraisemblance (RV) et de l'analyse des courbes caractéristiques de la performance d'un test (CPT) sont discutées. Nous proposons un guide de références portant sur vingt-deux instruments de mesure de la sévérité des traumatismes dans lequel on retrouve une définition de chaque instrument, une description du type d'échelle utilisé et des données sur leur fiabilité et validité.

Plus spécifiquement, la validité prédictive de quatre instruments de mesure utilisés à des fins de triage est évaluée à partir de leur RV et de leurs courbes CPT. Les valeurs utilisées aux fins de l'analyse représentent les meilleurs résultats publiés sur ces épreuves. Les écarts entre les différentes épreuves sont expliqués par l'utilisation de différentes échelles de mesure et par les différentes fiabilité et validité de contenu des instruments:

	RV+	RV-	A_{uc}
1. CRAMS	44.4 (33.0 - 59.7)	.04 (.01 - .13)	.98 (.97 - .99)
2. PHI	9.8 (8.9 - 10.8)	0	.95 (.93 - .97)
3. RTI	7.4 (6.2 - 8.9)	.06 (.03 - .14)	.83 (.77 - .88)
4. RTS	3.8 (3.7 - 3.9)	.04 (.02 - .08)	.83 (.80 - .87)

La capacité de prédire l'importance du traumatisme est évaluée par l'analyse des courbes CPT pour le RTS et le RTI à partir d'une même banque de données. Le RTI démontre sa supériorité sur le RTS.

Le RTS utilise soit une échelle ordinale ou une échelle à intervalles. L'analyse des courbes CPT ne démontre aucun avantage d'une échelle par rapport à l'autre. La fiabilité intra-observateur du RTS est élevée (coefficient de corrélation intra-classes de 0,95). L'évaluation de la validité de contenu démontre que l'âge, le nombre de blessures sérieuses et l'énergie d'impact contribuent à 20 % de la surface sous la courbe CPT (A_{uc}). Le mécanisme par lequel le traumatisme est infligé ne modifie que la valeur seuil alors que la coexistence d'une condition morbide, après ajustement en fonction de l'âge, n'a plus aucun effet sur A_{uc}.

Nous en concluons que les instruments de triage basés sur des données essentiellement physiologiques ont une faible validité de prédiction et que l'utilisation plus répandue du RTS est prématurée. De façon plus générale, il appert que l'application de principes de clinimétrie et de méthodes statistiques uniformes permet de raffiner les instruments de mesure de sévérité des traumatismes.

TO:

JUDY, ROBYN, RYAN, TARA

ACKNOWLEDGEMENTS

I would like to thank my supervisor James Hanley, PhD, Associate Professor of Epidemiology and Biostatistics, who was generous both with his suggestions and his time. Specifically I would like to thank him for writing the computer program using Delong's method which was essential for much of the data analysis.

Dr. David Mulder, Surgeon in Chief of the Montreal General Hospital and John Sampalis PhD, Epidemiologist, Department of Surgery, contributed to this thesis as members of my thesis committee and I am grateful for their participation.

The trauma registry data used in this thesis was provided by Dr. J. Vestrup, Director of Trauma Services at the Vancouver General Hospital and by Dr. J. S. Smith, Director of Trauma Services at the Milton S. Hershey Medical Center. The willingness of these two surgeons to share this data was essential for these investigations and I congratulate them on the high degree of academic openness that this action demonstrated.

Dr. Robert Marchand, who is interested in advancing the francophone cause, was kind enough to translate the abstract of this thesis. Marielle Olivier, the director of computer services in the Department of Epidemiology and Biostatistics was a willing consultant when computing problems arose. Finally I would also like to thank Vanessa Hill for the contribution of her word processing skills.

TABLE OF CONTENTS

Chapter One: TRAUMATOLOGY AND MEASUREMENT

Chapter Two: A TRAUMAMETRIC REFERENCE GUIDE

Chapter Three: STUDY OBJECTIVE AND RESEARCH DESIGN

Chapter Four: SPECIFIC METHODOLOGY AND RESULTS

Chapter Five: DISCUSSION

TABLES

Chapter Two

Chapter Three

Chapter Four

Chapter Four (Cont'd)

Chapter Four (Cont'd)

FIGURES

CHAPTER ONE

TRAUMATOLOGY AND MEASUREMENT

In North America, trauma ranks third after vascular diseases and malignant disorders, as a cause of death and is the leading cause of death in individuals between 1 and 45 years of age.[1,2] Regional trauma care systems have been introduced to reduce the mortality and morbidity associated with injury, by integrating pre-hospital, in-hospital and post-hospital phases of trauma care.[3,4] Instruments for the measurement of injury severity are used in the different phases of care to improve triage and clinical decisions and to facilitate both quality assurance programs and epidemiological studies.[3,5] The instruments developed for the measurement of trauma severity are integral to modern trauma care and are best assessed within the context of measurement theory.[6,7,8] The ability to determine outcome, predictive validity, is an instrument's most important quality and is the subject of this thesis.[3,5,9]

In this first chapter, issues of traumatology that will be reviewed are: trauma definitions, epidemiology, pathophysiology and management. The validity of injury severity instruments and the theoretical basis for this determination will be discussed in the section entitled measurement. The evaluation of diagnostic tests and the place that these evaluations have within measurement theory, will also be reviewed.

I. TRAUMATOLOGY

a) Definitions

According to Webster's dictionary, trauma is a bodily wound, or shock caused by an external agent;[10] in disciplines other than traumatology, injury is used as a general term but for the purposes of this thesis, injury will refer specifically to physical damage to a person, and will be used interchangeably with trauma.

Traumatology is the branch of surgery dealing with injury. The term accident, which implies that traumatic events are random and unavoidable, has been replaced in scientific work by the more precise term, unintentional injury.[11,12]

b) **Epidemiology**

While traumatic events may by unanticipated, epidemiological studies have demonstrated that the distribution of traumatic events is not random.[11,12] Injury rates are affected by demographic factors such as age, sex and race and are also influenced by temporal and geographic factors.[11,12,13]

The overall incidence of injury is unknown because many minor injuries do not come to medical attention; however it is known that trauma causes the death of 14,000 Canadians and 140,000 Americans annually. About one third of these deaths are due to motor vehicle related events, while homicide represents 15% of injury related mortality in the USA, in contrast to 4.3% in Canada.[2] The fact that for each fatal trauma case recorded, there are at least 3 cases of permanent disability, further emphasizes the magnitude of the burden of illness secondary to trauma.[1]

Injury related mortality rates vary greatly with both age and sex. High death rates are noted for young children, young adults and the elderly. The predominance of injury as a cause of mortality in the younger age categories explains why trauma leads all other disease processes as a cause of potential years of life lost;[2,13] however it must not be forgotten that injury is also important in the elderly. The North American injury related mortality rate of 39.3/100,000 for all ages climbs to 86/100,000 for those over 65 years and is almost 300/100,000 for those over 85 years.[14,15]

American studies of trauma related mortality determined that the unintentional death rate for males (67/100,000) is twice that of females (27/100,000). The intentional death rate for males is higher than that for females: the suicide rate is 19/100,000 for males and 6/100,000 for females and the homicide rate is 14/100,000 and 4/100,000 for males and females respectively.[13]

A study in Maryland found differences in the rate of trauma related mortality in different races. A high mortality from both unintentional injury (90/100,000) and suicide (13/100,000) was demonstrated in Native Americans . The suicide rate for whites was also high at 14/100,000 while the suicide rate determined for blacks and Asians was 6/100,000. The highest homicide rate at 35/100,000 was recorded in blacks.[13]

Trauma related mortality varies with economic, temporal and geographic factors. American data shows that the mortality rate varies from 71 per 100,000 in low income groups to 34 per 100,000 in high income groups. Injury-related deaths peak on Saturdays and in the month of July. Unintentional trauma deaths are twice as high in rural American areas as urban areas (75 vs. 37/100,000) while homicide is highest in large urban areas.[11,13]

The five leading causes of trauma requiring medical attention as reported in an emergency department based Ohio study were: falls (25.2%), cuts and piercing wounds (14.9%), being struck by or caught between objects (14.5%), motor vehicle crashes (12.1 %) and strenuous movements (8.5%).[15] Falls are the leading cause of non fatal injury and are second to motor vehicle crashes as a cause of fatality among all age categories, including the elderly.[15,16]

c) Pathophysiology

1) pathogen

Traumatic disorders result from the harmful transfer of environmental energy to the body. Positive energy sources include kinetic, chemical, thermal, electrical and radiation energy. Exposure to negative physical agents, results in altered energy transfer within the body and may result in trauma. Examples of negative physical agents include chemicals such as cyanide which blocks mitochondrial energy transfer and drowning which interferes with pulmonary gas exchange. [12,17]

2) conceptual model

The pathogenesis of injury is conceptualized within a three phase model, namely the pre-injury, injury and post-injury phases.[12]

The pre-injury phase involves the exposure of an individual to energy sources. As long as the performance level of the individual, which includes factors like skill, behaviour and physical tolerance exceeds that of the demands of a particular task, injury is avoided. Primary prevention of trauma occurs in the pre-injury phase.[12]

The injury phase occurs when the individual's performance level is exceeded by the demands of the task and an injury event takes place. This event may result from a decrease in the individuals performance level, an increase in the task demand, or a combination of decreased performance and increased task demand. The extent of the injury resulting from the descrepancy between the performance level and the task demand, depends on the magnitude of the energy load, the body surface area involved, the duration

of energy transfer, the impedance of intervening structures and the unique sensitivities of the tissues involved.[12]

The post injury phase is the time interval after the injury event when emergency, definitive and rehabilitative interventions occur. The prevention of morbidity and mortality by efficacious treatment in this phase is termed secondary prevention. [12]

3) pathogenesis[18]

The pathological components of injury are both local and systemic. The local component includes tissue destruction, blood loss, mechanical defects and superimposed infections. Injury is a dynamic process which continues to be inflicted until all of these local components have been corrected. The cumulative effects of the components of injury prolong both wound healing and the systemic effects of injury.

Systemically, trauma causes a marked increase in catabolism which alters its balance with anabolism. The resulting increase in energy expenditure is further increased when injury is complicated by infection. In major trauma, adequate nutritional energy must be provided to the patient to avoid the depletion of carbohydrate and lipid stores and to prevent the break down of structural protein.

There is a dynamic interplay between the systemic and local effects of trauma: the systemic alterations caused by local trauma affect wound healing while the persistence of the local injury results in continued systemic imbalance.

4) mortality

The pathological causes of injury related death have been studied in several autopsy series. In all of these series the leading cause of death is head injury.[19,20,21] In a series from San Francisco 50.1 % of deaths were due to brain injury, 29.1% were due to haemorrhage and vascular injury, and 9.8% of the deaths were due to sepsis. Fifty three percent of trauma fatalities occurred at the scene of the injury; 78% of the deaths occurring more than seven days after injury were due to sepsis.[19] The distribution of trauma deaths in relation to the time of injury is tri-modal, the immediate peak occurring in the first seconds to minutes, the early peak occurring in the first few hours and the late peak occurring days to weeks after injury.[1]

d) **Clinical and Diagnostic Strategies**

The clinical manifestations of trauma depend both on the specifics of the injury and the time interval between injury and definitive therapeutic intervention. An injury that occurs hours away from care will have a different presentation than the same injury that receives care minutes after the event.[7] Shires divides traumatized patients into three categories according to surgical requirements. The first category includes those individuals with an immediate threat to life due to interference with vital physiological function; this category of patient may require operative intervention within minutes of arrival in the emergency room. The second category includes patients with injuries that are not an immediate threat to life; these patients will require surgery but a few hours are available for clinical investigation. The third category includes patients with occult injuries where extensive investigations may be required to determine if intervention is required.[23]

The initial diagnostic strategy and subsequent treatment vary according to the threat to life and the organ systems that are injured. The threat to life varies according to the physiological system which is compromised. The American College of Surgeons Committee on Trauma has specified the priorities for diagnosis and immediate intervention in the trauma victim. The priorities are as follows:

A - airway
B - breathing
C - circulation
D - neurological deficit
E - exposure for complete patient evaluation.[24]

The management of these priorities which ideally occurs within the context of a trauma system may be performed by emergency medical technicians or medical doctors and may require interventions which range from simple first aid manoeuvres to sophisticated operative intervention.[4] Only after these priorities have been dealt with, is the diagnosis and management of other injuries appropriate.[24]

e) **Trauma Systems**

It has been accepted that organized systems are necessary for improvement of trauma care.[4,25] The support for this acceptance is based mainly on studies that compare trauma outcome before and after the introduction of trauma systems rather than on randomized controlled studies. For example, a study of general surgical trauma patients at Yale noted that the mortality rate of 16.1% for the year prior to the institution of a trauma service decreased to 11.8% in the year following the institution of a trauma service. The authors noted that no preventable deaths occurred during the year after the commencement of the trauma service, while at least 5 preventable deaths had occurred on the general surgical services in the previous year.[26] A San Diego study compared the survival of severely traumatized patients managed within a trauma system with

the probability of survival predicted by a national standard. The study showed that for penetrating trauma, the observed survival of 29% was superior to the expected survival of 18% and that for blunt trauma an observed survival of 20% was recorded when only 8% were expected to survive.[27] In an Ontario autopsy study an excess mortality of 47% was seen when comparing preventable deaths in patients managed without a trauma service to those in patients managed by a trauma service.[28] In spite of the methodological concern that factors other than a new trauma service could have caused these improvements, the evidence consistently supports the concept that regionalized trauma systems save lives.[4]

The American College of Surgeons committee on trauma defined four patient components to a trauma care system. These are access to care, pre-hospital care, hospital care and rehabilitation.[1]

Access to care requires that the trauma system be alerted when injury has occurred. Anticipatory measures to improve access to care include innovative communication technologies and public education. One suggested innovation was that off road vehicles and land vehicles used in remote areas carry electronic locating devices similar to those used in aircraft. [1,23,27]

Pre-hospital care involves personnel for initial resuscitation, treatment and triage of the trauma victim and also the equipment for extrication and transport of these patients.[1] The **hospital component** of a trauma system is organized with different hospitals being designated for the provision of differing levels of care. Level I centers provide tertiary care facilities, expect to treat 600-1000 trauma patients per year and are the systems center for trauma education, data collection and research. Level II facilities provide tertiary care with in-house surgery and anaesthesia capabilities, trauma quality assurance programs, and

treat 350 to 600 trauma patients per year but do not have research and education responsibilities. Other acute care facilities are classified as non-designated hospitals. Trauma systems have been designed for the urban USA but need to be adapted to the rural and Canadian situations, where distances for transfer of the trauma patient are often substantial. [3,29,30] The fourth component of a trauma system is **rehabilitation.** If permanent disability is to be avoided, this component should not be neglected.[1,4]

Trauma systems are very expensive and this fact has threatened their development. The economic returns of a trauma system were reviewed in a 1991 study from Tennessee. It showed that 89.5% of patients treated at a trauma center survive and 54.5 % of these patients return to an economically productive life. The authors of this study believe that the results of trauma systems justify their expense.[31]

II. MEASUREMENT

a) Rationale for the Measurement of Injury Severity

Measurement of injury severity is important for the pre-hospital triage of the trauma victim, as an aid to in-hospital clinical decision making, and is essential for valid administrative and scientific evaluations of all aspects of trauma care. Due to these varied requirements numerous instruments have been developed.[3,4,5,6] These instruments are formulated from different combinations of physiologic, anatomic, laboratory, demographic, and historic data.[8] Shires described categories for the trauma patient which define severity based on clinical requirements (see page 6).[23] The demanding task placed on injury severity instruments is that they identify the severely injured patient based on

information other than clinical outcome, to assure that correct decisions, that are critical to that outcome, are made.[3,22,25]

Early instruments of injury severity were developed for the field triage of trauma victims.[32,33] In the pre-hospital management of trauma, the decision concerning whether to transfer a patient to a trauma center or to a hospital not specifically designated for trauma, is not made by physicians, but by emergency medical technicians. Thus major trauma victims would bypass non-designated hospitals and be transported to a trauma center while minor trauma victims would be transferred to the nearest hospital, whether or not it was a designated trauma center.[4,6,22,34] As the time from injury event to death may be short, these instruments must facilitate rapid response. [1]

Decisions which are based on measurements of injury severity have important clinical and economic consequences. If an excess of minor trauma victims are transferred to a higher level trauma center, overtriaged, the resources of the trauma center will be overwhelmed with cases that could be well managed elsewhere. Overtriage could harm the non-designated hospitals, which would lose both experience and revenue as the result of the loss of patients. If major trauma victims are transferred to non designated hospitals, undertriaged, avoidable mortality may occur and the higher level of expertise of the trauma center will be under-utilized. Pre-hospital measurement instruments are designed to enable emergency medical technicians make appropriate decisions on the destination of the patient.[3,22,34]

Instruments used for in-hospital clinical decision making are of two types. One type is injury or organ specific; the other is a global measurement of trauma severity. The organ specific measures are used to aid in making specific clinical

decisions.[36,37] For example the CIS (colon injury score) is used to decide whether a primary anastomosis or a colostomy would be most appropriate for a patient with a colon injury.[36] Such organ specific instruments will not be evaluated in this thesis. The global in-hospital instruments attempt to quantify the patient's illness or track the dynamics of the disease process and its response to treatment. These instrument are not usually designed specifically for trauma but are designed for the assessment of severity in critical illnesses, which includes trauma.[3]

Evaluation of patient outcome is important for scientific evaluations of the care of the trauma victim. If injury prevalence and severity are not taken into account, evaluations of outcome in either preventative or therapeutic interventions can not be compared and may be biased. Injury severity instruments are used to adjust for differences in disease severity so that evaluations of different populations may compare "like with like".[38] In addition to the scientific implications, patient outcome is an important factor in fund allocation and public policy decisions. Trauma registries and the analysis of these data bases have become an essential and required component of trauma care. Trauma indices are central to outcome evaluations, quality assurance programs and hospital reimbursement.[3,5,6]

b) **Measurement Theory**

Despite the importance of injury severity instrumentation, the evidence concerning the validity of these instruments is often conflicting and confusing. This is particulary true of the triage instruments. This confusion is due to the lack of uniformity of statistical techniques, the frequent use of prevalence dependent measures of performance and varying outcome criteria.[39,40,41] Articles written by the proponents of specific injury severity instruments claim that their

instrument is effective and recommend its general adoption, while some authors claim that the present instruments are not helpful.[42,43,44,45] The analyses in these studies are not comparable making it difficult to judge the relative merits of the different instruments. The following section of this chapter will discuss the specifics of measurement theory and the evaluation of diagnostic tests; the next chapter will apply these principles to the published information on injury severity instruments.

1) Psychophysics and psychometrics [46,47]

Until the last decade there has been little interest in the theoretical basis for measurement in the health sciences. However as clinical research has become more complex, and instruments for the measurement of complex or abstract concepts have become necessary, measurement theory has become relevant. Because measurement in the basic medical sciences and in classical epidemiology presented no inherent difficulty, a theoretical basis for the management of methodological issues did not develop in medicine. It has therefore been necessary to borrow the theories of measurement developed in other disciplines, and apply them to the health sciences. The disciplines of psychology and education, which deal with subjective concepts, developed psychophysic and psychometric methodologies which are the theoretical bases for the measurement of abstract concepts.[47]

Pyschophysic methodology, which developed prior to psychometric methodology, demonstrated that humans can make consistent numerical estimates of the magnitude of physical stimuli, like the brightness of light or the loudness of sound. Psychophysic methodology also demonstrated that individuals can make consistent comparison of abstract phenomena, for

example, in the comparison of the brightness of a light to the loudness of a sound. [46,47]

Psychophysics studies the subjective measurement of stimuli which have a physical scale; psychometrics, an adaptation of psychophysical methods, assigns numerical estimates to qualities that do not have a physical scale. The psychometric methodology which includes evaluations of reliability and validity is now being applied to health measurement, including illness severity, where it has been shown that the consistency in judgments demonstrated in psychological measurement also applies to the rating of health and illness. [47]

2) Clinimetrics and Traumametrics

The application of measurement theory to clinical problems, in general, has been termed clinimetrics and in this thesis the specific application of measurement theory to the instruments of injury severity will be termed traumametrics.[48]

Clinimetric scales have been used in medicine for many years and recently quantitative methods have been increasingly applied to therapeutic problems. These indices are used to classify disease for prognostic purposes, to standardize therapeutic decisions and to increase objectivity when clinical signs were not sufficiently reliable. In addition to individual patient management problems, illness severity instruments are becoming increasingly important in evaluating outcome in specific categories of patients and also for public policy problems.[48,49,50,51,52]

Earlier clinical instruments were disease based and quite simple. More recently, disease based instruments have become more complex and emphasis has also been placed on the global measurement of illness severity. Instruments used in critical illness have evolved from the subjective to the more objective, and from a focus on a single system to multisystem measurements. Cullen found that single system severity instruments were unsatisfactory in the intensive care unit, because of their failure to account for individuals with multi-system disease.[52] The complexities involved in the measurement of multi-system failure, have necessitated the application of measurement theory to clinical problems.

Essential theoretical issues in the evaluation of illness severity instruments which are reviewed below include instrument definition, scaling techniques, reliability and validity.[3,5,7,8,9,53,54]

INSTRUMENT DEFINITION[7,8,53]

An instrument is defined by its attributes, means of application and purpose.

The first issue in instrument definition, which depends on the instrument's purpose, concerns the attributes or items utilized in the instrument.[8] No single item or attribute is usually sufficient to discriminate between alternative outcomes; hence to measure concepts like illness severity it is necessary to construct a composite index.[54]

The items in an instrument may be chosen on an empirical or a theoretical basis. Empirical techniques are used when an instrument has a practical purpose

and a recordable outcome which can be predicted. The theoretical approach is to choose items that appear relevant to a specific theory of the concept. Empirical scales only describe a condition, whereas theoretically based scales can both describe and explain a condition. For example an empirical assessment of lung function might evaluate respiratory rate while a theoretical assessment might evaluate arterial oxygen tension; the former is based on the observation that respiratory rate is often altered with dysfunction while the latter is based on the concept that the lung oxygenates blood. [46,47]

Attributes chosen for use in a severity instrument are usually chosen empirically and include demographic, historical, physiological and anatomic items. Only instruments which have physiological items can reflect the dynamics of a patient's changing condition. Selection of items for severity scales is often determined on the basis of data availability or instrument simplicity. Krischer states that age, sex and prior medical history should be included in the attribute profile of severity instruments; others believe that demographic detail should be excluded from an instrument's attribute profile but included in outcome analyses.[8]

The second issue in instrument definition, the means of application, includes the setting for instrument use, and the skill level of personnel using the instrument. In injury severity instruments the setting could be the site of the injury event, the interior of an ambulance or the medical records department and the personnel could be ambulance attendants, health care professionals or medical records technicians.[8]

The third issue in instrument definition is its purpose, which is simply the task the designer hopes that an instrument can perform.[8]

SCALING TECHNIQUES[46,47,53]

Analytical issues involved in instrument assessment include the type of scale utilized and the arithmetic functions used to determine scale aggregates. Scales are categorized according to a mathematical hierarchy. The four levels in increasing order of complexity are nominal, ordinal, interval and ratio scales. [47]

The lowest level, the nominal or categorical scale is simply a labelling process. Even if the label is numerical, no inference can be made from the relative size of the numbers involved. An example of a nominal scale would be gender. Nominal scales are not appropriate for measuring illness severity.

In ordinal scales, which are the most common type of scale used, numbers or labels reflect the increasing magnitude of the characteristic being measured. For example a patient could be mildly, moderately or severely injured or be graded from 1 to 3. Strictly speaking adding or subtracting ordinal scales is not appropriate nor is the use of parametric statistical techniques. Scales must meet at least the requirement of ordinality to be useful as illness severity instruments.

In interval scales, the next highest order, the units between numerical scores are equal and therefore can be added, subtracted and analyzed using parametric statistical techniques. Interval scales are mathematically limited because it is not possible to make multiplicative interpretations. For example, in an interval scale which rated injury from 1 to 5, the increase in severity from 2 to 3 would equal the increase from 4 to 5. Interval scales are preferable to ordinal scales in the measurement of illness severity but success in their development has been limited.

Ratio scales, which are the highest order, have the properties of interval scales but also have a natural zero value which makes multiplicative interpretations possible. In a ratio scale which rated injury from 0 to 5, not only would the increase in severity from 2 to 3 equal the increase from 4 to 5 but the injury rating of 4 would be twice as severe as an injury rating of 2.

Considerable debate has surrounded the common practice of assuming interval properties for scales which are ordinal. Some authors note that unless score distributions are severely skewed, one can analyze ordinal scales as if they were interval scales.[47] Another approach has been the development of methods which are used to develop equal appearing intervals. One such technique, Thurstone's method, determines the median rank for numerous scale items and then selects a limited number of items in a manner which results in equal intervals. [47]

Many of the severity indices are instances of statistical functions called additive value functions, a few are multiplicative value functions and others use vector analysis. The statistical problem is to provide a summation figure from the profile of attributes of an individual that is representative of health status. The grouping of individuals with differing profiles but identical summary scores is useful, as individuals with the same severity profile are often too few for statistical analysis.[8,53]

MEASUREMENT QUALITY

The quality of a measurement is assessed by evaluating two characteristics, reliability and validity. Reliability characterizes the consistency of an instrument

while validity evaluates the meaning of a measurement. Several ways have been developed to evaluate these characteristics. [46,47]

Reliability

The reliability or consistency of an instrument is the degree to which a measurement can be replicated. Absence of reliability reduces the clinical usefulness of an instrument and reduces the statistical power of both randomized and observational studies which use such an instrument. Low reliability in the measurement of a confounding variable can bias a study's conclusions. The upper bound of an instrument's validity is limited by its reliability. [46,47,55]

Gibson notes that severity instruments should comprise numerical ratings with clear and objective decision-rules for score derivation to enable the same rater, on different occasions, or different raters on the same occasion, to arrive at an "identical score" for the same patient.[7]

Two traditions exist for the assessment of instrument reliability. These are the psychometric tradition which was developed in the context of questionnaires, and the analysis of variance (anova) tradition which was developed within the context of reproducibility of the rater's results.[47] Test-retest reliability, parallel forms reliability and internal consistency are part of the psychometric tradition; inter-rater reliability is based on the anova tradition. Cronbach's alpha which is used for determination of internal consistency and the Intra Class Correlation which is used for evaluation of inter and intra-rater reliability can be shown to be quantitatively equivalent. An instrument which evaluates a subject by means of a series of items is conceptually equivalent to the summation of the observations made by different raters on the same subject. This statistical

equivalence implies theoretical equivalence of the different evaluations of reliability.[56]

The means of evaluating reproducibility, inter-rater and intra-rater reliability are the assessments that are most important in the evaluation of severity instruments. Inter-rater reliability compares the scores obtained by one rater with the scores obtained by another rater. The intra-rater reliability compares the score obtained by the same rater on separate occasions. The standard evaluations of inter-rater and intra-rater reliability are estimates of the intraclass correlation coefficient (ICC) for continuous measures and Cohen's Kappa (K) for categorical measures. Evaluations that are often used are simple agreement rates and Pearson correlation coefficients. An interesting alternative is graphic presentation of individual observations.[47,57]

The ICC, which ranges from 0 to 1, defines the portion of the measurement statistic which is explained by the trauma victim's condition. The remainder of the statistic is due to bias introduced by the raters, and random error.[47]

The K statistic evaluates observer agreement after correcting for chance agreement. Some authors believe that the K statistic unfairly penalizes raters who agree on the frequency of a condition.[58] The K statistic when weighted quadratically for partial agreement yields identical results to the intraclass correlation.[47]

Inter-observer and intra-observer agreement rates are reported in many clinical studies as a means of assessing reliability.[59] Agreement regarding normality of a patient is usually higher than agreement concerning abnormality.[60] Thus the agreement rate can be affected by the proportion of normal individuals in a

study population. Simple agreement rates have been criticized because they do not account for chance results.[47,59]

The Pearson product moment correlation coefficient is often used as means of reliability assessment. This is not appropriate as it evaluates the linear association between variables rather than the agreement level. Correlation may be high when agreement is nil.[57]

An alternative approach for reliability evaluation which is recommended by Bland and Altman is a graphic approach where the difference between the rater's score for each subject is plotted against the mean score for that subject. This allows a visual interpretation of reliability and the comparison of an instrument reliability at different levels.[57]

Gibson states that intra- and inter -rater reliability are the minimal evaluations needed in evaluating illness severity instruments and that evidence for this reliability should be presented by determination of either the intra class correlation coefficient or the kappa statistic.[7]

The acceptable level of reliability is not defined. Studies with a large sample size may tolerate an unreliable instrument; however instruments used for clinical decisions on individuals require a high level of reliability.[47]

An instrument's low validity may be secondary to its low reliability. Therefore a knowledge of an instrument's reliability can give insight into its validity. Severity instruments are often used in clinical decisions which indicates that the reliability requirements of these instruments are high.

Validity

Validity is the degree of confidence that can be attributed to inferences made from measurement results.[61] Psychometric theory has traditionally divided the concept of validity into three types: content validity, criterion validity and construct validity.[62] These divisions although somewhat artificial, are useful for validity assessment.

Content validation judges whether all the relevant domains of a trait are contained in the instrument. It is an evaluation of the information content of an instrument and the relevance of that information. Items on a scale must discriminate on the basis of the trait of interest and all domains of interest should be represented while domains from other traits should be excluded. The assessment of content validity is based on expert opinion and a literature review.[47] In this thesis the information evaluated for its validity includes domains represented by items in at least one instrument used for the measurement of injury severity. The content validity of these items will be evaluated by determining the increase in the performance of instruments which contain this information.

Face validity, a closely related concept, is a judgment of whether the items in an instrument appears reasonable to those who use the instrument.[47] In this thesis, this qualitative evaluation will be based on information in the literature and the author's opinion.

Construct validation uses underlying theory to develop or improve an instrument. A construct is a theory which is used to explain the relationship between various behaviours, measurements or outcomes. One technique of construct validation determines whether an instrument can differentiate between

the extremes of a particular trait. This is not useful in the development of a severity instruments because it is in the discrimination of the intermediate forms of an illness that an instrument would be most helpful. Correlation of an instrument with a trait that is felt to be related to the construct is convergent validity; correlation to a trait that is felt to be unrelated to the instrument is discriminant validity. If the construct is valid a correlation coefficient would be high in convergent validity and low in discriminant validity.[47]

Criterion validity, the most important quality of severity instruments, is demonstrated by comparing the instrument of interest to a criterion which is considered to be a gold standard. Criterion validity is called concurrent validity if another instrument is the gold standard and predictive validity if an outcome is the gold standard. An example of concurrent validation would be the comparison of contrast venography with radioactive scanning in the diagnosis of phlebitis. An example of predictive validity would be the comparison of the results predicted on a chest x-ray with the results of lung biopsies.[47]

Unfortunately there is no instrument which can be considered a gold standard for most illnesses, including trauma. Therefore the evaluation of predictive validity is the means of assessing criterion validity for injury severity instruments.

Predictive validity is often determined within data sets therefore the quality of the data is important. Validation should occur using data sets dissimilar from those used to develop the instrument to ensure that conclusions can be generalized beyond the original data set.[7] The validity of a severity measure that is designed for one outcome being applied to a different outcome is open to question. Often severity instrument evaluations fail to provide estimates of

standard error, making it impossible to determine the statistical significance of observed differences.[8]

Psychometric evaluations of criterion validity usually use correlation coefficients like Pearson's for interval data or phi for categorical data. In clinimetric evaluations of predictive validity the methodology commonly used is based on the statistical evaluations of diagnostic tests.[47,48] This methodology is an important part of this thesis and is reviewed in the subsequent section of this chapter.

STATISTICAL EVALUATION OF DIAGNOSTIC TEST PERFORMANCE

Severity instruments are developed and validated within data systems which predict clinical outcome(s). These processes require a data base that contains relevant clinical and laboratory indicators and the actual outcome for each patient.[54] Evaluation of the predictive validity of injury severity instruments can be performed with statistical methodology used in the evaluation of diagnostic test performance.[47,63] In the evaluation of severity instruments, the measurement of interest is compared with patient outcome. In traumatology, a consensus on which patient outcome constitutes the gold standard, has not been reached. Mortality, immediate surgery, admission to intensive care, disability, and the injury severity score have all been used either alone or in different combinations as the patient outcome.[22,34,39,42,43,64,65]

Three methods of the evaluation of diagnostic tests will be presented: the derivatives of the decision matrix (sensitivity, specificity and likelihood ratios), ROC curve analysis, and information theory.

A. Decision Matrix[63,66]

The decision matrix is a 2X2 table which presents binary diagnostic measures with binary outcomes, in a format which facilitates the analysis of the four cells. The combinations in the cells are true positive (a), false positive (b), false negative (c) and true negative (d).

Figure 1.1: DECISION MATRIX

	Outcome +	Outcome -
TEST +	a	b
TEST -	c	d
TOTAL	a + c	b + d

A decision matrix or a series of decision matrices can be developed for continuous data if these data are reorganized into a binary format; if only a single matrix is developed information is lost. Sensitivity and specificity are usually determined from a single decision matrix. Sensitivity, the true positive fraction, evaluates the test performance only among subjects who have a positive outcome a/(a+c). Specificity, the true negative fraction only evaluates the test performance in subjects who have a negative outcome d/(b+d).

Several authors find the likelihood ratios, derivatives of the decision matrix, superior to the individual measures of sensitivity and specificity because they allow easier interpretation of diagnostic tests. [22,66,67] A likelihood ratio is an expression of the change in the risk that a disease process is present for a given level of a diagnostic test. The likelihood ratio for a positive test (LR+), is the ratio of the true positive fraction to the false positive fraction (sensitivity/{1-specificity}). The likelihood ratio for a negative test is the ratio of the false

negative fraction to the true negative fraction ({1-sensitivity} /specificity). A high LR+ indicates that the likelihood of disease in a patient is increased while a low LR- indicates that the likelihood of the presence of disease is decreased. Unlike specificity and sensitivity which are determined from subsets of subjects with different outcomes (see figure 1.1), likelihood ratios are calculated from all subjects.

Using a nomogram the post-test probability of disease in a patient can be readily calculated from the likelihood ratio and the pre-test probability of disease.[66,68] The likelihood ratio is more applicable in the clinical situation because it indicates the change in the patient's likelihood of disease as opposed to the sensitivity and specificity which only give information concerning test performance.

Sensitivity and specificity, which only have meaning when reported as a pair, are cumbersome values for comparative purposes; difficult to interpret when applied to a patient; and make post-test probability calculations fairly complex. The likelihood ratios facilitate instrument comparison, are transparent in application to an individual patient and can be easily used to determine post-test probability of outcomes.[65,67]

In the trauma literature the term accuracy indicates the proportion of the study population with correct results and inaccuracy the proportion of cases with incorrect results.[69] These types of assessments have been termed "naive" because they depend more on the prevalence of disease within the population than the efficacy of the instrument.[70,

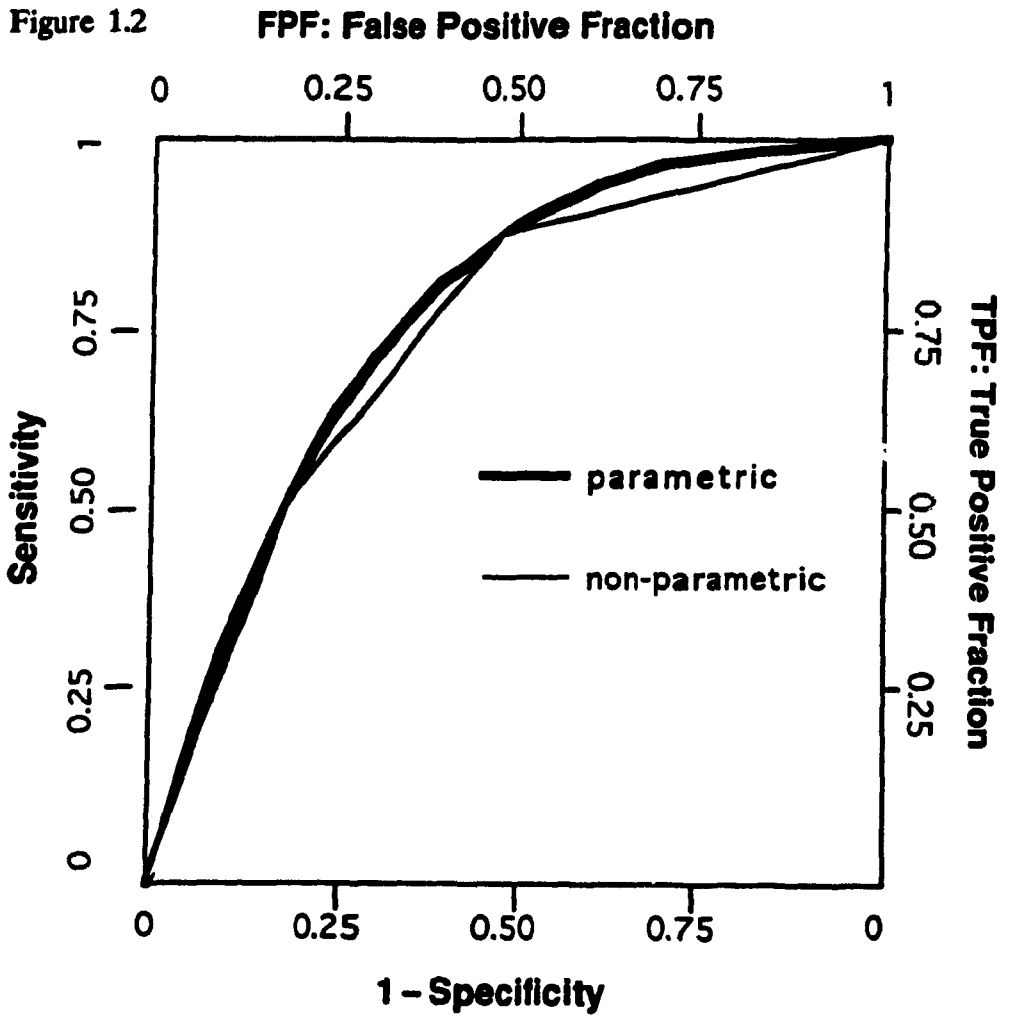
Because naive accuracy measures are poor descriptors of the predictive validity of an instrument, results based on them will not be reported in the review included in the next chapter. Sensitivity and specificity, the most frequent measures reported will be presented as will likelihood ratios when available. It is noted that the trauma literature rarely reports confidence intervals for the components of the decision matrix and the one study that did report likelihood ratios calculated the confidence intervals incorrectly.[22,71,72]

B. The Receiver Operator Characteristic (ROC) Curve [63,66,73,74,75]

The components of the decision matrix and their derivatives, the likelihood ratios, do not depend on the prevalence of disease in the study population, but do have several drawbacks. They lose information when measurement simplification is required and do not account for either the raters confidence threshold or decision criterion (cutoffs). ROC curve analyses, like the components of the decision matrix are independent of disease prevalence but unlike them, can evaluate ordinal and continuous instruments and are independent of both decision criteria and observer confidence threshold.

The ROC curve which demonstrates the continuous tradeoffs between proportions of true positives and false positives can be conceptualized as an infinite series of decision matrices. The curve is constructed by plotting true-positive ratios (sensitivities) against false-positive ratios (1-specificity). (See figure 1.2.)

Figure 1.2


FPF: False Positive Fraction
0
0.25
0.50
0.75
1
Sensitivity
TPF: True Positive Fraction
parametric
non-parametric
1 – Specificity

The area under the ROC curve which evaluates the location of the entire curve rather than any particular decision criteria or cutoff on the curve estimates an instrument's performance. The area ranges from 0.50 (50%) which is a chance result to 1.0 (100%) which indicates perfect discrimination.

The area can be calculated using parametric assumptions as is done in the computer program ROCFIT, or by using non-parametric assumptions within standard statistical programs.[76,77,78] The parametric method of area determination which fits a smooth curve gives a slightly larger estimate than the non-parametric estimate which uses a trapezoidal technique to calculate the area.[76,79](see figure 1.2). When the area under the curve is estimated using parametric assumptions it is referred to as A_z. In this thesis the term A_{uc} will be used as a generic term for the estimate of the area under the curve.

Standard errors for an estimated A_{uc} can be determined and 95% confidence limits calculated so it is possible to compare different instruments using different or identical data sets. When instruments are compared within the same data set, i.e. on the same patients, the paired nature of the data reduces the variance, therefore a downward correction of the standard error is appropriate and allows a more powerful comparison.[77] The statistical power of ROC curve analysis to detect differences between instruments can also be maximized if cases of intermediate difficulty are evaluated. Poor power to detect differences occurs if cases are very easy, because A_{uc} for all instruments will approach 1 and statistical power will also be a problem if the cases are very difficult because all instruments will then approach the null, which for A_{uc} is 0.50.[76]

ROC analysis, sensitivity, specificity and likelihood ratios are relatively stable to changes in the prevalence of the outcome in different data bases unless there

is differential selection. This bias occurs if outcome detection is more likely in those with a positive test than those with a negative result.[80] This is an unlikely problem in trauma research.

An ROC curve evaluates continuous or ordinal measures without information loss, and is not biased by subjective evaluation or by disease prevalence. Therefore the performance of an instrument, as measured by A_{uc} is unaffected by the choice of a cut off or the observer's confidence threshold and is stable to changes in the prevalence of the outcome in different data bases.

C. Information Theory

In the context of information theory, information is defined as a reduction in uncertainty. The greater the difference between the certainty of a diagnosis before a measurement is taken, compared with after, the greater the information content or gain of a test.[63]

In the trauma studies originating from Washington Hospital Center in Washington DC, the authors calculate information gain for trauma instruments by comparing the probability of survival of trauma patients who have been assessed by the instrument, to the overall prevalence of mortality.[81] They use a methodology which they call PER (prevalence, expected, relative) which estimates the relative information gain, by calculating the ratio of the information gain for the instrument of interest to the perfect instrument.(This methodology is equivalent to the K statistic.[82,83]) This format accounts for the population prevalence of the outcome of interest. The ratio rather than the difference was used because if absolute information gain was evaluated an outcome prevalence of 0.50 would have a large potential information gain while

an outcome prevalence close to 1.0 would have a small potential information gain, irrespective of the efficacy of the instrument.

In this thesis ROC curve analyses of injury severity instruments will take place within the context of information theory. The absolute performance gain is defined below.

As the null for A_{uc} is .50 an absolute difference between the A_{uc} determined for two instruments underestimates the difference in performance by one half. The performance gain of one instrument over another, and the 95% Confidence Interval of the performance difference is equal to twice the difference in A_{uc}.

Absolute Performance Gain = $2 (A_{uc}2 - A_{uc}1)$
95% Confidence Interval = APG ± 2(1.96*SE absolute difference)

SUMMARY

Trauma is an important public health problem. Our understanding of the epidemiology and biology of trauma is reflected in the attributes included in the injury severity instruments. Important clinical, scientific and administrative decisions are made on the basis of these instruments. Illness severity instruments have a well established place in clinical medicine but it is only recently that they have been evaluated within a theoretical framework. Evaluation of injury severity instruments, which are essential to trauma care should occur within a theoretical framework, to permit an understanding of their performance, and how that performance may be enhanced.

CHAPTER TWO

A TRAUMAMETRIC REFERENCE GUIDE

Instruments for measuring trauma severity are being used to assist those making life and death decisions for the trauma victim, accreditation decisions for hospitals, economic decisions for health care systems and scientific conclusions in clinical and epidemiologic research. The purpose of this literature review is to document the traumametric qualities of severity instruments which were either designed specifically to measure injury severity or were designed to measure critical illness, and which have been applied to the measurement of injury severity. Injury severity instruments have flaws, but rather than dismiss the currently available instruments lightly, a critical evaluation may lead to strategies for their improvement and refinement.[3,5,6]

Using the published information available on each instrument, the definition, scaling techniques, reliability and validity will be assessed. The measurement qualities of the pre-hospital, hospital and outcome instruments will be reviewed in separate sections. Each section will include: 1. an overview of instrument nomenclature and the measurement qualities evaluated, 2. a review of the measurement qualities of individual instruments and 3. a general commentary.

The nomenclature of injury severity instruments is confusing due to use of acronyms and the similarity of their names. The overview in each section will clarify the nomenclature and record which evaluations have been performed. The review of the individual instruments critiques the evaluations of the instrument qualities and makes inferences concerning the instruments. This comprehensive review of information on individual instruments has been compiled for reference purposes. **For the convenience of the reader the instruments which are referred to later in the thesis will be indicated with an asterisk.** The last part of each section will discuss general

issues concerning the present understanding of the quality of pre-hospital, hospital or outcome instruments.

I. PRE-HOSPITAL INSTRUMENTS

A) Overview

The pre-hospital instruments were designed to help paramedical personnel identify major trauma victims and triage them to trauma centers. The scaling techniques used in these instruments include ordinal and interval scales. As is shown in Table 2.1 most of the scales are ordinal.

Table 2.1: PRE-HOSPITAL INSTRUMENTS
Nomenclature and Scaling Mechanism

	Instrument Name	Scale
1. RTI	Revised Trauma Index	Ordinal
2. TI	Triage Index	Interval
3. TS	Trauma Score	Ordinal
4. RTS	Revised Trauma Score	Interval
5. CRAMS	Circulation, Respiration, Abdomen, Motor, Speech	Ordinal
6. PHI	Pre-hospital Index	Ordinal
7. PTS	Pediatric Trauma Score	Ordinal
8. Kane's	Kane's Checklist	Ordinal*
9. MOI	Mechanism of Injury	Ordinal*
10. AI	Anatomic Injury	Ordinal*

*Dichotomous Scale

The reliability of most pre-hospital instruments is untested and instructions for the attainment of reliability with these instrument are uniformly lacking. The reliability of the TS and CRAMS was evaluated using overall agreement not corrected for chance. The TI and RTS had the goodness of fit of a logistic regression model reported as an evaluation of reliability. This is not an assessment of inter-rater agreement or reliablity as understood within the context of measurement theory. No assessment of reliability of the other instruments are reported in the literature.

Table 2.2 shows the information content for and the type of predictive validity assessments which have been performed on each instrument. The components of the decision matrix were the usual means of evaluating predictive validity, although likelihood ratios and ROC curve analysis were occasionally used.

Table 2.2: PRE-HOSPITAL INSTRUMENTS
Content and Predictive Validity Assessments

	INFORMATION CONTENT					PREDICTIVE VALIDITY			
	Phys	Anat	Age	Mech	Morb	Matrix	LR	ROC	Infor
1. RTI	*	*		*		*			
2. TI	*					*			*
3. TS	*					*	*	*	*
4. RTS	*					*		*	*
5. CRAMS	*	*				*	*	*	
6. PHI	*	*				*	*	*	
7. PTS	*	*				*			
8. Kanes		*	*	*		*			
9. MOI & AI		*	*	*		*			

Column Headings: Phys - Physiologic, anat - Anatomic, Mech - mechanism of injury, Morb - co-morbidity, Matrix - decision matrix, LR - Likelihood ratio, ROC - receiver operator characteristic curve, Infor - information theory.

B) Review of Individual Pre-hospital Instruments

More than a dozen instruments for the pre-hospital measurement of injury severity are presented in the literature.[3,5,84] In this review ten of these instruments will be presented in two categories, scales and checklists. The scales are ordinal or higher order instruments, while the checklists are dichotomous ordinal instruments which define major trauma as a positive response to any single item included on the checklist. The scales which are prominent in the literature and therefore will be reviewed are the Trauma Index, Champion's Trauma Indices, the CRAMS Score, the Pre-hospital Index and the Paediatric Trauma Score. Following this three checklists: Kane's checklist, the Mechanism of Injury and Anatomic Injury will be discussed. As noted earlier, instruments essential for further understanding of the thesis will be indicated with an asterisk.

SCALES

*1) *The Revised Trauma Index*[32,39,85]

The trauma index which was first introduced in 1971 was evaluated in a revised form in a 1974 Japanese study but was not further studied until 1990 when it was evaluated in the revised form at Pennsylvania State University in Hershey. The date and reason for the revision are not recorded in the literature. The favourable conclusion concerning the RTI reached in this recent study, requires its re-evaluation.[32,39,84]

a) Definition

i) Purpose and Application - The original trauma index was designed to be used as a triage instrument both by paramedical personnel in the field and by physicians in community hospital emergency departments. The revised trauma index (RTI) has only been evaluated for use as a triage instrument by paramedics.

ii) Attributes - The RTI consists of five attributes, two anatomic items (body region involved and wound type) and three physiologic items (central nervous, cardiovascular and respiratory status).

b) Scaling Techniques

The RTI is an approximate ordinal scale with individual item scores that do not increase sequentially. These items are coded 1,3,5, or 6 with increasing magnitude indicating more severe injury. The only difference between the Trauma index and the RTI is the item coding. The highest score for each attribute is used to compute the total score. The RTI categorizes patient injury severity as minor (3-9), moderate (10-14), severe 15-19, and critical (>20).

The instruments developers report a linear relationship between the aggregate of the trauma index and their impression of severity suggesting ordinality. The five groups of variables are assumed to be on a continuum but are not always from a single underlying attribute. For example the items for the evaluation of respiratory status include apnoea and chest pain which can not be considered to be on a continuum.[7] The five components of the trauma index have been arbitrarily weighted and are aggregated as a simple arithmetic sum.

c) Reliability

No evidence is presented on inter or intra-rater reliability. Gibson notes that the statement that the variables included in the Trauma Index are easily understood by the non-physician does not mean it can be reliably applied.[7]

d) Validity

i) Content Validity - Both the trauma index and the RTI include physiological, anatomic and mechanism of injury items. They do not include demographic or comorbidity items. The range of items included in these instruments is wider than other indices thereby giving them a high content validity.

ii) Predictive Validity - The original study used hospital admission as the outcome of interest. The data set that was used to validate the instrument was generated from the same institution as the data set to develop the instrument. Also there was no assurance given that the outcome in the validation study (admission) was blinded to the trauma index total.[7,32]

A second study of this index performed in Japan in the pre-hospital setting assessed the validity of the trauma index with verbal rather than numerical values. The study concluded that the trauma index and the status of the patient one week after injury, were correlated.[85]

In the 1990 Pennsylvania study, the RTI was evaluated using two outcomes, death and the Injury Severity Score (ISS - discussed below in section III). In prediction of death the RTI had a sensitivity of 95% and a specificity of 87%. When an ISS dichotomized at 15 and then 20 was used to define major trauma, the sensitivity of the RTI was 48.3 % with a specificity of

92.3% for the former and the sensitivity was 69.7% with a specificity of 90.8% for the latter definition.[39]

e) Commentary

The trauma index was one of the earliest indexes in civilian practice. The items included in this instrument were selected from 60 variables. Only one publication objectively evaluates the predictive validity of this instrument. An assessment of the reliability of the RTI has not been published. On the basis of the most recent evaluation of the RTI, the authors of the study recommend its use as a triage instrument.

*2 ***Champion's Trauma Indices (Triage Index, Trauma Score, Revised Trauma Score)***

The above instruments all originated from the Department of Surgery, Washington Hospital Center and are adaptations of each other. The Triage Index, the first of these instruments, was rapidly replaced by the trauma score (TS); the latter has been thoroughly evaluated and as a result of these evaluations it in turn was replaced by the revised trauma score (RTS). The triage index and the RTS were developed with two aggregation techniques one for use in patient triage, the other for outcome evaluation.

2a) ***Triage Index and Triage Score***[81]

a) Definition

i) Purpose and Application - The Triage Index was developed to assist paramedical personnel identify life threatening states, to permit the

evaluation of the therapeutic performance of a trauma service, and to compare this performance with other trauma services.

This instrument was designed for field use by paramedical personnel with investigative modalities limited to physical examination, vital signs and an electrocardiogram rhythm strip.

ii) Attributes - The five attributes included in the triage index are respiratory expansion, capillary refill, eye opening, verbal response and motor response. The last three attributes are components of the Glasgow Coma Scale. These attributes were chosen because they correlate with survival and they represent the three systems (central nervous system, respiratory system and cardiovascular system) whose failure is most often the cause of injury related mortality.

b) Scaling techniques

The five attributes are coded as follows: respiratory expansion 0,2 or 3, capillary refill 0 or 2, eye opening from 0 to 3, and both verbal response and motor response each scored from 0 to 4. The triage score which is meant for triage purposes is simply the arithmetic sum of the above items. The Triage Index is determined by weighting the five attributes, according to coefficients produced in a logistic regression evaluation of survival. The values in the triage scale are ordinal but the weighting techniques used in the triage index result in an approximation of an interval scale.

c) Reliability

Individual attributes of the triage index were evaluated for inter-rater reliability. A comparison of the individual attributes which were obtained from the same

charts by different observers showed a disagreement level of 6%. A definition for disagreement was not presented, and no mention was made of a correction for chance agreement. The reliability of this instrument when data are collected directly from the patients has not been evaluated.

d) Validity

i) Content Validity - The items chosen for inclusion were chosen from the three physiological systems which are most commonly associated with injury mortality, therefore suggesting high face validity. The absence of co-morbidity, mechanism of injury, demographic or anatomic attributes restricts the trauma index's content validity.

ii) Predictive Validity - The triage index was validated against a second data set, generated by the same institution. The outcome evaluated was mortality. The result was considered a false positive if the predicted probability of death was over 50% and the patient survived, and false negative if the patient died with a probability of death of greater than 50%. Using this cutoff the sensitivity for prediction of death was 96% and the specificity was 80%.

Information gain was evaluated in these patients using PER.(P is the prior probability of death, E is the information gain and R is the relative information gain). The R value which compares the instrument to a perfect instrument was .83 in the design set and .78 in the validation set. The PER technique is mathematically the same as the Kappa statistic.[54] No studies on the triage index have been published since the original article.

iii) Commentary

The triage index is important as the basis for the development of the other Washington scales. In the development of this scale the PER techniques were used to evaluate the instrument and logistic regression coefficients were used to develop an interval scale. The reliability of this instrument has not been satisfactorily evaluated. Inter-rater agreement was determined only on chart based data and the statistical evaluation did not adjust for agreement by chance. The sensitivity and specificity of the index were reasonably high but no studies on populations from other institutions have been published.

*2b) ***Trauma Score***

The trauma score is a modification of the Triage index and is its successor.[86,87]

a) Definition

i) Application and Purpose - The trauma score is designed to be used in the pre-hospital setting by paramedical personnel. It has been shown to be useful in both blunt and penetrating trauma,[63] as well as in paediatric age groups.[65,69,86,87] The purposes described by the designers include patient triage, patient outcome prediction and standardizing disease prevalence in comparative studies. This instrument has been used with the injury severity score (ISS) to evaluate patient outcome as a part of the TRISS methodology.[88]

ii) Attributes[70] - The attributes used in the TS, as in the Triage index are from the neurologic, cardiovascular and respiratory systems. The neurologic attributes which were similar to the Glasgow Coma Scale (GCS) in the

triage index were replaced by the corresponding GCS items in the Trauma Score. The cardiovascular items were expanded to include systolic blood pressure and capillary return and the respiratory items to include respiratory rate and respiratory expansion.

b) Scaling techniques[86]

The scale has a range from 1 to 16 with 16 indicating the minimum derangement. The neurologic attributes are adapted directly from the GCS but are coded from 1 to 5. Respirations per minute are coded from 0 to 4, respiratory expansion is coded 1 or 0, systolic blood pressure from 0 to 4 and capillary return from 0 to 2. The final score is a simple arithmetic sum of the attributes with no attempt at development of an interval scale as was done with the Triage Index.

c) Reliability[87]

In a Washington DC study[66], reliability of the attributes of the trauma score was assessed in the typical pre-hospital setting, by assessing inter-observer agreement levels. An experienced nurse researcher's results were compared to those of paramedics with differing levels of training. The nurse-researcher aggregated the components to obtain the trauma score. In the 16 point scale there was complete agreement in 91%, a disparity of 1 point on 4.9% and of 2 points in 0.7%. No adjustment for chance agreement was performed. The distribution of the Trauma Score was presented and it was noted that the results were highly polarized with 111/207 patients at 16, which is the least injured category and 31/207 being recorded at the lowest value of 1 which is the most injured category.[89]

A San Francisco study[70] compared the TS obtained by the paramedic in the field to the score obtained by a senior surgical resident in the emergency department.[86] The Kendall's rank order correlation coefficient was 0.55 and the uncorrected agreement level, defined as a TS deviation of 1 or less was reported to be 84.9%. The authors reported that patients with a TS disparity of more than 2 on the second reading had a poorer prognosis than those with no disparity. The authors suggest that some of the disagreement was due to the responsiveness of the TS to the dynamics of the patient's condition but the fact that a patient with an improved trauma score also had a worse than average prognosis makes this finding difficult to interpret.

d) Validity

i) Content Validity - Like the Triage Index the face validity of the trauma score is high, as the attributes included in the instrument are from the physiological systems associated with injury related mortality. Like the triage index, only physiologic items are included, so the content of the index is not comprehensive.

ii) Predictive Validity - Many of the studies of pre-hospital instruments use the Injury Severity Score (see ISS section III), as a measure of outcome. In a San Francisco study, using the ISS of greater than 20 as an outcome the sensitivity of the TS was 43.3% at a cutoff of 12 and 63.3 at a TS cutoff of 14. The specificities at these two cutoffs were 96.9% and 88.4% respectively.[87] Numerous studies have confirmed the low sensitivity and high specificity of the TS.[22,34,45]

e) Commentary

Of the pre-hospital injury severity instruments, the trauma score has been the most thoroughly evaluated. The TS is an ordinal scale. The evaluations of reliability have not been rigorous but do suggest that TS results are reproducible. The low sensitivity of the TS results in many false negatives but few false positives. The low sensitivity is a problem with this instrument as many authorities believe that it is unacceptable to "undertriage" the severely injured individual. One study reported that false negative patients were physiologically stable and could safely have their true status determined in a non-designated community hospital prior to transfer to a trauma center.[75,88]

*2c) ***Revised Trauma Score***[43]

The trauma score was revised because two items in the TS were difficult to observe at night and because the Major Trauma Outcome Study showed that the trauma score was underestimating the severity of injury in some patients with head trauma.

a) Definition

i) Purpose and Application - The revised trauma score has the same purpose and application as its predecessors which were designed to triage trauma victims, predict patient outcome and standardize trauma prevalence and severity for comparative studies.

ii) Attributes - The neurologic attributes are the GCS items, the respiratory attribute is respiratory rate and the cardiovascular attribute is systolic blood pressure. Respiratory expansion and capillary refill which are in the TS

were excluded from this instrument because neither could be observed at night and respiratory expansion was difficult to observe at all times.

b) Scaling Techniques

In the RTS the intervals for the Glasgow coma scale used in the TS, were re-coded according to probability of death, which the authors state are accepted by neurosurgeons. The blood pressure and respiratory rates were then assigned similar numerical codes which corresponded to the probability of death. Two formats of the Revised Trauma Score (RTS) are described. The triage-RTS (t-RTS), the arithmetic sum of the three attributes, is an ordinal scale for pre-hospital triage, and the RTS, the sum of the weighted coded values of the three attributes, is an interval scale. Like the triage index the weights were determined by a logistic regression survival evaluation within a data set defined for the design of the RTS.

c) Reliability

The proponents of the RTS note that its "predictive reliability' is superior to the TS. "Predictive reliability" is a measure of the goodness of fit of the models generated by the two instruments using logistic regression survival evaluation. The assessment of goodness of fit, using the Hosmer-Lemeshow statistic was much better for the RTS than the TS. This suggests increased precision of the model generated using RTS, but it is not a measure of agreement. No studies of inter-rater reliability or other measures of agreement have been published concerning the RTS.[38]

d) Validity

i) Content Validity - The information content of this instrument, like the TI and the TS, is entirely physiological; none of the other domains of interest are included.

ii) Predictive Validity - The predictive validity is presented in terms of both ISS and death as outcomes. The data concerning prediction of ISS include only sensitivity which was 94.8% for the TS as compared to 97.2% for the RTS. Sensitivity is meaningless without an estimate of specificity. Again using mortality as an outcome, the authors compared the sensitivity of the T-RTS and TS and found the T-RTS more sensitive(59% vs 48%) but less specific (81% vs 91%). than the TS when using mortality as an outcome.[43]

e) Commentary

The revision of the trauma score was motivated by the finding that some head-injured patients with high probability of survival, were dying and in response to complaints that some attributes were difficult to visualize in the pre-hospital situation. This adaptation does not seem to have reduced the diagnostic accuracy of the instrument and has the advantage of easier use and better modelling properties, than the TS. The improved goodness of fit may be due to the fact that the trauma score was never weighted within a regression model as were the triage index and the RTS. The inter-rater reliability of this instrument has not been evaluated.[89]

*3. ***CRAMS Scale***[42,90]

a) Definition

i) Application and Purpose - CRAMS was developed to triage major trauma victims to a trauma center and other traumatized patients to a non-designated facility. The instrument was designed to be used by paramedical personnel.

ii) Attributes - The CRAMS Scale includes four physiologic items (capillary refill, respiration, motor and speech) and one anatomic item(abdomen/chest injury).

b) Scaling Techniques

The five items are scored from 0 to 2 so that a score of 10 results when all items are normal. CRAMS is an ordinal scale which is aggregated as an arithmetic sum. The weighting of the items is arbitrary.

c) Reliability

In a prospective 1985 study in Utah the level of the CRAMS scale determined in the field was compared with the level determined in the emergency room when the patient arrived. The comparison showed that 89.5% of the totals were identical, 9% were within one point of each other and 1.55% differed by 2 to 4 points.The agreement between field and emergency room determination of major versus minor trauma using the Crams scale was 99.4%.[42]

The proportion of normal records (CRAMS = 10) was 0.65; the proportion of minor trauma (CRAMS $\geq$ 7) was 0.95. The inter-observer agreement was high; however no adjustment for chance agreement was performed.

d) Validity

i) Content Validity - The physiologic variables are similar to the variables in the Champion Indices which give it equivalent face validity. One anatomic item, truncal injury, adds to CRAMS content validity.

ii) Predictive Validity - In the Utah study a CRAMS Scale of less than six was associated with a mortality of 62%, while those with a score of 7 or greater had a mortality of 0.15%. Ninety-eight percent of those with a score or 6 or less either had immediate surgery, were transferred to the ICU or died in the emergency room. The American College of Surgeons states that it is expected that about 5% of trauma cases will be major trauma. The CRAMS Scale identified 5.5% of cases in the Utah study as major trauma. In a study from San Diego at a cutoff of 8, CRAMS had a sensitivity of 66% and a specificity of 82% when predicting mortality [44]

e) Commentary

CRAMS is an ordinal scale which has been evaluated for both its reliability and validity. It may perform as well as other triage instruments but has been criticised because it lacks the interval properties thought to be useful in outcome evaluation and because the required examination of the chest and abdomen may make it more difficult to use.[43,92]

*4) ***PHI (Pre-hospital Index)***[64,93]

a) Definition

i) Purpose and Application - This instrument was designed for use by paramedical personnel. Its purpose is to define an objective and accurate basis for triage decisions in the pre-hospital phase of trauma care.

ii) Attributes - The PHI has five components, four of them physiologic (Systolic blood pressure, pulse rate, respiratory status and level of consciousness) and one anatomic (the presence or absence of penetrating trauma to the thorax or abdomen).

b) Scaling Techniques

The four physiologic items are coded from 0 to 5; however not all the numbers in the sequence are used. Penetrating trauma is scored as either 0 or 4. The weighting of items is arbitrary.

The simple arithmetic sum provides the summary score. The authors classify patients with a score of 3 or less as "minor trauma" and those with a score of greater than 3 as "major trauma".

c) Reliability

The reliability of this instrument has not been evaluated.

d) Validity

i) Content Validity - This scale uses 4 physiologic items which evaluate the three most important systems, central nervous system, respiratory and cardiovascular systems. In addition to these physiologic items, one anatomic factor, the presence or absence of penetrating trauma increases the comprehensiveness of the instrument.

ii) Predictive Validity - In the original pilot study on the PHI, the mortality was 0% for those measured at 3 or less and 27% for those with a measure of 4 to 24. In a subsequent prospective multi-center study no mortality was noted in individuals with scores of three or less but a mortality of 23% was recorded in those with a score higher than 4. Using 3 as a cutoff point and

death or emergency surgery as an indicator of major trauma, the PHI had a sensitivity of 92.7% and a specificity of 93.3%.

e) Commentary - This instrument had high specificity and sensitivity in a multi-center study. Like CRAMS it was not designed for outcome evaluation. The reliability of this instrument has not been evaluated.

5) ***PTS (Paediatric Trauma Score)***[94,95]

a) Definition

i) Purpose and Application - This system was developed to assure comprehensive evaluation of the injured child in the pre-hospital situation and for the field triage of the paediatric patient. The PTS was evaluated on children up to 18 years in the original study and children up to 14 years in a subsequent study.

ii) Attributes - The attributes include three anatomic items, (weight, open wound, fracture), and three physiologic items (systolic blood pressure, airway maintenance, and central nervous system status).

b) Scaling Techniques

Attributes are scored -1, +1 and +2; the weighting of these attributes is arbitrary. The score is a simple arithmetic sum of the scores on the six attributes. The instrument is an approximation of an ordinal scale.

c) Reliability

This property has not been assessed in the PTS.

d) Validity

i) Content Validity - This instrument contains anatomic and physiologic data but no demographic, mechanism of injury or co-morbidity items. The use of body weight, and blood pressure measurements that are relevant to assessing trauma in children, makes sense and therefore increases the face validity of the PTS.

ii) Construct Validity - The PTS was compared with the ISS in two populations one from a regional paediatric trauma center, the other from a national paediatric trauma registry. The mean ISS for patients with a PTS <6 was 30 while the ISS for patients with a PTS >6 was 6. This study indicated that the PTS identifies major trauma similarly to the ISS (convergent validity).

iii) Predictive Validity - One study recorded that a PTS of <2 was associated with a 100% mortality and that a PTS of <6 was associated with an "increased potential for morbidity and mortality". A subsequent study using the ISS >15, as an outcome, determined that at the optimal cut point of 8 the sensitivity and specificity of the PTS were 0.78 and 0.75 respectively.

e) Commentary

The recognition that children are different than adults and that physiological variables must be adjusted to the size of the child is an important contribution. The definition of instrument attributes and the scaling techniques used are arbitrary and imprecise. The predictive validity has been evaluated and is similar to the RTS after the coding in the RTS was adjusted to vital signs appropriate for children.

CHECKLISTS

*6) *(MOI) Mechanism of Injury, and *7) (AI) Anatomic Injury*[40,41,91]

a) Index Definition

i) Purpose and application - MOI and AI are used alone or in combination with pre-hospital scales to assist paramedical personnel make triage decisions.

ii) Attributes - The attributes included in MOI lists are falls of > 15 feet, motor vehicle crashes with extrication time over 20 minutes, intrusion into the passenger space, ejection , death of other passengers, pedestrian thrown more than 15 feet, pedestrian under 12 years, or field evidence of high energy transfer.

Attributes included in the Anatomic Injury checklist include penetrating trauma from head to mid-thigh, major amputation, severe burn, scalping injury, paralysis, or head and neck trauma with airway obstruction.[41]

b) Scaling Technique

These are dichotomous ordinal scales with any positive item resulting in the patient being labelled as a major trauma victim.

c) Reliability

These checklists have not had a reliability evaluation.

d) Validity

The studies evaluating MOI and AI do not report sensitivity, specificity, or likelihood ratios. The evaluations report the correct classification proportion a quantity which is dependent on the prevalence of major trauma in the study population. A study from Portland Oregon showed an appropriate triage of 36% with overtriage as high as 43%.[88] The authors felt this was acceptable as the TS without MOI information would have missed as much as 36% of the major trauma.[40]

e) Commentary

The inadequacy of the pre-hospital scales has resulted in the use of MOI and AI data to help improve the sensitivity of physiological instruments. This is done at the expense of specificity and whether the MOI and AI improve the performance of triage scales is not clear.[91]

8) ***Kane's Checklist***[22,32]

a) Definition

i) Purpose and Application - The instrument was designed by Kane to assist paramedical personnel in the field identification of patients who would benefit from a trauma center.

ii) Attributes - Six anatomic, two physiologic, two mechanism of injury and one demographic item(s) are included in this list.

b) Scaling Techniques

This instrument is a checklist which is a dichotomous scale. Any positive item results in the patient being labelled as a major trauma victim and triaged to a trauma center.

c) Reliability

Kane's checklist has not been assessed for reliability.

d) Validity

i) Content Validity - The checklist included anatomic and mechanism of injury items which are not included in many of the scales. Several of these items like penetrating head injury were added to the checklist because the authors felt they had high "face validity". The checklists in contrast to the scales include age as item within the instrument.

ii) Predictive Validity - When major trauma was the outcome of interest and was defined as an ISS < 16, major truncal surgery, or death within 6 hours of emergency department admission, the checklist had a sensitivity of 81% and a specificity of 77%. These results were generated from the same data set which was used to develop the instrument. In a smaller study, Kane's checklist was shown to have a sensitivity of 85% and a specificity of 65% when major trauma was used as the outcome.[22]

e) Commentary

In the first published study of Kane's Checklist, it performed as well as the ordinal scales. This checklist performed well in a small study using a different data set. The reliability of this instrument has not been evaluated.

C) General Commentary on the Pre-hospital Instruments

Although considerable data on the predictive validity of different pre-hospital instruments are presented in the literature, the use of different cutoffs and outcomes make interpretation and comparison of these data difficult. Confidence intervals and other standard statistical techniques are rarely reported.

Only two studies have compared the efficacy of pre-hospital severity instruments using ROC analyses in the same data set.[44,45] Because of the paired nature of these evaluations there is no possibility of bias due to patient selection. In Nebraska, Oranato compared CRAMS and TS, which were reconstructed retrospectively from information obtained from a computerized record of ambulance charts.[45] They evaluated the instruments' ability to identify major trauma victims which were defined as patients who had acute life or limb threatening injuries, were dead on arrival or required Cardio-pulmonary resuscitation. CRAMS and TS were compared on 5130 (27% of the total) patients who had enough information to calculate the totals for the two scales. No demographic differences were noted between the trauma victims excluded from the study or those included in the study. In this study the TS had a sensitivity of 0.37 and CRAMS had a sensitivity of 0.20. The specificities were 0.77 and 0.87 respectively. ROC curves were calculated for both of these instruments and compared to the impression of the ambulance attendants. The performances of both instruments were inferior to the ambulance attendants impression of injury severity. This study has shown performances which are lower than any other published evaluations of these two instruments. It is not known whether the performance of the instruments was poor, or whether this reflects the poor performance of the retrospective reconstruction of the instruments.

In a prospective study from San Diego, Baxt et al. compared CRAMS, RTS, TS, and PHI against several endpoints using ROC curve analysis.[44] With death as an endpoint all the instruments had a sensitivity of at least 85% with a specificity of at least 85%. However when other endpoints, which are presumably less obvious were used the performance of all of the instruments deteriorated. When the ISS was used as an endpoint, no instrument reached a sensitivity and specificity of 70% and when patient deficit was used as an outcome sensitivities and specificities of 65% were not achieved. The authors concluded that adequate information to predict outcomes that were not obvious, would not be available in the pre-hospital situation.

These comparative studies of pre-hospital instruments suggest that the instruments are not satisfactory at the present time. The San Diego study, which has none of the methodological problems of the Nebraska study questions the value of the three most widely used triage instruments. These findings require further scrutiny.

II. HOSPITAL SEVERITY INSTRUMENTS

A) Overview

Hospital Severity Instruments have been designed as an aid to clinical decision making and to assure comparability in observational studies. The scaling mechanism of the in-hospital critical care instruments is usually ordinal. CARE uses a mechanism which give a pictorial assessment of severity and another which plots the trajectory of patient status.

Table 2.3: HOSPITAL INSTRUMENTS
Nomenclature and Scaling Mechanism

	Instrument Name	Scale
1. APACHE	Acute Physiology and Chronic Health Evaluation	Ordinal
2. APACHE II	Acute Physiology and Chronic Health Evaluation Two	Interval
3. SAPS	Simplified Acute Physiology Score	Ordinal
4. CARE	Cardiorespiratory Physiologic Studies	Undefined

APACHE II has had its inter-rater reliability assessed. No other reliability studies have been performed on the other instruments.

The content and predictive validity of the hospital instruments are presented in Table 2.4. Unlike pre-hospital and outcome instruments, these instruments have had their predictive validity evaluated with ROC curve analysis and inferential statistics. The use of these instruments with trauma patients in particular, has been limited.

Table 2.4: HOSPITAL INSTRUMENTS
Content and Predictive Validity Assessments

	INFORMATION CONTENT					PREDICTIVE VALIDITY		
	Phys	Anat	Age	Mech	Morb	Matrix	LR	ROC
1. APACHE	*				*	*	*	*
2. APACHE II	*		*		*	*	*	*
3. SAPS	*						*	*
4. CARE	*							

Column Headings: Phys - Physiologic, anat - Anatomic, Mech - mechanism of injury, Morb - co-morbidity, Matrix - decision matrix, LR - Likelihood ratio, ROC - receiver operator characteristic curve

Only a few instruments have been specifically designed for use with trauma patients after admission to hospital. Several instruments have been developed for the evaluation of the critically ill patient in the intensive care unit, and some of these have been applied to trauma patients. Three such instruments will be reviewed along with one instrument which was developed specifically for the critically ill trauma patient. Instruments mentioned later in the thesis are indicated with an asterisk.

B) Review of Individual In-Hospital Instruments

*1 *APACHE*[96,97]

a) Definition

APACHE stands for acute physiology and chronic health evaluation.

i) Purpose and Application - This instrument was designed to determine the severity of illness in groups of critically ill patients. The authors note that randomized studies are unacceptable in many clinical situations, and that APACHE could improve the internal validity of observational studies.

ii) Attributes - The APACHE system consists of two parts, a physiology score to evaluate the acute illness and an evaluation of pre-existing health. The acute illness portion consists of 34 physiological variables from 7 physiologic systems: cardiovascular, respiratory, renal, gastrointestinal, haematological, metabolic and neurologic system. Pre-existing health is evaluated on the basis of the patients activities of daily living, physician attendance, medication use and past medical history.

b) Scaling Techniques

The physiologic score is based on the sum of the weights given to the 34 variables. The weights vary for 0 to 4. Chronic health is classified from A to D. Although chronic heath was labelled alphabetically the progression from A to D is an ordinal progression from the least severe to the most severe. Patients are designated according to the physiologic total and chronic health classification in a two tiered evaluation.

On the basis of clinical opinion the variables were weighted with the objective of developing an interval scale. If an item is missing from a patient's record it is assumed to be irrelevant to the patient's evaluation and therefore given a normal value.

c) Reliability

No evaluations of the reliability of this instrument have been published.

d) Validity

i) Content Validity

The items in the index were selected on the basis of a literature review and clinical opinion. The inclusion of co-morbidity data with physiologic data is a strength of this instrument.

ii) Construct Validity - APACHE was correlated with mortality with r=0.47 and amount of therapy received r = 0.59. Within the midrange of the instrument (15-20) each point on the scale was associated with a 2% increase in mortality. The pre-existing disease classification when evaluated within regression models that included the physiologic score contributed to the prediction of death only when the patient was in the most severe (D) category.

iii) Predictive Validity - Using death as an outcome, the sensitivity of APACHE was 97% with a specificity of 49%.

e) Commentary

The APACHE classification system requires extensive information, much of which would only be collected in an ICU. The assumption of a normal value when data are missing will not be consistently true. How missing values would bias the instrument is unknown.

*2) ***APACHE II***[96,97,98,99]

a) Definition

i) Purpose and Application - The APACHE II system was developed to replace the APACHE system which was complex and lacked multi-

instituti onal validation. Like its predecessor, APACHE II's purpose is to classify critically ill patients in intensive care.

ii) Attributes - The attributes are classified as Acute Physiological Score (APS), Age Score and Chronic Health Score. The APS was reduced from 34 variables in APACHE to 12 variables in APACHE II. The reduction was based on R square contribution to a regression model which used hospital mortality as an outcome. Three of these 12 variables were reweighted on the basis of coefficients generated by the regression model. The age score was added to this instrument because age is known to contribute to mortality independently of disease severity. The presence of system failure or immunocompromise which was the D classification in the original APACHE instrument was the only factor used in evaluating chronic disease in APACHE II.

b) Scaling Techniques

In APACHE II the items were weighted on the basis of coefficients developed in a regression model, rather than only on clinical opinion, which was the technique used in APACHE. This makes the authors' claim to an interval scale more tenable. The APACHE II aggregate is the arithmetic sum of APS points, age points and chronic health points.

c) Reliability

The authors report a 96% agreement in the chart abstraction of physiologic data and that "disagreements were minor" for the determination of pre-admission health status data. Formal reliability testing has not been reported.

d) Validity

i) Content Validity - APACHE II has demographic data (age), physiologic data and co-morbidity data (system failure or immunocompromise). The addition of age increases its information content over APACHE.

ii) Predictive Validity - The authors found a high correlation of the APACHE II score with death. They noted that the overall risk of death was correlated with the specific disease being evaluated. The areas under the ROC curve (A_{uc}) for APACHE II and the original APACHE were similar (.863 and.851) as were the R square values (.319 and .310). The predictive validity of APACHE II varies with the disease state evaluated.

3) ***SAPS (Simplified Acute Physiology Score)***[99,100]

a) Definition

i) Purpose and Application - SAPS was also developed as a simplification of the original APACHE system. The developers of this instrument wanted to avoid the assumption used in APACHE that missing data should be designated as normal. SAPS uses data collected in the first 24 hours after admission to facilitate clinical decisions, multicenter studies and outcome comparisons.

ii) Attributes - The attributes include thirteen of the physiologic items from the APACHE system, with the addition of age. No pre-admission co-morbidity information is included in this instrument.

b) Scaling Techniques

Each of the 14 items in SAPS is coded from 0 to 4. The arithmetic sum of the items is the final SAPS score.

c) Reliability

No evaluations of this instrument's reliability have been published.

d) Validity

i) Content Validity - Age is included in SAPS, but co-morbidity data are not; thus it covers a higher number of relevant information domains than APACHE, but less than APACHE II. SAPS includes more physiologic items than APACHE II.

ii) Predictive Validity - SAPS had a sensitivity and specificity of 69% for mortality at a cutoff of 12. The ROC curve of SAPS and APACHE were reported as not significantly different although A_{uc} values and standard errors were not published.

e) Commentary

SAPS is a system very similar to APACHE II. Weighting of the items was based on expert opinion with no objective attempt to develop an interval scale. It is a simpler instrument than APACHE II as co-morbidity items are not included. It has not been evaluated for use with trauma patients alone.

4) *CARE (Cardiorespiratory Physiologic Studies)*[38]

a) Definition

i) Purpose and Application - The CARE system uses pattern recognition systems to track the status of critically ill trauma patients and to help make management decisions in individual patients. The information for the items included in this instrument are only available in ICU patients who are being invasively monitored.

ii) Attributes - Twelve attributes related to cardiovascular, respiratory, and metabolic status are included in this instrument.

b) Scaling Techniques

The 12 variables have been standardized to the mean and standard deviation of a control group. The items are placed radially on a circle diagram with 9 concentric circles. The 5th circle is the R state or normal and the circles interior to or exterior to the R circle, are the negative or positive Z scores. From these diagrams four states are recognized: A- compensated stress, B- metabolic imbalance, C- respiratory decompensation and D - cardiac decompensation. An individual patient's "physiologic trajectory can be plotted on a graph using distance from the R state to quantify the patient's physiologic state. The A to D states would be considered a nominal scale; the physiologic trajectory would be considered an ordinal scale.

c) Reliability

No evaluations of this instrument's reliability have been published.

d) **Validity**

It has been shown in a before/after study, which did not control for disease severity that trauma mortality decreased from about 25% to 10% after the institution of CARE in an ICU situation.

e) **Commentary**

CARE is an interesting concept using pattern recognition to assist in management of trauma patients. The reliability of this instrument has not been evaluated and no satisfactory validation studies have been performed.

C. General Commentary on Hospital Instruments

The hospital instruments have been evaluated with ROC curves, the use of inferential statistics and consistent endpoints. The physiologic profile system of CARE has not been compared to any of the APACHE derived instruments. In a multicenter study of 691 trauma victims evaluated with TS, ISS and the APACHE II, all three instruments contributed significantly to a model predicting mortality. The area under ROC curves for the three instruments were not significantly different and ranged from .79 for the ISS, .81 for the TS and .85 for APACHE II.

III. OUTCOME INJURY SEVERITY INSTRUMENTS

A. Overview

The outcome instruments are used to evaluate trauma care retrospectively. There are two general categories of instruments, those that derived from the Abbreviated Injury Score and those based on the International Classification of Diseases (ICD). The outcome instruments, especially the Injury Severity Score, have also been used to evaluate the predictive validity of pre-hospital instruments. Table 2.5 reviews the nomenclature and indicates that both ordinal and interval scales are used in the outcome instruments.

Table 2.5: OUTCOME INSTRUMENTS
Nomenclature and Scaling Mechanism

	Instrument Name	Scale
1. AIS	Abbreviated Injury Scale	Ordinal
2. ISS	Injury Severity Score	Ordinal
3. TRISS	Trauma Score (or RTS) and ISS	Interval
4. ASCOT	A Severity Characterization of Trauma	Interval
5. RESP	Revised Estimated Survival Probability	Ordinal
6. AI	Anatomic Index	Interval
7. PMC	Patient Management Category	Ordinal
8. OPS	Outcome Predictive Score	Ordinal

True reliability assessments have been used infrequently in the assessment of this category of instrument. The reliability of the AIS and by extension the reliablity of its derivative the ISS has been evaluated. The Washington Hospital center uses the

assessment of goodness of fit as an assessment of reliability for TRISS and ASCOT. As noted earlier this is not an assessment of inter-rater reliability.

Table 2.6 indicates that the information content of several outcome instruments is quite comprehensive and that ROC curve analysis has been used more frequently than on triage instruments.

Table 2.6: OUTCOME INSTRUMENTS
Content And Predictive Validity Assessments

	INFORMATION CONTENT					PREDICTIVE VALIDITY		
	Phys	Anat	Age	Mech	Morb	Matrix	LR	ROC
1. AIS		*						
2. ISS		*				*		*
3. TRISS	*	*	*	*		*		*
4. ASCOT	*	*	*	*		*		*
5. RESP		*					*	
6. AI		*				*		

Column Headings: Phys - Physiologic, anat - Anatomic, Mech - mechanism of injury, Morb - co-morbidity, Matrix - decision matrix, LR - Likelihood ratio, ROC - receiver operator characteristic curve.

B. Individual Outcome Injury Severity Instruments

Six outcome instruments will be reviewed in detail; four of these are based on the AIS and two on ICD codes. Three other approaches will be briefly discussed under the heading "others".

1) ***Abbreviated Injury Scale (AIS)***[101,102,103,]

a) **Definition**

i) Application and Purpose

This scale was developed in 1971 to rank injuries according to severity, and to standardize terminology used to describe injury. Its original purpose was for rating tissue damage sustained in motor vehicle accidents to facilitate the evaluation of the relationship of injury incidence, and mechanism, to vehicle design. The AIS classifies individual injuries but does not assess the effect of multiple injuries in the same patient. Two instruments, the injury severity score and the anatomic profile, have been derived from the AIS to evaluate the multiply-injured patient. These instruments are presented later in this section. The AIS has been modified to include penetrating injuries and the immediate consequence of injury, to enable the instruments which are derived from the AIS to evaluate both blunt and penetrating trauma. The AIS is used after hospital separation and is obtained by the extraction of data from the hospital charts.

ii) Attributes - For the purpose of scoring the body is divided into seven regions: 1. external; 2. head; 3. neck; 4. thorax; 5. abdomen/pelvis; 6. spine; 7. extremities. The attributes are anatomic descriptions of injuries which are defined in the 73 page AIS dictionary.[102]

b) Scaling

AIS scores are based on integers from 1 to 6. AIS is an ordinal scale for each of the seven body regions coded as 1-minor, 2-moderate, 3-serious, 4-severe, 5-critical, 6-unsurvivable and 9-unknown. The AIS is not simply a ranking of expected mortality, as this would make it impossible to distinguish moderate and

minor injuries, which do not threaten life. The AIS does not aggregate scores from the seven body regions.

c) Reliability[103]

Mackenzie studied the inter-rater reliability of the AIS, by comparing AIS scores obtained by surgery residents, registered nurses, paramedics and medical records technicians from the same charts. It was noted that correct determination of the number of injuries was a problem with paramedics recording fewer injuries, than the other occupational categories. When the number of injuries was not accounted for, the reliability as determined by the weighted Kappa statistic was moderate (0.41 to 0.60); however when specific injuries were compared the agreement on the coding was substantial (K 0.61-0.80).

d) Validity

The AIS is a consensus derived anatomical system of injury description. The maximum AIS (MAIS) has been useful in research concerned with vehicle design change, but has not been found useful in trauma research.

e) Commentary

The AIS is important in the research of injuries associated with vehicle design, but its importance in the measurement of injury severity is due to the instruments derived from the AIS.

2) *Injury Severity Score (ISS)*[103,104,105,106,107,108,]

a) Definition

i) Application and Purpose - The ISS was developed to numerically describe the overall severity of injury. It is applied retrospectively by chart review to patients who have injuries involving one or more body regions. The ISS was originally described for blunt trauma victims, but has been adapted for use in penetrating trauma and also for use with children.

ii) Attributes - The attributes consist of the three most severely injured body regions as determined by the AIS. The body is evaluated as 6 regions thus differing from the AIS system which defines 7 regions. The regions defined by the ISS are: 1. head or neck 2. Face, 3.Chest 4. Abdominal or pelvic contents 5. Extremities or pelvic girdle. 6. other external.

b) Scaling Techniques

The ISS is the sum of the squares of the three regions with the highest AIS scores. An AIS score of 6 for any region is considered unsurvivable and is coded at the maximum ISS score, which is 75. The AIS does not have a linear relationship with mortality and therefore the ISS was developed. The simplest higher order relationship, a quadratic, was found to have a linear relationship to mortality and therefore was chosen as the scaling mechanism. The scaling techniques used in determining the ISS has no statistical or biological explanation. In 1974 Baker suggested that the quadratic relationship may reflect fundamental aspects of response to injury, but no such relationship has been delineated since that time. Including AIS scores from more than three body regions, did not improve the function of the ISS. The ISS is an approximation of an ordinal scale.

c) **Reliability**[103]

The reliability of the ISS depends on the reliability of the AIS (see above). Mackenzie noted that a lower reliability in the AIS may be acceptable for a single injury but could become unacceptable after magnification due to the quadratic transformation, used in determination of the ISS.

d) Validity

i) Content Validity

The ISS included only anatomic items from the AIS; no physiologic, demographic or mechanism of injury or co-morbidity attributes are included. The AIS originally did not describe penetrating injury but in 1985 was altered to include these items thereby improving the face validity of the use of ISS in cases of penetrating trauma.

ii) Predictive Validity - Baker in 1974 demonstrated that the ISS had a linear relationship with mortality and that the ISS was correlated with time interval until death. Bull in a series of 1333 road traffic accidents demonstrated that the ISS was related to mortality, time to death, mean hospitalization time and severity of disability. Bull also determined that the LD 50 (lethal dose for 50%) for the ISS was 40 for young adults, 29 for middle age and 20 for the elderly.

Copes found it necessary to review the ISS in 1988, following the changes in the AIS which were introduced after the evaluations of the ISS by Baker and Bull. Copes notes that the ISS is not strictly monotonic in its prediction of mortality. He believes that this is due to the heterogeneity within the same ISS summary scores. He also noted the loss of information inherent

in the fact that a body region could have multiple injuries but only the most severe injury in that region is included in the summary score.[107]

ROC curve analysis of the ISS measurement of critically ill patients was performed as part of an evaluation of APACHE. The A_{uc} for the ISS when mortality was the outcome was 0.79. Sensitivity and specificity figures have not been published for the ISS and could not be derived from the published ROC curve.[98]

Goldberg et al. in a multivariate analysis of RESP showed that the ISS was a significant predictor of mortality, ventilatory need and length of hospitalization. In their study ISS explained 31% of the variance associated with injury mortality; Baker in her original description stated that 49% of the variance was explained by the ISS.[9,104]

e) Commentary

The ISS has weaknesses in its scaling mechanisms and in the predictive validity in patients with multiple injuries to one body region. It is however widely utilized both as an assessment of injury severity and a means of assessing and comparing triage instruments. Because there is no instrument in traumatology which can be considered a gold standard, the ISS often fills that role.

*3) ***TRISS***[109,110,111,112]

a) Definition

i) Purpose and Application - The TRISS system's purpose is to be a standard objective system for the evaluation of trauma care. The TRISS score is

determined from physiologic information recorded in the emergency department and anatomical, demographic and mechanism of injury information extracted from the patient's medical record.

ii) Attributes - The attributes of TRISS and Revised-TRISS are the Trauma Score (TS) or Revised Trauma Score (RTS), patient age category, and the Injury Severity Score (ISS). Patients with blunt injuries are analyzed separately from patients with penetrating injury.

b) Scaling Mechanisms

The coded values for the TS or RTS and the ISS are determined as noted under their description as individual instruments. Age is recorded as 0 for under age 55 and 1 for over age 54 and the patients are categorized as having suffered either blunt or penetrating injury. The aggregate score of TRISS is the probability of survival, as determined using a logistic function. Coefficients for blunt trauma and coefficients for penetrating trauma are calculated for the three co-variates included in TRISS. These coefficients originate from a North American standard, the Major Trauma Outcome Study (MTOS). The coefficients for blunt trauma are used for both blunt and penetrating trauma in children.

TRISS is used to audit the outcome of individual patients by reviewing patients who died when TRISS predicted a probability of survival of greater than 50%, and by reviewing patients who live when TRISS predicted a survival of less than 50%. The proponents of this evaluation system label this process PRE (see Information Theory - Chapter 1).

The mortality in a specific institution can be evaluated by comparing the actual number of deaths at that institution to the expected number of deaths as determined by TRISS. An absolute value of Z greater than 1.96 indicates a statistically significant difference from the MTOS standard results which were used to determine the coefficient used in TRISS. The proponents of this evaluation system label this process DEF (definitive).

A statistic called M has been developed to determine the match of the predicted probability of survival of the study population to that of the MTOS. The lower the value of M, which ranges from 0 to 1, the less confidence one can place in the Z scores.

c) Reliability

The reliability of the TS and the AIS (the basis for ISS) have been evaluated individually as noted when those instruments were reviewed. One study of TRISS evaluated the inter-rater reliability of AIS and found the intraclass correlation to be 0.78. The reliability of the RTS has not been evaluated and the reliability of TRISS as a composite index has not been determined.

d) Validity

i) Content Validity - TRISS combines physiologic (RTS), anatomic (ISS), demographic (age) and mechanism of injury (penetrating vs blunt) information to determine a patient's probability of survival. Because many of the domains of interest in a injury severity instrument are included and several attributes in TRISS are instruments that are recognized individually, the face and content validity of TRISS are high. Co-morbidity information is not included and the mechanism of injury information is non-specific.

ii) Predictive Validity - Pre-charts which have either the TS or the RTS on the ordinate and the ISS on the abscissa show that the majority of trauma deaths occur above the 50% probability isobar and almost all individuals below the isobar survive. Sensitivity and specificity of TRISS for penetrating trauma has been reported to be 80.7% and 98.6% respectively while for blunt trauma the sensitivity is only 58.5% with a specificity of 99.3%. ROC curves and likelihood ratios have not been published for TRISS.

TRISS has been criticised because it is limited in its ability to predict mortality in low falls, does not account for multiple severe injuries to a single body region and does not account for the heterogenicity of penetrating trauma ie. stab wounds vs. fire arms. Other critics have warned that the Z-statistic evaluation of outcome will miss important differences in small populations or in populations with a low proportion of severely injured patients.

e) Commentary

TRISS combines the most widely used physiologic score with the most widely used anatomic score and despite some deficiencies is the primary method for evaluating the performance of trauma systems. This instrument reports good sensitivity with high specificity in one study.

4) Revised Estimated Survival Probability Index *(RESP)*[9,113]

a) Definition

i) Purpose and Application - RESP, which is calculated from the international classification of diseases(ICD) codes, is proposed as an alternative to the ISS. ICD coding is available on the face sheet of all hospital charts and therefore does not require the extensive survey of the chart which is necessary for determination of the ISS. RESP is recommended as a means of adjustment for severity mix in trauma outcome analyses but is not designed for evaluation of individual patient outcome.

This instrument is applied retrospectively by medical records technicians who review the ICD codes on the face sheets on the chart of discharged trauma patients.

ii) Attributes - The items in RESP are selected from the ICD-8 codes which indicated traumatic injury. Burns, effects of chemical substances and radiation and iatrogenic diseases are excluded leaving ICD codes 800.0-939.0,950.0-959.9 and 991.0-996.9.

b) Scaling Techniques

The age specific single condition survival rate was calculated for each ICD trauma category, using a data set of 77,600 patients. The final probability for each patient is the product of all the single condition survival probability rates, recorded in the medical record.

c) **Reliability**

The reliability of RESP has not been formally addressed. The level of agreement of the ICD-8 face sheets has been evaluated by the Institute of Medicine. A review by the institute showed an uncorrected agreement of only 63.4% on the primary diagnosis. It is noted that the order of diagnosis which is not important in the determination of the RESP, was included in this evaluation and contributed to the low level of agreement noted in the study.

d) Validity

i) Face Validity - Trauma codes in the ICD were not developed with grading of injury severity as a specific objective and some codes cover a very heterogeneous group of injuries. An index based only on ICD codes must be viewed with some scepticism. The information content of ICD codes can be considered to be anatomical.

ii) Predictive Validity - In a 1984 study RESP was determined from the face sheets of medical records and after re-extraction using the full chart. The two RESP probabilities were evaluated independently and also compared to the ISS using regression techniques with mortality, length of hospital stay and need for ventilatory assistance as outcomes. Likelihood ratios for RESP were inferior to the ISS and the 95% confidence intervals for the likelihood ratios for the RESP scores obtained from the face sheet included the null value of 1.

e) Commentary

RESP was developed to avoid the full chart extraction process necessary to determine the ISS. Validation studies show that RESP only approached the predictive validity of the ISS when the entire chart was reviewed. The authors

note that a high standard of quality control is necessary if hospital discharge face sheets are to be used in determination of RESP. They speculate that RESP will improve with the increased precision available in the ICD-9 coding of injuries. RESP was designed to assess outcome of trauma victims without the use of extensive chart reviews. The most recent study demonstrates that this objective has not been achieved.

5) ***Anatomic Index***[114,115]

a) Definition

i) Purpose and Application - The anatomic index was described to determine the probability of death in patients with blunt injury. The application of this instrument is through a retrospective hospital chart review of the ICD codes.

ii) Attributes - The attributes are the injury codes of the international classification of diseases.

b) Scaling Techniques

The probability of death conditional on associated injuries was determined for each injury code. The conditional probabilities were used to rank the injuries for each individual and then the probability of mortality for each injury was determined. This second probability is termed the effective probability. The Anatomic Index is the effective probability of the most serious injury in an individual patient as ranked by the conditional probability.

c) **Reliability**

The reliability depends on the reliability of ICD codes which as was noted in the discussion concerning the RESP index, has not been formally assessed. The agreement level which was uncorrected for chance, was only 0.63.[113]

d) **Validity**

Using the PER technique to evaluate mortality prediction, the information gain of .60 was determined for the AI when contrasted with a "perfect index". In this evaluation the AI was considered to predict death if the effective probability was greater than 0.50. The sensitivity of AI was 88.5% and the specificity was 70.3%. These results were determined from a subset of the study population set aside from the data used to create the instrument and has not been independently validated since this instrument was introduced in 1980.

e) Commentary

The AI is dependent on ICD coding which itself can be unreliable. The instrument calculates the probability of death from the most serious injury. The most serious injury in this instance is really used to define a patient's injury profile. This concept is receiving renewed interest in measurement research. The AI has not been updated for newer ICD codes or validated in independent patient populations. The concept of injury profiles based on the most serious injury is interesting.

6) *ASCOT (A Severity Characterisation Of Trauma)*[110,116]

a) Definition

i) Purpose and Application - ASCOT has been described to improve some of the deficiencies of TRISS. Specifically, the ISS, which is the anatomic component of TRISS is replaced by the Anatomic Profile (AP). The proponents of ASCOT state that the ISS underestimates the seriousness of injury in trauma victims with multiple injuries to one body region. ASCOT is determined in a chart review performed by medical records technicians.

ii) Attributes - The attributes include physiologic (RTS), anatomic (AP), demographic (age) and mechanism of injury (blunt or penetrating) items. RTS and mechanism of injury are coded in the same way as in TRISS. Age is coded as 0 for less than 55, as 1 for 55-64, 2 for 65-74, 3 for 75-84 and 5 for >85. The Anatomic Profile (AP) which is a derivation of the AIS consists of 4 profiles: A - Central nervous system with AIS levels 3-5, B. Thoracic and throat regions AIS levels 3-5, C. Other serious injuries AIS levels 3 to 5, and face AIS levels 1 or 2, D.Other Injuries AIS levels 1 or 2.

b) Scaling Mechanisms

Originally clustering techniques using vector analysis of seven dimensional space was used to develop 80 patient profiles for the AP; in a more recent article, logistic regression was used to predict survival and develop coefficients for the three components of the RTS, the four components of the AP, and for the coded age. Unlike the ISS, all AIS scores are included in the determination of the AP. The value for each body region is the square root of the sum of squares for all the injuries. Coefficients for the components of ASCOT are

determined separately for blunt and penetrating trauma. No evidence supporting the use of ASCOT in paediatric patients has been published.

The summary score for ASCOT is the probability of survival determined by adding the products of the coefficients previously determined and the coded values of the covariates.

c) Reliability

The Hosmer - Lemeshow statistic which evaluates goodness of fit of a logistic regression model, was used to evaluate predictive reliability. This statistic which is not an evaluation of reliability, was substantially lower for ASCOT than for TRISS in both penetrating and blunt trauma, indicating that ASCOT is a better model than TRISS. The authors presume a Hosmer-Lemeshow statistic below 15.5 indicates good statistical agreement. For penetrating trauma the statistic for ASCOT was 12.65 but for blunt trauma it was 24.8.

No evaluations of inter-rater reliability or other assessments of agreement have been performed.

d) Validity

i) Face and Content Validity - The information content of this instrument is similar to TRISS but it retains information concerning multiple injuries to the same body region and also includes more age categories.

ii) Predictive Validity - In its one published evaluation ASCOT had similar sensitivity and specificity to TRISS. For blunt trauma the sensitivity was 63.3% with a specificity of 99.2% while for penetrating trauma the sensitivity was 86.1% with a specificity of 98.7%.

e) Commentary

The theoretical advantages of ASCOT are that it accounts for multiple injury to the same region and that the four components of the AP are not lost in a summary figure as they are in the ISS. ASCOT may be an improvement on TRISS but this remains to be proven.

7) ***Others***

These instruments are briefly presented because of specific facets which make them interesting contributions. Patient management categories are also used in this thesis to define major trauma.

a) **Preventable Death** [117,118,119]

While not strictly an instrument, trauma care outcome can also be evaluated by assessing preventable deaths. In this method of evaluation, experts use either clinical and autopsy information or autopsy information alone to determine the proportion of deaths that would have been preventable in an optimal situation. This method has been particularly important at drawing attention to the problem of the high mortality associated with low injury severity in rural areas of North America.

*b) **Patient Management Categories**[120]

A combination of ICD-9 specific diagnosis and face sheet procedure codes has been used to develop 126 patient management categories. This system allows the identification of the most severely injured patients , "tertiary" trauma patients, whether or not they are seen at a trauma center. It was noted that this method accurately identified 97.8% of trauma victims that were included in the Pennsylvania trauma registry. The use of PMC's has been criticized as a means of evaluating trauma severity because it uses treatment as a means of classifying

disease severity retrospectively. An interesting advantage is that it identifies patients that should be, but are not included in a trauma system.[120,121]

c) **Outcome Predictive Score (OPS)**[122]

The OPS addresses the prediction of death and major infection in the trauma victim. Infection is the most common cause of late death in the trauma victim. This instrument uses the ISS, age, degree of bacterial contamination and the expressing of the monocyte HLA-DR antigen to predict major infection and death. According to the authors HLA-DR antigen expression is an inheritable trait the lack of which increases the risk of infection and subsequent death.

C. General Commentary on the Outcome Instruments

The outcome instruments that are based solely on ICD codes have either been shown to be invalid or have not been revalidated since their original presentation. The use of ICD codes has more recently been used with procedural codes to develop patient management categories which allow evaluation of patients outside a formal trauma system. The universal use of ICD codes continues to be an appealing means of evaluating the trauma patient.

The outcome measures that depend on the AIS have become the standard of evaluation of care within trauma registries. Problems with the ISS have been noted. Many of the problems noted with the ISS are inherent in TRISS. The predictive validity of TRISS seems to be high in the one study that reports this value but it is surprising that for an instrument that is this widely used, so little evaluation has occurred. Statistical presentations that allow for ready comparison of instruments and evaluation of measurement qualities are a problem with most outcome instruments.

The newest AIS derived outcome instrument, ASCOT will require time to evaluate. ASCOT depends on AIS-90 which many trauma registries have not as yet incorporated.

III. CONCLUSION

Injury severity instruments are used to make decisions that are critical to the life of trauma victims and to the function and economic well being of the trauma system. It is not possible to properly evaluate the predictive validity of these instruments because many of the studies evaluating the instruments do not use statistics or techniques that facilitate comparison. The few studies that permit direct and unbiased comparisons have brought into question the usefulness of these instruments. If this is the case then these instruments which are so central to trauma care should be improved or replaced. In addition to the deficiencies in the statistical evaluation of these instruments, this review demonstrates that many instruments have not been evaluated at the standard expected within the context of measurement theory. The definition, scaling mechanisms, reliability and content validity all effect an instrument's predictive validity. If we are to improve the predictive validity of instruments, these qualities should be assessed and when found deficient the instrument should be altered.

CHAPTER THREE

STUDY OBJECTIVE AND DESIGN

I. OBJECTIVE

The objective of this thesis is to compare several severity instruments used in traumatology and assess the importance of different qualities in instrument performance.

II. RESEARCH QUESTIONS

a) What is the predictive validity of the ISS? Is the predictive validity different in blunt and penetrating trauma?

b) Which of four pre-hospital instruments has the highest predictive validity?

c) Do scaling techniques, reliability or content validity effect the predictive validity of severity instruments?

III. DATA SOURCES

The above questions will be answered by analizing data from the: (a) literature; (b) Vancouver General Hospital (VGH) trauma registry; and (c) Pennsylvania State University at Hershey Pennsylvania (PSU) trauma registry.

a) **Literature Data**

Several published articles concerning the measurement of injury severity report data that permit the calculation of likelihood ratios and ROC curves. The studies used for these comparative evaluations are presented in Table 3.1. ROC analysis was also used to evaluate the ISS using data published by the MTOS on 14,876 trauma victims[107] and to evaluate the Mechanism of Injury using information on a study of 500 patients reported in a study from San Jose.[913] The use of published data facilitated the evaluation of instruments that are not included in the VGH and PSU data sets (see below) ; it also permitted re-evaluation of the authors' conclusions concerning their instrument.

Table 3.1: LITERATURE DATA SOURCES FOR PRE-HOSPITAL INSTRUMENTS

Instrument	N	Mortality	Study Center
CRAMS	2110	3.5%	Salt Lake City, Utah[42]
PHI	3581	2.9%	Multi Center Trial[64]
RTS	2100	6.1%	Washington, DC[43]
RTI	2340	4.2%	Hershey, Pennsylvania[39]

b) **Vancouver General Hospital (VGH) Data**

Data from the first year of the Vancouver General Hospital trauma registry were used for some aspects of this thesis. The Vancouver General Hospital, a university hospital for adults, is a designated Level I trauma center with an immediate catchment area of 1.2 million. It also is a referral center for the entire province of British Columbia. One third of the trauma patients are transfers from outside Vancouver.

Patients are entered into the VGH trauma registry by emergency department nurses. The nurses have been advised to include all patients who have a disease process caused by external force. Hospital admissions that are coded by ICD-9 as external trauma and which were not initially included in the registry are reviewed and included in the registry. Trauma cases are excluded if the injury event occurred more than 6 days prior to admission. A final decision concerning the entry of a questionable case into the registry is made by the surgeon who directs the trauma service.

The VGH data base includes 50 variables on 1578 patients treated between November 10, 1989 and November 9, 1990. Outcomes variables evaluated in this study were death, discharge disposition, admission to intensive care, length of hospitalization and immediate surgery. The components of the RTS were collected in the field and within 15 minutes of arrival in the emergency room. These data were used for the comparison of scaling techniques and the determination of reliability. Variables used in the evaluation of content validity were age, ISS, class of injury and mechanism of injury. The Injury Severity Score (ISS) was collected only for patients who died or were in hospital for more than two days; therefore the maximum VGH data set, available for analysis when the ISS was used, was 1161 of the 1578 subjects (73.6%). The missing variable that most affected the data available for analysis was the Glasgow coma scale. Because this information is part of the revised trauma score (RTS), pre-admission RTS was available on 893 (56.6%) and admission (RTS) was available on 518 (32.8%) of the data set. For the reliability study both RTS results were available for 367 trauma victims. The admission RTS was supplemented with the emergency RTS to increase the observations available for analysis to 1041 (66.1%). The missing Glasgow coma scale reduced the population available for analysis. The subset of the VGH data used

in the analysis is indicated with the different analyses which are presented in Chapter 4. The mean age, mortality and ISS for these subsets (see Table 3.2) do not indicate a selection bias.

Table 3.2: COMPARISON OF VGH DATA SUBSETS

	Mean AGE	Mean ISS	Mortality
1. All cases	42.5	14.3	5.8%
2. Admission RTS	45.3	14.0	6.6%
3. Combined RTS	44.5	14.6	6.7%

c) **Pennsylvania State University at Hershey Data (PSU)**

Data from the Milton S. Hershey Hospital trauma registry were also available for analysis. The catchment population which is concentrated within a 70 mile radius of this hospital is 1.2 million. The hospital is a university hospital which is designated as a Level I trauma cente

All age categories are included in the registry as are the transferred patients who come from the immediate geographic vicinity. Entry into the trauma registry is based on the ICD-9 categories for external trauma.

The data from PSU included 43 variables on 3079 patients collected from 1987 to 1991. Both the RTS and the RTI (revised trauma index) are included in this data set. The assessment of scaling mechanisms using the RTS was re-evaluated using the PSU data set. Outcome variables evaluated were mortality, patient management category and the ISS. Variables used in evaluating content validity included mechanism of injury and number of serious injuries. Missing data

reduced the data set to 2979 (96.1%). The mortality, mean age and mean ISS are presented in Table 3.3.

Table 3.3: DESCRIPTION OF PSU DATA

N	Mean AGE	Mean ISS	Mortality
2979	42.5	14.3	5.8%

IV. ANALYSIS

a) Likelihood Ratios

The pre-hospital instruments were compared by determining both the likelihood ratio positive (LR+) and the likelihood ratio negative (LR-) of each instrument at the cutoff level recommended in the literature. Because of the skewed distribution of the statistic, Taylor series confidence intervals were calculated using a logarithmic transformation.[123] The likelihood ratios of the instruments were then assessed for statistical and clinical differences.

b) ROC Curve Analysis

ROC curves and the area under the curve (A_{uc}) were determined using ROCFIT which was developed by C Metz at the University of Chicago and by using either Delong's Method or PROC Logistic within the statistical analysis system (SAS).[76,78]

ROCFIT was used for unpaired comparisons of data from the literature. This program calculates A_z with a standard error and parameters for graphic representation of the curve. ROCFIT is used with grouped data which makes it convenient for calculations using data from the literature.

Studies comparing different instruments performance on the same data set (VGH or PSU) were performed using Delong's method. This non-parametric method takes into account the paired nature of the analysis thereby reducing the size of the standard error and resulting in a more powerful evaluation. This program calculates the area under the curve but does not estimate the parameters necessary for graphic representation of ROC curves.[77,78]

The Logistic Procedure in SAS calculates what it calls a Rank correlation, "c", for assessing the predictive ability of a model. C is the same statistic as A_{uc} and will be referred to as A_{uc} in this thesis. This fact allows one to use the multivariate analysis capabilities of the logistic procedure to determine A_{uc}. Because a standard error is not determined by the Logistic Procedure, Delong's method was used for significance testing and paired analyses.[124]

ROC curve analysis was used to evaluate the predictive validity of the ISS and several triage instruments. It was also used to compare scaling techniques, the predictive validity of pre-admission versus admission RTS and the content validity of different instrument attributes. As noted in chapter one content validity judges whether the relevant domains are included in the instrument. A review of the trauma literature (see chapter 2) suggests that the relevant domains in a severity instrument include physiologic, anatomic, co-morbidity, mechanism of injury and demographic (age) information.

c) **Reliability Determination**

Reliability was evaluated in the 367 VGH patients. SAS was used to calculate correlation coefficients (PROC CORR), plot graphs (PROC PLOT) and do the two way anova (PROC ANOVA) necessary for the intraclass correlation coefficient(ICC). The confidence intervals for the ICC were calculated using Satherthwaite's approximation.[125,126]

d) **Logistic Regression**

The Logistic procedure was used to determine both the weighting and the significance of attributes added to the RTS in the evaluation of content validity. The score statistic is a chi-square evaluation of the adequacy of the fit of the model after adjusting for the number of covariates. This statistic is presented with the logistic regression models in chapter 4.

e) **Statistical Significance**

In this thesis statistical significance is defined as p values less than 0.05 and confidence intervals at the 95% level. All p values are two sided evaluations.

V. OUTCOMES

The outcome evaluated must be defined to permit meaningful comparison. The outcomes for the different data sources are described below.

a) **Literature Data** - Several outcomes are used in the literature but the only outcome that was used in all the articles for the four instruments of interest was mortality. Other outcomes are therefore not used in the ROC curve analysis because simultaneous comparison using different outcomes is not valid. As the

power of ROC curve analysis is reduced when obvious cases predominate subsets of the literature data were used to evaluate pre-hospital instruments. One subset was defined by excluding the least injured and most injured patients the other subset by excluding only the least injured patients.

b) **VGH Data** - In the VGH data set, mortality, need for surgery, length of hospitalization, length of intensive care admission and separation disposition were outcomes that could be associated with major trauma. The ISS was used as a means of assessing the outcomes that would be most useful.

Clearly mortality is the most important outcome as well as the easiest outcome to measure. Mortality was the outcome most frequently used.

Disability is another outcome of importance and an hierarchal index of disability was developed in the VGH data set. The outcomes in this index were death (D0), long term care (D1), home with nursing or rehabilitation (D2) and home without assistance (D3). For the purposes of ROC analysis this outcome was dichotomized to "death or institutionalization" (DD0) and "home with or without assistance" (DD1). Analysis of variance of the Disability Index demonstrated that the mean ISS for the four levels of this index were different in magnitude and that these differences were highly significant ($p<.0001$). Bonferroni corrected T-tests of the different means demonstrated that all possible main effects were significantly different. (see Table 3.4) Two sample t-test comparison of the dichtomized outcome showed similar differences (see Table 3.5).

Table 3.4: DISABILITY INDEX

	N	Mean ISS	SE	P-value
D0	92	35.89	2.36	
D1	128	17.16	1.34	
D2	367	13.22	0.50	
D3	574	10.96	0.01	<.0001

Table 3.5: DICHOTOMIZED DISABILITY INDEX

	N	Mean ISS	SE	P-value
DD0	220	24.99	1.40	
DD1	941	11.84	0.28	<.0001

Transfer from the emergency department directly to the operating room in combination with mortality was not selected as an outcome of interest in the VGH data. Other studies that have used surgery as an indicator of major trauma have included only neurological or truncal surgery that was performed immediately. In this data set it was not possible to precisely define the type of surgery performed. One third (33.2%) of the patients in this data set either died or went from the emergency department to the operating room. It is likely that many of the immediate surgery patients had extremity trauma and that they were housed in the emergency department while waiting for an operating room.

Admission to ICU was evaluated as an outcome and thought to be useful (Table 3.6). The Mean ISS for those admitted to ICU suggests that admission to ICU indicates major trauma. Exclusion or inclusion of mortality from this assessment

resulted in unimportant changes therefore the simpler (inclusive) outcome was used.

Table 3.6: MEAN ISS FOR VGH ADMISSIONS TO ICU

ICU	N	MEAN ISS	SE	P-value
No	1034	11.99	0.33	
Yes	127	33.35	1.31	<.0001

The Length of Hospital stay correlated less to the ISS than did direct transfer from the emergency department to the operating room (Pearson Correlation Coefficient = .266 se = .03)

Three outcomes will be used in assessing predictive validity in the VGH data. These will be mortality, disability index and admission to the ICU. These outcomes all have face validity and have also had construct validation with the ISS.

c) **PSU data**

Three outcomes were evaluated using the PSU data: mortality, Patient Management Categories (PMC), and the ISS. Mortality was used for the content validity studies and the ROC evaluation of the RTI.

PMC define the seriousness of a patient's injury based on a combination of ICD-9 codes and the treatment received. PMC are rated as minor, serious or tertiary. The outcome was dichotomized to major trauma (tertiary) non-major trauma (minor and serious). The ISS was dichotomized as minor trauma < 15 and major trauma > 15. These three outcomes were used to directly compare

the RTS and the RTI in the same data set. This allowed a comparison of the predictive validity of these instruments using outcomes of differing difficulty.

CONCLUSION

Uniform statistical techniques will facilitate the ranking of the performance of injury severity instruments, thus enabling the evaluation of these differences within the context of measurement theory.

CHAPTER FOUR

DATA ANALYSIS: SPECIFIC METHODOLOGY AND RESULTS

This chapter commences with an ROC analysis of the ISS. Following this, four triage instruments are evaluated using likelihood ratio and ROC analyses and then explanations for these results are sought through the analysis of instrument scaling mechanisms, reliability and content validity.

I. ROC EVALUATION OF THE ISS

The objective in this analysis is to determine the predictive validity of the ISS. ROC curves for the ISS were calculated from data collected on 14,876 trauma victims by the MTOS, using mortality as an outcome.[107] In this data set injuries are categorized into blunt and penetrating trauma and age under 50 years or over 49 years. The A_{uc}'s were determined for each of these four age/injury type categories, as well as a summary figure for all 14,876 trauma victims and are presented in Table 4.1.

Table 4.1: ROC CURVE ANALYSES OF THE ISS

Mechanism	Age	N	Mortality	A_z	SE
Penetrating	<50	3,424	7.3%	.9026	.0081
Penetrating	≥50	279	19.6%	.9238	.0170
Blunt	<50	8,214	5.1%	.8980	.0074
Blunt	≥50	2,544	10.0%	.8750	.0124
ALL	ALL	14,876	6.4%	.8942	.0081

The A_{uc} for the ISS is similar for both age groups and both mechanisms. The estimates of A_{uc} for penetrating trauma are larger than for blunt trauma; however

these differences are not significant. The ISS evaluated in this study was derived from the AIS-85 which was changed to include more information concerning penetrating trauma. Further evaluation of the ISS is presented later in this chapter in the section that evaluates content validity.

II. A TRAUMAMETRIC EVALUATION OF TRIAGE INSTRUMENTS

A) Comparison of Triage Instrument Performance

The objective in this analysis was to use likelihood ratio and ROC curve analysis to evaluate triage instruments according to predictive validity.

1. Likelihood Ratio Analysis

Likelihood ratios were determined from published data for CRAMS[42], PHI[64], RTS[43] and the Trauma Index[39] at the cutoff points recommended in the literature. The outcome evaluated was mortality (see Table 4.2).

Table 4.2: LIKELIHOOD RATIOS OF FOUR TRIAGE INSTRUMENTS

Index	N	%Death	Positive Likelihood	Negative Likelihood
CRAMS	2110	3.5%	44.4 (33.0 - 59.7)	0.04 (.014 - .126)
PHI	3581	2.9%	9.8 (8.87 - 10.8)	0.00
RTI	2340	4.2%	7.4 (6.2 - 8.9)	0.06 (.025 - .137)
RTS	2100	6.1%	3.8 (3.7 - 3.9)	0.04 (.017 - .075)

(95% Confidence interval)

The post-test probabilities of mortality were calculated for each instrument. The results for a positive test are presented in Table 4.3 and for a negative result in Table 4.4.

Table 4.3: POST-TEST PROBABILITIES AFTER A POSITIVE RESULT

		PRE-TEST			POST-TEST	
	N	Probability	Odds	LR+	Odds	Probability
Crams	2110	3.5%	.036	44.4	1.610	62%
PHI	3581	2.9%	.030	9.79	.292	23%
RTI	2340	4.2%	.044	7.43	.326	25%
RTS	2100	6.1%	.065	3.8	.247	20%

Table 4.4: POST-TEST PROBABILITIES AFTER A NEGATIVE RESULT

		PRE-TEST			POST-TEST	
	N	Probability	Odds	LR-	Probability	Odds
Crams	2110	3.5%	.03627	.04	0.0015	.15%
PHI	3581	2.9%	.02987	-	-	-
RTI	2340	4.2%	.04384	.06	0.003	.3%
RTS	2100	6.1%	.06496	.04	0.003	.3%

2. *ROC Curve Analysis*

Sufficient data were available in the literature to use the program ROCFIT to calculate A_{uc} values for CRAMS, PHI, and the triage-RTS; data provided by PSU allowed direct comparison of the RTI and the RTS using Delong's method. Comparative analyses which use the RTS as the reference instrument include the simple differences in the areas and the absolute performance gain (see Information Theory Chapter One). To permit comparison of the studies from the literature, mortality, the only outcome which was consistently reported, was the outcome used. The PSU data base permitted evaluation of mortality, patient management categories and the injury severity score as outcomes.

a) ROC evaluation of CRAMS, PHI and the RTS

Using the program ROCFIT the A_{uc} was calculated for CRAMS, PHI and RTS. The results presented in Table 4.5 are based on analysis of literature data.

Table 4.5: ROC ANALYSIS OF CRAMS, PHI AND RTS

Instrument	N	Mortality	A_{uc}	SE
CRAMS	2110	3.5%	.9960	.0014
PHI	3581	2.9%	.9873	.0028
t-RTS	2100	6.1%	.9665	.0052

CRAMS and PHI were contrasted with the RTS by calculating the difference in the A_{uc} (A_{uc} diff) and the absolute performance gain (APG)(see Table 4.6).

Table 4.6: COMPARISON OF CRAMS AND PHI WITH THE RTS

Instrument	A_{uc} Diff	APG	Z-Stat	P-Value
CRAMS	2.95%	5.9%	5.48	<.0005
PHI	2.08%	4.16%	3.5	<.0005

CRAMS and PHI have an A_{uc} which is significantly larger than the RTS. The magnitude of the differences between the RTS and both CRAMS and the PHI is small. The power of ROC analyses is reduced when the study population is weighted with easy cases.[76] These populations are heavily weighted with obvious cases; therefore to increase the ability to discriminate between these instruments further ROC evaluation was performed after exclusion of obvious cases from the extremes of health and injury. The cases that were excluded were the cases which were defined as the least injured or the most injured by the triage instruments themselves. Specifically CRAMS scores of 0 and 10; a PHI of 0, 1, 23, and 24 and t-RTS of 1 and 12 were defined as obvious and excluded. The results of this ROC analysis are presented in Table 4.7. This table also describes the proportion of the original population that is included in the subset studied, and the mortality rate in the subset.

Table 4.7: ROC ANALYSIS AFTER EXCLUSION OF OBVIOUS CASES

Instrument	N	Proportion	Mortality	A_{uc}	SE
CRAMS	707/2110	33.5%	6.9%	.9825	.0052
PHI	1031/3581	28.8%	9.2%	.9506	.0096
RTS	591/2100	28.1%	21.2%	.8334	.0193

The A_{uc} of the RTS dropped markedly when evaluated in a population that did not include easy cases. A second evaluation was performed using the subset that removed only the most healthy subjects (see Table 4.8). Specifically a CRAMS of 10, PHI of 0 and 1, and t-RTS of 12 were defined as the healthy extreme and excluded from the study population. The proportion of the original population in this subset and the mortality of the subset are also presented in this table.

Table 4.8: ROC ANALYSIS AFTER EXCLUSION OF HEALTHY CASES

Instrument	N	Proportion	Mortality	A_{uc}	SE
CRAMS	732/2110	34.7%	10.1%	.9880	.0040
PHI	1041/3581	29.1%	10.1%	.9550	.0090
RTS	725/2100	34.5%	35.9%	.7705	.0166

The ROC analyses of CRAMS, the PHI and the RTS using all cases (see Figure 4.1a), the subset with the easy cases excluded (see Figure 4.1b), and the subset with the healthy extreme excluded (see Figure 4.1c) are presented graphically. It is clear that the difference between the RTS and the other two instruments is small when obvious cases are included but is large when the obvious cases are excluded.

Figure 4.1a

Pre-hospital instruments:- Outcome: Mortality
All results included

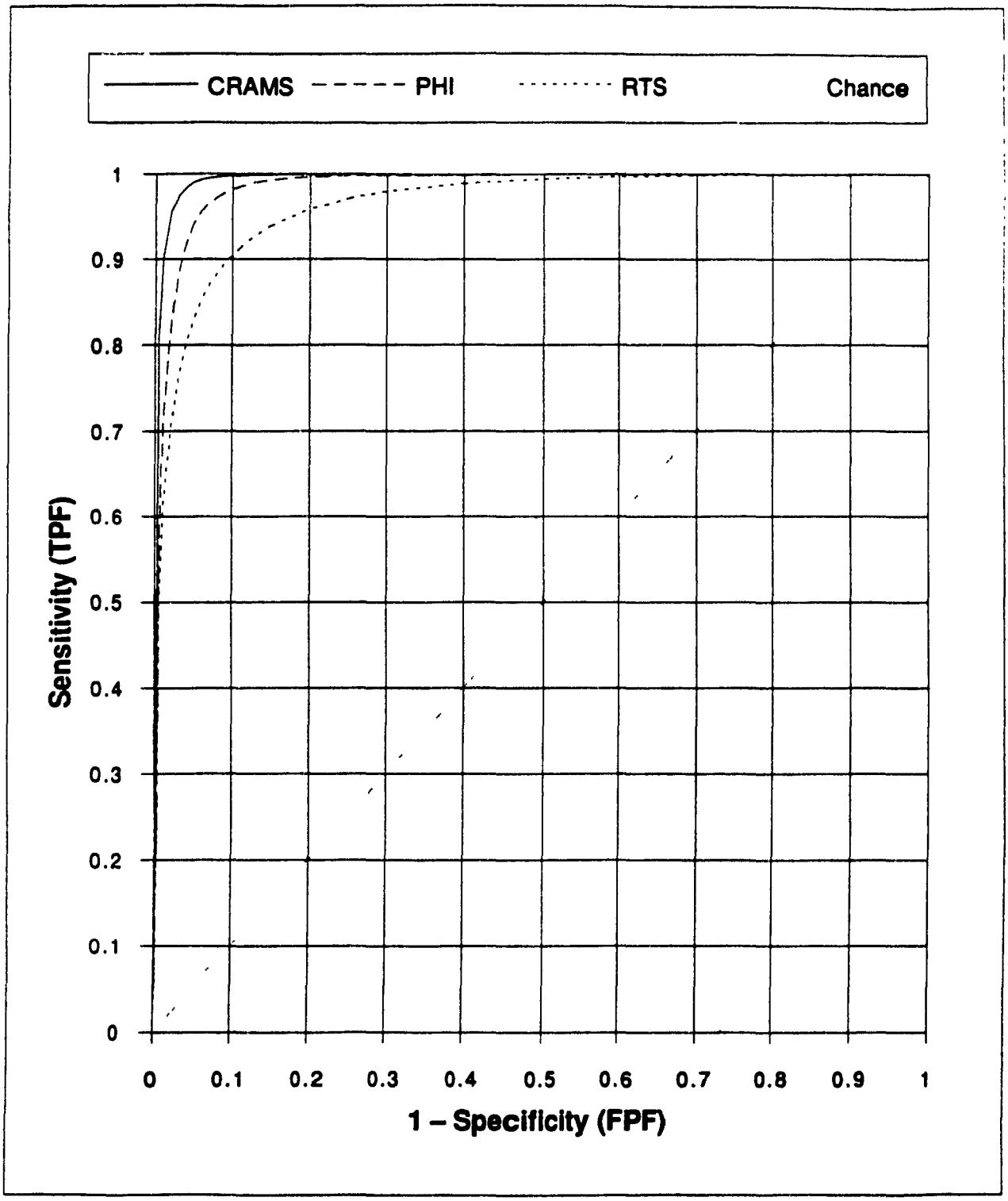


Figure 4.1b

Pre-hospital instruments:- Outcome: Mortality
Easy cases excluded

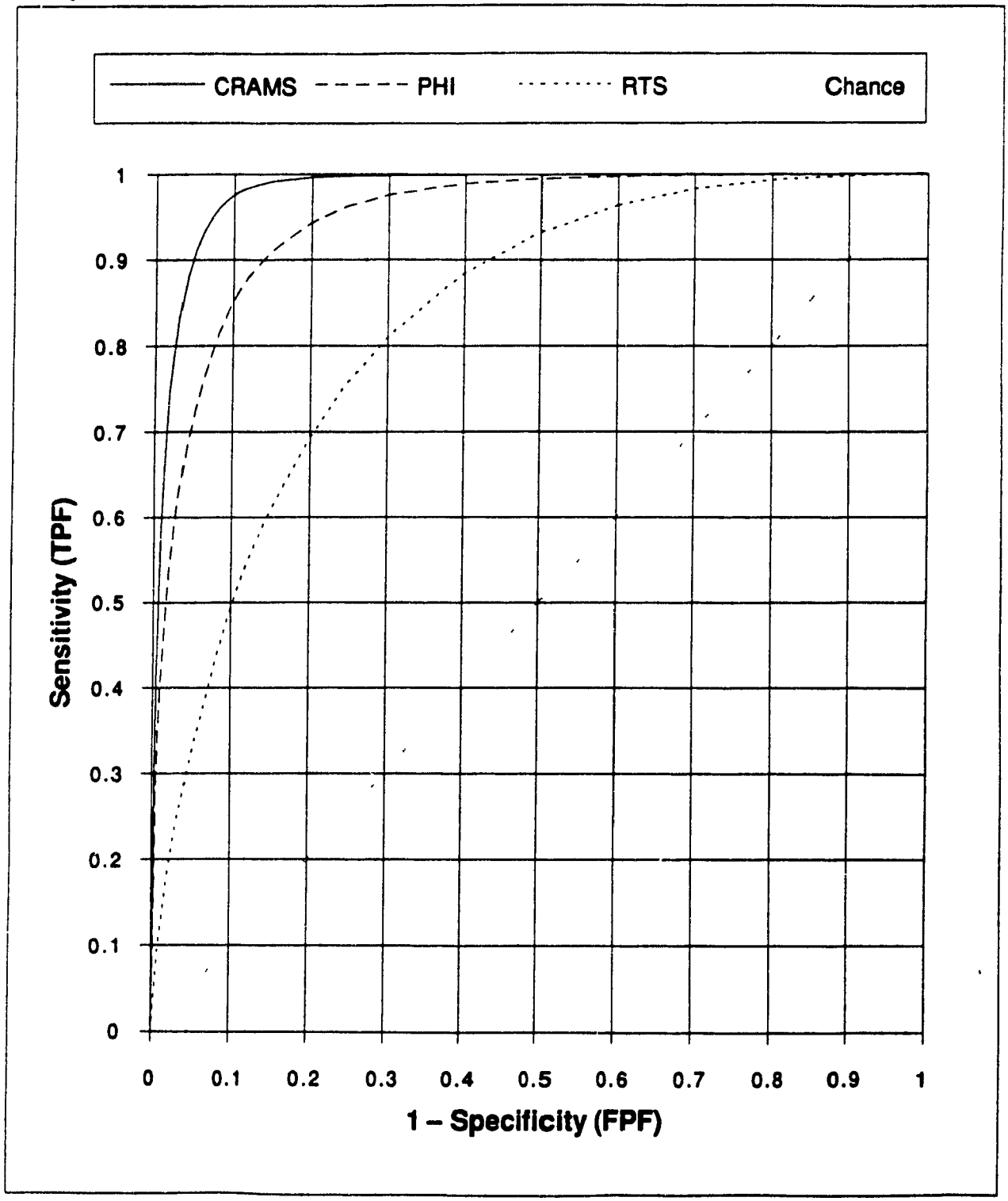


Figure 4.1c

Pre-hospital instruments:- Outcome: Mortality
Healthy extreme excluded

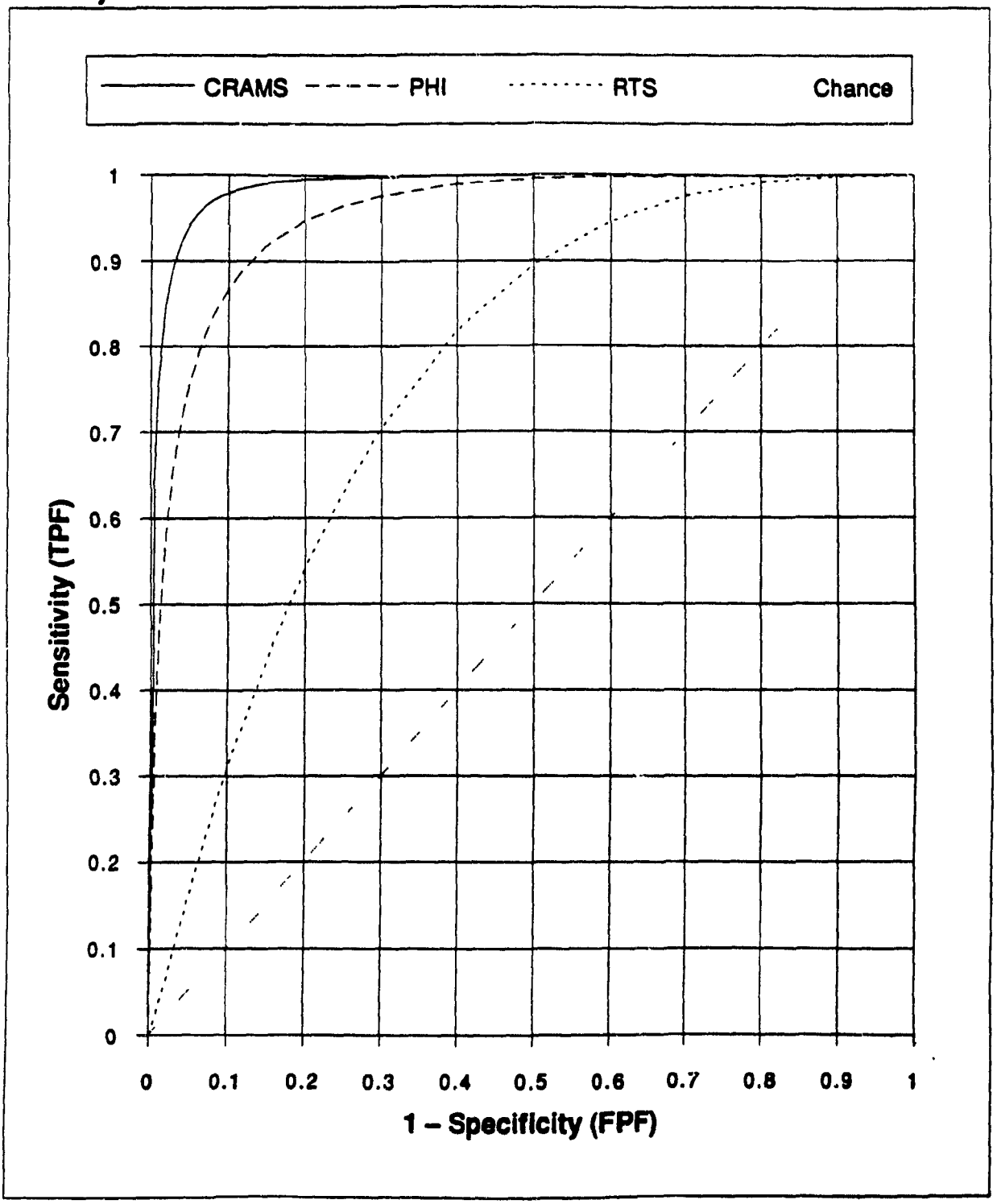


Table 4.9 demonstrates the differences in the area under the curve (A_{uc} diff) and the absolute performance gain in the different evaluations.

Table 4.9: COMPARISON OF CRAMS AND PHI WITH THE RTS (95% Confidence Intervals)

	All Cases	Both Extreme Excluded	Health Extreme Excluded
CRAMS			
A_{uc} diff	2.95 (1.9 - 4.0)	15.0 (11.0 - 18.8)	21.8 (18.4 - 25.1)
APG	5.9 (3.8 - 8.0)	29.9 (22.0 - 37.6)	43.5 (36.8 - 50.2)
PHI			
A_{uc} Diff	2.08 (0.9 - 3.2)	11.72 (7.5 - 15.9)	18.5 (14.8 - 22.2)
APG	4.16 (1.8 - 6.4)	23.4 (15.0 - 31.8)	36.9 (29.5 - 44.3)

b) **ROC analysis of the RTI and the RTS**

Delong's method was used to determine the A_{uc} of the RTS and RTI. This evaluation was performed within the same data set using mortality or major trauma as outcomes. Major trauma was defined by patient management categories (PMC) or the injury severity score. A tertiary injury was the PMC termed as major trauma and an ISS score of greater than 15 was also used to define major trauma. The difference in the area under the curve (A_{uc} diff) and the absolute performance gain (APG), (Table 4.10) were calculated using the RTS as the reference instrument.

Table 4.10: COMPARISON OF RTS AND RTI

Outcome	RTS A_{uc}	RTI A_{uc}	A_{uc} Diff	SE*	APG	Zstat
Mortality	.863	.886	.023	.018	.046	1.3
Tertiary Injury	.649	.746	.097	.0095	19.4	10.1
ISS > 15	.703	.878	.175	.0099	35.0	17.5

SE* is the Standard Error of the A_{uc} difference. The Standard Error of APG equals 2(SE).

In this study no difference was demonstrated between the RTS and RTI when mortality was the outcome assessed but when serious injury as defined by PMC or ISS was used as the outcome the RTI was superior.

A comparative analysis of pre-hospital instruments demonstrated that CRAMS was superior to the PHI which was superior to the RTS. When these instruments were evaluated using population subsets that excluded the obvious cases the superiority of CRAMS and the PHI over the RTS was accentuated. A comparison of the RTS with the RTI on the same data set demonstrated no significant differences when mortality was the outcome evaluated but when the less obvious end points of Tertiary Injury or ISS>15 were the outcomes studied, the performance of the RTI was markedly superior to the RTS.

Evaluation of the triage instruments using likelihood ratio and ROC analyses consistently demonstrated that the pre-hospital instrument restricted to physiological information, the RTS, was inferior.

B) Evaluation of the Measurement Qualities of Triage Instruments

The objective in the previous section was to evaluate the performance of several injury severity instruments. In this section the objective of the analyses is to evaluate the reasons for the differences in performance by evaluating three measurement qualities. The qualities to be evaluated are scaling mechanism, reliability and content validity.

a) Comparison of Interval and Ordinal Scaling Techniques.

The objective in this analysis was to determine if scaling mechanism effects predictive validity of a pre-hospital instrument. The predictive validity of the Triage RTS (t-RTS), an ordinal instrument, whose aggregate score is the arithmetic sum of its components, was compared to the RTS which is weighted by regression co-efficient's to produce an interval scale.

This evaluation was performed using the pre-admission values on 893 patients in the VGH data base and on 2959 patients in the PSU data base. Delong's calculation, which accounts for the paired nature of the evaluation, was used to determine the difference between the areas under the ROC curve for pre-hospital RTS and t-RTS values. End points that were evaluated were mortality for both the VGH and PSU data and ICU stay and Disability for the VGH data (Table 4.11).

Table 4.11: COMPARISON OF INTERVAL AND ORDINAL RTS PERFORMANCE

Outcome (%)	N	A_{uc} Diff	SE	P-Value
1. Mortality (6.7)	893	0.00025	0.0062	0.97
2. ICU (6.9)	893	0.0012	0.0054	0.82
3. Disability (17.0)	893	0.00013	0.0026	0.96
*4. Mortality (4.7)	2959	0.0014	0.0024	0.57

*PSU data (other evaluations are from VGH data)

There was no significant difference in A_{uc} when comparing the alternative formulations of the instruments over three different outcomes.

A second evaluation which used mortality as the outcome was performed on both the VGH and PSU data sets (Table 4.12). This was performed after exclusion of the most and least injured patients from the data sets, that is patients with a t-RTS of 1 or 12.

Table 4.12: COMPARISON OF INTERVAL AND ORDINAL FORMATS OF RTS (Extreme Cases Excluded)

Outcome (%)	N	A_{uc} Diff	SE	P-Value
1. Mortality (15.8)	158	0.38	1.9	0.84
*2. Mortality (18.0)	479	1.6	0.9	0.07

*PSU data (other evaluation is from VGH data)

In the second evaluation of interval vs ordinal scales no significant difference was noted although in the Pennsylvania data a small increase in A_{uc} of the interval scale over the ordinal scale approached statistical significance.

These evaluations of the ordinal and interval format of the RTS did not demonstrate any improvement in the performance of the interval scale over the ordinal scale.

b) **Determination of the Reliability of the RTS**

The reliability of an instrument affects its validity. The objective in this evaluation was to determine the reliability of the RTS.

The reliability of the RTS was evaluated on 367 subjects within the VGH data set, who had a pre-admission RTS determined first at the accident scene (70.6%) or the primary care facility (29.4%) and subsequently in the Vancouver General Hospital emergency department. The two measurements were compared using determinations of overall agreement, correlation co-efficient, plots of the differences against the mean of the two results and estimates of the intraclass correlations. The mortality in this population was 8.4% and the mean ISS was 16.6.

Although not an issue of reliability, it was noted that the predictive validity of the RTS determined at the two locations was not significantly different.

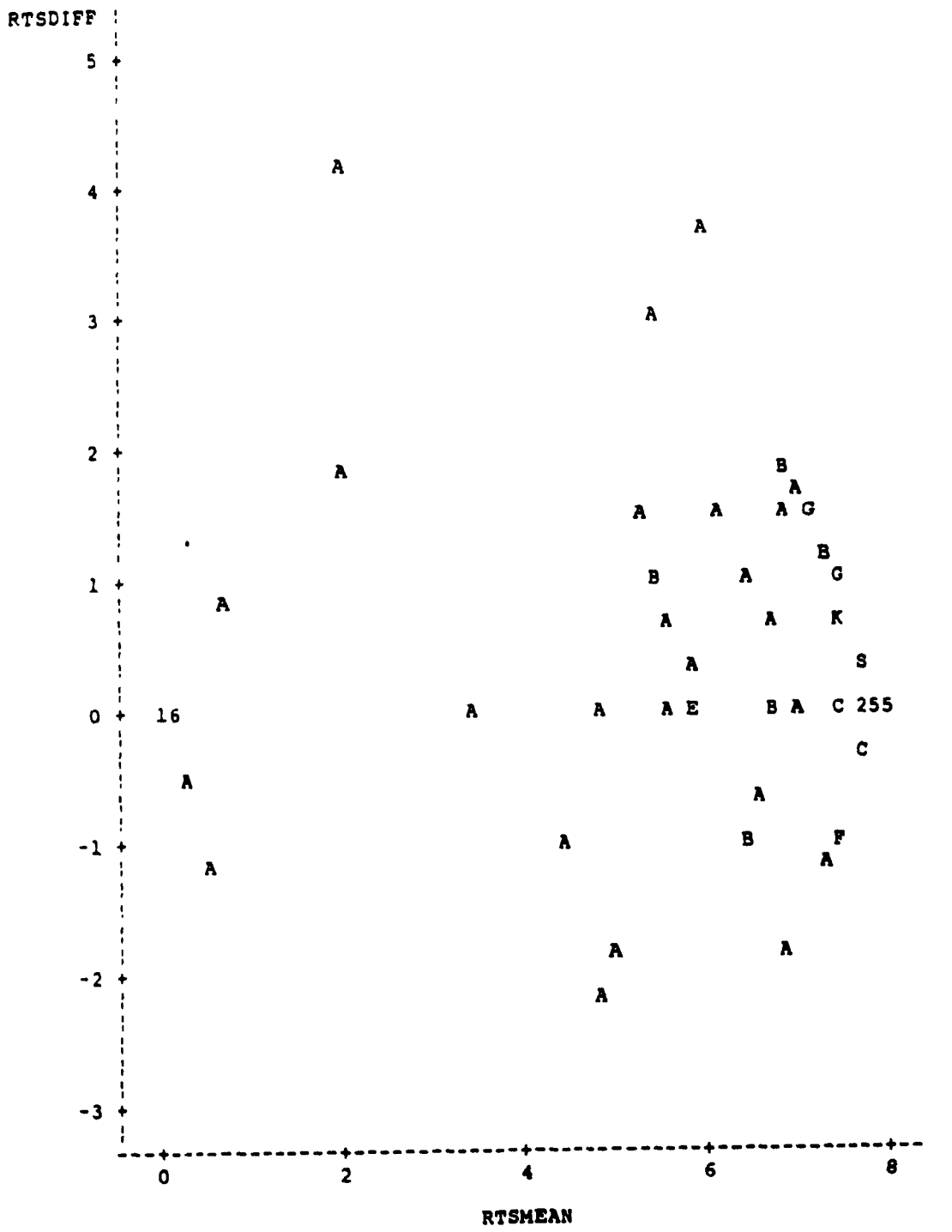
i) Inter-rater Agreement - The agreement level and two correlation coefficients were calculated and are presented in Table 4.13.

Table 4.13: INTER-RATER AGREEMENT AND CORRELATION OF RTS

Comparison	N	Result
Agreement	367	0.776
Pearson Correlation	367	0.952
Kendals Tau-b	367	0.620

ii) Plot of Difference vs Mean RTS [57] - The plot of the difference between each individuals two RTS values and the mean of those two values is presented in Figure 2. It is noted that the distribution of the mean results is bipolar with the largest category having a mean of 7.8408 (n=255) and the second largest category having a mean of 0 (n=16). It is noted that there is more scatter between the two extremes than near these two extremes.

Figure 4.2 RELIABLITY PLOT



iii) Intraclass Correlation - The intraclass correlation was determined from an ANOVA model which included terms for the patient and the raters. The reliability was calculated with a random effects model and the confidence intervals for the reliability were calculated using Satherthwaites approximation.[125,126] The ICC was determined for all patients and also for a subset of patients which excluded those with the two highest scores (7.8404 & 7.55) and the two lowest scores (0 & .2908). This was done to evaluate the ICC of patients with intermediate RTS values. The ANOVA Tables for the evaluation of all 367 patients and for the intermediate subset of 73 patients are presented below (Tables 4.14 and 4.15).

Table 4.14: TWO WAY ANOVA FOR ALL PATIENTS (N = 367)

Source	DF	Mean Square
Patient	366	6.5265*
Rater	1	2.3776*
Error	366	0.1557
Total	733	
Mean RTS 7.18		* p <.0001

Table 4.15: TWO WAY ANOVA FOR INTERMEDIATE SUBSET OF PATIENTS (N=73)

Source	DF	Mean Square
Patient	72	4.6005*
Rater	1	9.737*
Error	72	0.6741
Total	145	
Mean RTS 6.43		* p <.0001

As is noted in Table 4.16 the ICC for all patients was high but when the obvious cases were removed the reliability dropped from 0.95 to 0.71.

Table 4.16: INTRA-CLASS CORRELATION

Population	N	R	(95% CI)	Mortality	Mean ISS
All Patients	367	.95	(.943 - .961)	8.4%	16.6
Intermediate Subset	73	.71	(.593 - .803)	12.3%	25.1

c) **Content Validity of Instruments with Non-physiologic Domains**

The objective in this analysis was to determine which attributes improved the content validity of a physiologic instrument. In this thesis attributes that increase the predictive validity of an instrument are said to increase the relevant information content or content validity of that instrument. Content validity was evaluated by adding non-physiologic attributes to the RTS which was used as the source of physiologic information. The only outcome evaluated was mortality. The attributes studied were age, anatomic information, mechanism of injury and co-morbidity. Evaluation of the additional attributes was performed with logistic regression and ROC analyses. These attributes were all coded as dichotomous (0,1) variables unless indicated otherwise. In the tables which present logistic models all the coefficients are statistically significant unless indicated otherwise.

Models which were significant in the logistic regression evaluation were tested for statistically significant changes in the A_{uc} using Delong's technique. Point estimates of the A_{uc} are calculated by Proc Logistic. The Score statistic allows comparison of the goodness of fit of different models after correction for the number of variables in each model. The higher the Score statistic the better the fit of the model. Both the VGH and PSU data were used for the analysis of

content validity. In addition, likelihood ratio and ROC analysis of published data from San Jose were used to evaluate mechanism of injury information.

i) Demographic Information (Age)

A dichotomous (0,1) age category of below 55 and 55 and above was evaluated alone and with the RTS. Other age categories were evaluated but did not provide further insight. This evaluation was performed using the VGH data.

Table 4.17:
AGE CATEGORY CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS — Outcome: Mortality
(Coefficients and Statistics from Logistic Regression Models)

	RTS	Age	Score Statistic	A_{uc}
1.		0.948	19.6	.607
2.	-0.93		448.1	.839
3.	-1.01	1.85	452.3	.906

The p values for all the coefficients equalled 0.0001. The differences between the A_{uc} values were significant ($p=0.002$).

These studies showed that age alone would not perform well as an instrument of injury severity but that an instrument that included the RTS and age had a better performance than the RTS alone.

A logistic regression evaluation was also performed to determine if the performance of the RTS was different for young (age category = 0) or old patients (age category =1). This was done by evaluating the statistical

significance of a term, which is the product of age and the RTS, into the logistic regression model. Such a product term is called an interaction term. An interaction term should have biological meaning and statistical significance before it is included in a regression model.[127] In this case there was no significant age/RTS interaction.

ii) Mechanism of Injury (MOI) Information

San Jose Study - Re-analysis: 500 consecutive patients seen at San Jose Hospital had TS, CRAMS, MOI, ISS, and mortality recorded. Patients were classified as seriously injured if they died, were hospitalized for more than three days, had a TS of ≤14 or an ISS > 15.

206 individuals were classified as seriously injured. The TS and CRAMS were evaluated alone and in combination with the MOI information. Examples of the MOI information used by these authors included motorcycle crash and vehicular intrusion into the passenger space. The authors concluded that the use of MOI information with either CRAMS or the TS, fortified the instrument. To facilitate evaluation of these conclusions likelihood ratios and ROC curves were determined from their data and are presented in Table 4.18.

Table 4.18: LIKELIHOOD RATIO EVALUATION OF MECHANISM OF INJURY

Criterion	Sensitivity	Specificity	LR+	LR-
MOI	.90	.10	1.0	1.0
TS	.45	.94	7.5	0.585
TS + MOI	.75	.40	1.25	0.625
CRAMS ≤8	.66	.82	3.66	0.41
CRAMS + MOI	.93	.30	1.32	0.233

The addition of MOI information changed the sensitivity and specificity of the TS and CRAMS but the positive likelihood ratios (LR+) are much smaller for the combined instruments.

The specificity of each scale can be evaluated at any level of sensitivity using ROC curves. ROC curves were calculated for CRAMS and the TS alone and the specificity of the instruments were determined at the sensitivity the authors had found after the addition of MOI information. This allows direct comparison of the sensitivities obtained with the addition of MOI as noted in the table above with the sensitivities noted in the ROC curves for the TS and CRAMS without MOI data.

Table 4.19: COMPARISON OF SPECIFICITY AT DEFINED SENSITIVITY

Instrument(s)	Sensitivity	Specificity
TS + MOI	75	40
TS (ROC)	75	50
CRAMS + MOI	90	30
CRAMS (ROC)	90	33

ROC evaluation (Table 4.19) did not show an improvement in the performance of CRAMS of the TS with the addition of the MOI information used in this study.

Injury Class Information: The mechanism of injury has been categorized into different classes of injury within the VGH trauma registry. These classes are blunt, penetrating and thermal injury. Injury class was evaluated in the VGH data base using logistic regression to evaluate possible improvement in the predictive validity of the RTS with the addition of this MOI data. The pre-admission RTS was supplemented with the admission RTS when this data was missing to result in a population of 1043. The possibility of an interaction between age and injury class was investigated by placing an interaction term in one of the regression models (data not shown). The coefficients of the model were evaluated for statistical significance and the models which included the injury class were compared with the RTS alone (Table 4.20) and the RTS with age (Table 4.21). Comparisons were made on the basis of the Score statistic and A_{uc}.

Table 4.20 presents logistic regression models with the injury classes alone, the RTS alone and then injury class and RTS in a full model. The only class of injury that had a significant coefficient when the RTS was included was blunt trauma.

Table 4.20: INJURY CLASS CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS — Outcome: Mortality
(Coefficients and Statistics from Logistic Regression Models)

RTS	B	P	T	Score Statistic	A_{uc}
	-1.1	-2.4	-0.2**	33.7	66.2
-0.93				448.1	83.9
-1.04	0.88	-2.4*	1.1**	455.5	89.3

* .05 < p < 0.1 ** p > 0.1

Table 4.21 presents logistic regression models which include the RTS and age, as a contrast to a model which includes class of injury, age and the RTS. None of the injury class variables were significant when age was included in the model.

Table 4.21: INJURY CLASS CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS and AGE — Outcome: Mortality
(Coefficients and Statistics from Logistic Regression Models)

RTS	AGE	B	P	T	Score Statistic	A_{uc}
-1.01	1.85				452.3	90.6
-1.11	1.75	0.5**	-2.7*	1.18*	470.6	92.1

* 0.05 < p < 0.1 ** p > 0.1

In another model (data not shown) the interaction of age and injury class was not significant.

This evidence suggests that injury class information does not significantly improve the performance of the RTS.

Energy Impact Information: In another evaluation of mechanism of injury the VGH data were reorganized according to the energy of impact associated with the type of injury.[128] Injuries related to falls from greater than 6 meters or to firearms were designated as high energy impact (HE); lacerations, low falls and stab wounds were designated as low energy impact (LE). Thermal injuries (T) were left unchanged and undesignated injuries were the reference category. The components of the blunt and penetrating mechanisms were reclassified as low energy (LE) and high energy (HE) injuries. These covariates were evaluated alone and with the RTS and AGE. These comparisons were done in the subset of patients who had their RTS determined in the pre-hospital situation. A model which included an age/low energy interaction term was also evaluated.

In an evaluation of a model with only the energy intensity covariates, LE was highly significant ($p = .0001$), HE approaches significance ($p = .06$) and T was not significant (data not shown). In Table 4.22 the evaluation of the contribution of the energy impact information to the RTS is presented. In this situation HE was significant ($p = .02$) and LE approached significance ($p = .06$).

Table 4.22: ENERGY IMPACT CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS — Outcome: Mortality
Coefficients and Statistics from Logistic Models

RTS	LE	HE	T	Score	A_{uc}
1. -0.93				448.1	83.9
2. -0.96	0.40**	1.6	1.0**	453	87.8

* $0.05 < p < 0.1$ ** $p > 0.1$

In Table 4.23 the evaluation of age, RTS and energy intensity is presented. With the addition of age, LE is no longer significant ($p = .29$) however when an interaction term for low energy and age is included in the model all the terms are significant at the .05 level and the interaction term has a p value of .008. This evaluation suggests that patients in the young age category who have a low energy impact injury have a lower mortality than average and that the old age category patients with a low energy impact injury have a higher than average mortality. Thus low energy impact injuries can only be included in this model when an age/low impact interaction term is also included.

Table 4.23: ENERGY IMPACT AND AGE CONTRIBUTION
Reference Instrument: RTS and AGE — Outcome: Mortality
Coefficients and Statistics from Logistic Models

RTS	Age	LE	HE	T	Age*Le	Score	A_{uc}
-0.01	1.85					452.3	90.6
-1.0	2.0	-0.03**	1.92	1.0		472	91.6
-1.1	1.0	-2.3	1.8	1.2	3.3	473	92.4

* 0.05 p < 0.1 ** p > 0.1

This fourth model which includes high and low energy impact, RTS, age category and an interaction term for age and low energy impact has the highest score statistic and the largest A_{uc}. It is noted that the interaction of low intensity injury with age, makes clinical sense.[92] This regression model was compared with the RTS alone and with the model which included age and RTS (Table 4.24). The data presented show that age and energy impact improve the performance of the RTS by over 20%.

Table 4.24: PERFORMANCE GAIN WITH ENERGY IMPACT INFORMATION

Reference	N	Mortality	A_{uc} Diff	SE*	APG*	P-Value
EI|RTS	893	.066	10.3	2.6	20.6	0.0001
EI|Age, RTS	893	.066	2.2	0.95	4.4	0.018

Analysis of the data from Pennsylvania confirmed the significant contribution of age and high energy mechanisms (firearms) to the basic physiological instrument. Unlike the Vancouver data, the Pennsylvania data did not allow the separation of the falls into high energy and low energy

etiologies thus precluding a second evaluation of the age/low energy impact interaction.

iii) Anatomic Information

ISS - The ISS score was used as an example of anatomic information to add to the physiological information of the RTS. The ISS is not available in the pre-hospital situation so this form of anatomic information would not be of practical value but could give an indication of the usefulness of similar information.

Table 4.25: ISS CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS — Outcome: Mortality
Coefficients and Statistics from Logistic Models

ISS	RTS	Score	A_z
0.08		273.6	82.1
	-0.89	368.2	83.6
0.04	-0.71	373.6	83.7

All the coefficients noted in Table 4.25 are statistically significant ($p < 0.001$); however the increase in the A_{uc} associated with these values was not statistically significant ($p > 0.1$). In this study the ISS did not improve the predictive validity of the RTS.

Models which included the ISS, age and the RTS were evaluated.

In Table 4.26 the coefficients for these terms are statistically significant but the differences in the A_{uc} were not. The interaction term of age and ISS was not significant and the Auc did not change when this term was included.

Table 4.26: ISS CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS and AGE — Outcome: Mortality
Coefficients and Statistics from Logistic Models

ISS	RTS	Age	Age/ISS	Score	A_{uc}
	-0.97	1.68		371.2	89.3
0.05	-0.78	1.94		391.0	91.4
0.06	-0.78	2.4	.67*	394.1	91.4
*p = .41					

Another assessment was performed using the ISS as a dichotomous variable with the demarcation value being an ISS of 15. As is shown in Table 4.27 an interaction between the ISS score was significant in the model without RTS but the A_{uc} estimate was unchanged. When the model included the RTS and the age/ISS interaction term neither the interaction nor the A_{uc} difference was statistically significant.

Table 4.27: ISS CATEGORY CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS and AGE — Outcome: Mortality
Coefficients and Statistics from Logistic Models

ISS > 15	RTS	Age	Age* ISS < 15	Score	A_z
4.7		0.88	2.57	153.2	83.8
	-0.89			368.2	83.6
	-0.97	1.68		371.2	89.3
2.5	-0.8	1.46	1.73**	376.4	91.2

** $p > 0.1$

Serious Injuries - The number of serious injuries was recorded in the PSU data base. This anatomical information is available in the prehospital situation and is therefore relevant. The number of serious injuries, which ranged from zero to ten was recategorized as "none", "one" or "multiple" and coded as 0, 1 and 2. Serious Injuries (SI) were evaluated alone and as a means of adding anatomical information to the RTS and the RTS and Age (Table 4.28). As is shown in Table 4.28 Serious injury data added significantly to a model with RTS and a model with RTS and age.

Table 4.28: SERIOUS INJURY CONTRIBUTION TO INSTRUMENT PERFORMANCE
Reference Instrument: RTS and AGE Outcome: Mortality
Coefficients and Statistics from Logistic Models

	RTS	SI	Age	Score	A_z
1.	-0.98			1135	86.3
2.	-0.94	.54		1135	89.2
3.	-1.17		2.4	1186	92.3
4.	-1.13	.71	2.5	1188	94.1

In Table 4.30 the evaluation in the differences in the A_{uc} of the models was tested comparing the different variables conditional (|) on the presence of the other variables. It is shown that serious injury increased the performance of the RTS alone, and that a model that included the RTS, age and serious injury performed significantly better than models that excluded any of these variables. This final model had an absolute performance gain of 15.6%.

Table 4.29: PERFORMANCE GAIN WITH SERIOUS INJURY INFORMATION
Outcome: Mortality

Comparison	N	Mortality	A_{uc} Diff*	SE*	APG*	P-Value
SI|RTS	2964	4.7%	2.9	1.1	5.8	.001
Age|RTS,SI	2964	4.7%	4.9	1.4	9.8	.0003
SI|RTS,AGE	2964	4.7%	1.8	0.6	3.6	.0002
SI,AGE|RTS	2964	4.7%	7.8	1.5	15.6	.00000008

*A_{uc} Diff - Difference from Reference Instrument
SE - Standard Error of A_{uc} Standard Error of APG = 2(SE)
*APG - Absolute Performance Gain

iv) Co-morbidity

The presence or absence of 7 previous disease conditions was recorded in the PSU data. These were recategorized to the presence or absence of any previous disease.

The 95% confidence intervals for A_Z for comorbidity alone was marginally above the null of 50% (Table 4.30). But co-morbidity did add to the A_{uc} of the RTS alone. As is shown in Table 4.31 this association approached statistical significance. However when adjusted for age the coefficient for co-morbidity was not a significant contributor to the model. There was no significant interaction between age and co-morbidity (data not shown).

Table 4.30: CO-MORBIDITY CONTRIBUTION TO INSTRUMENT PERFORMANCE

RTS	Age	Co-morbidity	Score	A_{uc}
1.		.123	1151.7	50.7 - 57.5
2. -0.98			1134.0	86.3
3. -1.03		.329	1151.7	89.0
4. -1.17	2.4		1186.0	92.3
5. -1.17	2.3	.930**	1187.7	92.4

** p = .335

Table 4.31: CO-MORBIDITY CONTRIBUTION CONDITIONAL ON RTS

N	Mortality	Auc Diff	SE	ZStat	P-value
2964	4.7%	2.7	1.3	-2.2	.056

SUMMARY

This chapter includes an ROC evaluation of the most widely used measure of outcome the Injury Severity Score and comparative evaluations of four triage instruments' ability to predict mortality and serious injury. The importance of scaling techniques, reliability and content validity to predictive validity were studied and the results presented.

CHAPTER 5

DISCUSSION

In 1980 a panel of 30 experts used the delta technique to rank different qualities used in evaluating injury severity instruments.[5,9] The qualities in the order of importance were predictive, construct and face validity, inter-rater reliability, data availability, separation of illness severity and care quality, and simplicity.

The predictive validity of trauma instruments is the most frequent quality evaluated, construct validation studies have been more limited, (unless one considers the prediction of ISS values as construct validation), and issues of face validity and content validity have received little attention.

Feinstein refers to face validity and content validity as issues of sensibility and states that they are evaluated qualitatively rather than quantitatively, as is done in other measures of validity.[48] However, in this study content validity was evaluated quantitatively by determining the effect of an instrument's information content on predictive validity.

Ranked after the validity qualities is inter-rater reliability which is important because the level of reliability effects predictive validity.

Simplicity and data availability are ranked at the lowest end of the spectrum. These qualities are in opposition to the issues of content validity. The simpler the instrument and the less data determined to be available, the lower the content and face validity of an instrument.

The lack of uniform statistical definitions in the evaluation of predictive validity has been a problem with many trauma studies. This study demonstrates that likelihood

ratio and ROC analyses of instruments for the measurement of injury severity have the advantage of allowing direct comparison of instruments, give more meaningful results than accuracy and triage rates and are less cumbersome than the measures of specificity and sensitivity. ROC analysis comprehensively evaluates an instrument's performance as this type of analysis does not depend on a single cutoff point (see Chapter One).

The sources of data used in this thesis were the trauma literature, the Vancouver General Hospital trauma registry and the Pennsylvania State University trauma registry. The literature data had the advantage of permitting the evaluation of instruments not included in the trauma registries but did not allow multivariate analysis or evaluation of instruments within the same population. The trauma registries which are both hospital based do not have a defined population base. Missing data reduced the sample size in the VGH trauma registry, which was in its first year of operation. The PSU trauma registry did not have this problem. The trauma registries permitted paired evaluation of instruments within the same data sets which results in unbiased comparisons. Different information was collected in the two registries which permitted evaluation of different issues but usually did not allow cross validation of results.

I. INSTRUMENT PERFORMANCE

The study by Baxt et al. is the only study prior to this study that prospectively evaluated triage instruments using ROC curve analysis.[44] In that study no difference could be demonstrated between CRAMS, PHI and the RTS in predicting either mortality or disability. The authors concluded that the performance of these instruments is poor and that there is insufficient information in the pre-hospital situation to improve their predictive validity. Moreover they felt that the instruments

contributed little, as the prediction of mortality in most cases was obvious. This is an important study because if its conclusions are accepted then the present efforts in measurement of injury severity in the pre-hospital situation should be abandoned.

In this thesis likelihood ratio and ROC curve analyses were used to evaluate the predictive validity of conventional trauma severity instruments and determine whether scaling techniques, inter-rater reliability or content validity are important in attaining high predictive validity.

ROC curve analysis was used to evaluate the ISS using MTOS data on 14,876 trauma victims (see Table 4.1). Performance of the ISS in the entire population (A_{uc} = .894 ± 0.005) was not different from its performance in any of the sub-categories based on trauma mechanism or age. The A_{uc} for penetrating trauma was not statistically different from blunt trauma which indicates that changes made to the AIS-1985, because of deficiencies in evaluation of penetrating trauma, have equalized the predictive validity of the ISS for both blunt and penetrating trauma. It was surprising, that the ISS, which is calculated with the benefit of hindsight does not predict mortality better than this evaluation demonstrated.

Likelihood ratios (LR) were used to compare the published data on four triage instruments (see Table 4.2). The studies analyzed were all large (N > 2000) and were evaluated at the cutoff points recommended by each instrument's proponents. The negative likelihood ratio confidence intervals overlapped, so that it was not possible to distinguish the performance of these four instruments on this basis. In contrast, there was a dramatic gradation in the positive likelihood ratios with CRAMS, PHI and RTI all being superior to the RTS. The RTS is the most widely used triage instrument and is the only one of these four instruments which is limited to

physiologic information.[88] When post-test probabilities (Tables 4.3 & 4.4) are determined the differences are not as dramatic. This is a reflection of the differences in the mortality prevalence in the underlying population. Evaluations that depend on disease frequency have the potential of obscuring the actual performance of the instruments.

ROC evaluation was performed on three of these instruments (Tables 4.5 & 4.6 & Figure 4.1a,b,c). Using mortality as an endpoint it was noted that CRAMS was superior to PHI which was better than the RTS. The magnitude of this statistically significant difference was not large and all instruments had an A_{uc} above 95%. This indicated that all of the instruments performed well when mortality was used as an outcome.

The ratings given by the instruments themselves were used to identify the individuals with obvious outcomes. When the obvious minor and grave injuries were excluded it was noted that CRAMS and PHI maintained there predictive validity above 95% (Table 4.7 & Figure 4.1b) but that the performance of the RTS dropped (A_{uc} = 83% ± 2%). This difference in instrument performance was accentuated when only the healthy extreme was excluded (Table 4.8 and Figure 4.1c). Table 4.9 summarizes the ROC studies which show that there is a small but significant difference for CRAMS and PHI over the RTS, when all observations are included but when the obvious extremes of health and injury are excluded this difference is markedly increased. The absolute performance difference between CRAMS and the RTS when the healthy extreme is excluded was 43.5% (Table 4.9) It could be argued that using an instrument to exclude the obvious cases would bias the instrument against itself. Examination of the mortality rates in these populations and their subsets shows that the mortality increased by less than 2 fold for CRAMS and PHI and almost 3.5 times for the RTS. One would expect that if the RTS was biased against itself, that the

mortality increase would be less than the increase noted for the other two instruments, rather than the contrary. The evaluation of the subsets demonstrated that the small significant difference noted in the unrestricted populations reached an important magnitude in the "non-obvious" subset.

The likelihood and ROC curve analysis do not confirm the findings of Baxt et al. that all instruments are essentially the same, rather they support the concept that non-physiological information improves an instrument's performance.

Comparison of the RTS and RTI using Pennsylvania State University data demonstrated no statistical difference between these two instruments ability to predict mortality (Table 4.10). However when major trauma, as defined by patient management category or by the injury severity score, was used as an outcome the absolute performance gain of the Revised Trauma Index over the Revised Trauma Score for the respective outcomes was 19% and 34%. Both the RTI and RTS are better predictors of mortality than the less obvious outcome, major trauma. However it was clear that when a less obvious outcome is used, the instrument which is not limited to physiologic information performed better.

The differences in predictive validity between the triage instruments that have been presented here contradict some aspects of Baxt's study of pre-hospital triage instruments.[44] This thesis supports the idea that instruments have similar abilities when predicting mortality in an unrestricted population; however when the easy cases were excluded or outcomes that are more difficult to predict were utilized the instruments with non-physiologic information had a superior performance. As the Revised Trauma Score is the most widely used instrument, it is generally assumed that it must be the superior instrument. Analyses presented in this study contradict

this assumption and therefore it is concluded that the widespread adoption of the RTS is premature.

II. MEASUREMENT QUALITIES

Further investigation was performed to ascertain whether scaling mechanism was important, what the reliability of the RTS was, and what information in addition to physiologic information improves the predictive validity of a physiological instrument.

Scaling Mechanisms

The RTS has both an ordinal and an interval format. The triage-RTS is aggregated by adding the coded values of the Glasgow coma scale, systolic blood pressure and respiratory rate resulting in an ordinal scale. The RTS, an interval scale, is obtained by adding the values of these attributes after they have been weighted by coefficients determined by logistic regression.

Comparison of these two formats of the same attributes on the same data permits assessment of the importance of scaling mechanisms. This evaluation used mortality, ICU admission and disability as an outcome in the VGH data and mortality in the PSU data.

In the evaluations using an unrestricted population the difference in performance of the two formats was insignificant (Table 4.11). When the extremes of minor and major injury were excluded no significant difference was noted in either the VGH or the PSU data however the performance of the interval scale in the PSU data which was higher by 1.6% did approach significance ($p = 0.07$) (Table 4.12). These results

show that for the overall population the interval format of the RTS is not an improvement over the ordinal scale. The PSU data suggested that for the non-obvious patients, the interval scale may improve predictive validity, but this trend was not replicated in the VGH data set. No evidence has been presented which confirms the importance of interval scaling on the predictive validity of triage instruments; the ordinal format of the RTS functions as well as the interval format. Therefore it is concluded that ordinal scales are not necessarily at a disadvantage when compared to interval scales.

a) **Reliability**

Evaluation of the reliability of the RTS was performed comparing the pre-admission to the admission results which were included in the VGH data (Table 4.13). Overall agreement and two correlation coefficients were calculated. The overall agreement was .776 (± 0.022) and the Pearson correlation coefficient was 0.95. Pearson correlation was high but it must be emphasized that this is a measure of linear association not agreement.

The 95% confidence intervals for Kendals tau-b (0.53-.71) for the RTS overlaps the value of 0.55 published for its predecessor, the Trauma Score. This measure of association evaluates the concordance of the ranking of two measurements not the agreement or the reliability of an instrument. Plotting the mean of the measurement against the differences allows one to visualize the trends in reliability at different measurement levels and also to visualize the strength of agreement. It is clear from the plot (Figure 4.2) of these studies that the RTS is most reliable at the two extremes and that there is more scatter at intermediate levels.

The intraclass correlation coefficient (ICC) which is a measure of agreement evaluates the similarity of pairs of measurements on the same individual The ICC was determined for all patients and also for the subset of patients which excluded the extremes of health and injury (Table 4.16). The ICC of 0.95 dropped to 0.71 when only the intermediate subset was evaluated. The upper limit of the validity of an instrument is determined by the square root of its reliability.[129] This would reduce the maximum validity or the RTS from .97 to .84 for the non-obvious cases. It is noted in the anova tables (Tables 4.14 & 4.15) that the coefficient denoted as rater was significant. This demonstrated that there was a significant difference in whether the RTS was a pre-admission or the admission determination. As noted in the chapter 4 there was a trend for the RTS determined in the emergency department to be higher than the pre-admission RTS. The A_{uc} for the admission results was 2.8% higher but the 95% confidence intervals for the difference crossed zero. The ICC for the PHI, CRAMS and RTI have not been determined in any study. It is likely that reliability decreases for all triage instruments when the less obvious cases are evaluated.

b) **Content Validity**

The predictive validity is recognized as the most important quality of instruments for illness severity. The major difference between the RTS and the other instruments which were shown to have superior predictive validity in several analyses presented in this thesis is the information content of the instruments. If it is a lack of content validity, that limits the predictive validity of the RTS, then the addition of non-physiologic information to this instrument should improve the predictive validity and also determine what information would be useful in a new instrument. Logistic regression models were used as

means of adding information content to the basic physiologic information in the RTS. The attributes assessed were age, mechanism of injury, anatomy and co-morbidity. The null for ROC curve evaluation is an A_{uc} of 50%; therefore the absolute performance gain of the information added to the RTS will be used to evaluate instrument performance.

i) Age

The first attribute evaluated, age, is not included in any of the injury severity instruments. As is shown in Table 4.17 age significantly improves the A_{uc} of the RTS by 6.7 %.The absolute performance gain of 13.4% for age was the largest gain noted for any single attribute added to the RTS.

ii) Mechanism of Injury (MOI)

MOI information has been touted in the literature as a means of fortifying other instruments. A paper originating in San Jose which expressed such an opinion was re-evaluated by calculating Likelihood ratios and ROC curves from the published data. The LR+ and the LR- values for the four mechanisms of injury that the authors added to CRAMS and the TS were both at the null of 1. When they added this information to these instruments the result on the overall likelihood was a reduction of the performance of these instruments at the cutoff recommended (TABLE 4.18). ROC curve analysis allowed determination of the specificity associated with the sensitivities resulting from the use of MOI information (Table 4.19). It was shown that the use of this information changed the cutoff point resulting in a higher sensitivity with a lower specificity but the predictive validity of the instrument was not improved. The conclusion of the authors that MOI information had fortified the instruments is contradicted in this evaluation.

Class of injury which had been categorized into the traditional blunt and penetrating trauma was also evaluated using the VGH data. Penetrating trauma tended to decrease mortality and blunt trauma tended to increase mortality, which was the opposite of what would be expected. When age was added to the model none of the injury class variables were significant(Table 4.21). It was concluded that these classifications were not helpful in predicting the risk of death from injury.

Re-evaluation of the same data after they were reorganized according to the energy impact associated with the injury was performed using the VGH data and the PSU data when applicable. Injuries caused by firearms and falls of over 6 meters were categorized as high impact injuries and low falls, lacerations and stab wounds were categorized as low impact injuries. Not all observations in the VGH data set could be classified as high or low impact injury as this information was not available on all, therefore analysis was performed on the whole data set using dummy variables, which credited undesignated patients only with the RTS value. This would likely result in an underestimate of both the magnitude and the significance of a MOI coefficient. Nevertheless it was demonstrated that high impact injuries were both significant and that the absolute performance gain was 7.9% (Auc diff = 3.95% Table 4.23).

In the VGH data low impact injuries which included lacerations, stab wounds and falls from less than 6 meters, interacted strongly with age. Models that included evaluation of low impact injuries without accounting for the age interaction showed no statistical significance for low impact injuries (Table 4.23); however when an age/low impact interaction term was included in the model (Table 4.24) low impact and the age/low impact

interaction were significant. Low impact injuries are associated with a decreased mortality in individuals under 50 and with an increased mortality for those over 50 years of age. This finding is in agreement with other trauma studies which document the poorer prognosis in the elderly.[130] Evaluation of a model with high and low impact of injury and age showed an absolute performance gain of 20.6% over the RTS alone. Energy of impact information showed an absolute performance gain of 4.4% over age and RTS. The PSU data confirmed the importance of high impact injuries but was not organized in a way that allowed evaluation of the age/low impact interaction. Because an estimate of the type of impact was not available on many subjects in the VGH data the magnitude of this effect may be underestimated but this study does demonstrate that injury impact information improves the predictive validity of the model and when age and an age/low impact interaction term are included, the instrument performance is improved by over 20%.

iii) Anatomy

Using the VGH data the importance of anatomic content was evaluated using the ISS as a source of anatomic information. The number of severe injuries which was recorded in the PSU data was categorized as none, one or multiple severe injuries and used as another means of evaluating the importance of anatomic content. The ISS is ascertained by chart review and therefore this information is not available in the pre-hospital situation; the number of severe injuries was also collected after admission but it would seem reasonable that this information could be determined by a paramedic in the pre-hospital situation.

In the first evaluation the ISS was used as the source of anatomic information and mortality was the outcome used. The predictive validity of the ISS and the RTS alone were similar and there was no increase in A_{uc} when these two instruments were combined (Table 4.26). The addition of the ISS to the model which contained age and RTS did not significantly improve the A_{uc} and no significant interaction was noted between age and the ISS. (Table 4.27). An interaction of age and the ISS was sought after dichotomizing the ISS below and above 15 but again the interaction of age and ISS was not significant (Table 4.28).

The number of severe injuries was also used as a means of adding anatomic information to the RTS. The data were available in integers from 0 to 10 but were recategorized to 0, 1 and multiple.

This was done because the frequencies became low for categories higher than one and it was felt that this type of information could be ascertained in the pre-hospital situation. Table 4.29 shows that the Severe Injury category increase A_{uc} over the RTS and also increases the A_{uc} over RTS and age. Table 4.30 shows that these differences are all highly significant and that age and severe injury together increase A^{uc} by 7.8%. The use of severe injury with age resulted in an absolute performance gain of more than 15%.

iv) Co-Morbidity

The PSU data records an indication of pre-existing disease as follows: vascular, respiratory, renal, diabetes, therapy, unspecified and spleen. These categories were amalgamated into one category, the presence of absence of preexisting disease. The A_{uc} for these disease alone (50.7% - 57.5%) was

very close to the null of 50%. The increase in the A_{uc} of the RTS alone when co-morbidity information was added approached statistical significance but when the RTS was adjusted for age no difference was detected. Co-morbidity as recorded here was strongly correlated with age which is associated with both the risk of co-morbidity and with an increased mortality rate. In the APACHE scoring system, it was found that pre-existing disease in itself did not contribute to an increased mortality but that pre-existing organ failure did. In the Apache II instrument chronic disease is only included as a risk factor if it is associated with organ failure. It is likely if more specific information was available that organ failure would have contributed to mortality but co-morbidity as recorded in the PSU data is confounded by age. This evaluation supports the inclusion of information concerning age in pre-hospital triage instruments.

SUMMARY

The measurement of injury severity requires attention to the principles of measurement theory. It has been stated that pre-hospital measurement is limited by the data available in that situation and that existing pre-hospital instruments are equivalent. This study found that pre-hospital instruments had an A_{uc} at least as high as the ISS which is determined retrospectively. It was shown that triage instruments which are not restricted to physiologic information have a better predictive validity than the RTS which only has physiologic information. This was most clearly demonstrated when non-obvious outcomes were evaluated. Interval scales were not shown to have an advantage over ordinal scales and it was shown that the reliability of the RTS was reduced when the obvious cases of extreme injury or minor injury were excluded from the analysis.

Investigation of the information content that would increase the predictive validity of a physiologic instrument showed that age was the single attribute that increased instrument performance by the largest magnitude. High impact injury significantly increased instrument performance, while low impact injury only increased performance when both age and the age/low impact interaction terms were included in the model. Low impact injury was associated with a below average mortality in the young and an above average mortality in those over 54 years. Anatomical information when recorded as zero, one or multiple severe injuries also improved the predictive validity of the RTS and the instrument which included age and the RTS.

Injury Class, recorded as blunt and penetrating injury, does not improve prediction, and co-morbidity in itself does not predict mortality when adjusted for age.

In conclusion, likelihood ratio and ROC curve analyses demonstrated that all triage instruments are not equal. It is clear that when more difficult endpoints than mortality are used, instruments with increased information content, perform better.

This study concurs with the consensus conference which placed validity as the most important instrument quality. The lowest ranked quality presented by that conference was simplicity. The problem with physiologic instruments is that their excess simplicity has resulted in low face and content validity. This study demonstrated that age, impact of injury and number of serious injuries, information which is available in the pre-hospital situation, should be included in severity instruments to improve the decisions based on these measurements.

Future research based on the findings in this study should include the prospective pre-hospital collection of information on the patients age, impact of injury and number of serious injuries and an evaluation of the predictive validity of instruments that include these attributes. Outcomes evaluated should allow discrimination between instruments.

Instruments for measuring trauma severity are an integral part of regional trauma care systems. This study shows that standardized statistical techniques and the application of the principles of measurement theory are effective strategies for the refinement of these instruments.

REFERENCES

1. Trunkey DD: Trauma - A Public Health Problem in Early Care of the Injured Patient. Moore EE (ed.), BC Decker Inc. Toronto, Ontario, 1990.

2. Mulder DS: Entre Amis: 1985 Presidential Address, American Association for the Surgery of Trauma. *J Trauma*, 26:207-215, 1985.

3. Cales RH: Injury Severity Determinations: Requirements, Approaches, and Applications. *Annals of Emergency Medicine*, 15:1427-1433, 1986.

4. Flint L: Regionalization of Trauma Care in Early Care of the Injured Patient. Moore EE (ed.), BC Decker Inc. Toronto, Ontario, 1990.

5. MacKenzie EJ: Injury Severity Scales Overview and Direction for Future Research. *American J Emerg Med*, 2:537-548, 1984.

6. Champion HR, Copes WS, Sacco WJ and Flanagan ME: Injury Severity Scoring In: Early Care of the Injured Patient. Moore EE (ed.), 4th Edition, BC Decker Inc. Toronto, Ontario 1990.

7. Gibson G: Indices of Severity for Emergency Medical Evaluative Studies: Reliability Validity and Data Requirements.

8. Kirscher JP: Indexes of Severity: Conceptual Development. *Health Services Research*, 11:143-157, Summer 1976.

9. Goldberg JL, Goldberg J, Levy PS, et al.: Measuring the Severity of Injury: The Validity of the Revised Estimation Survival Probablity Index. *J Trauma*, 24:420-427, 1984.

10. Fung YC: The Application of Biomechanics to the Understanding and Analysis of Trauma. In The Biomechanics of Trauma Nahulm, AM Melvin J (ed), Appleton-Century-Crofts. Norwalk, Connecticut 1985.

11. O'Neil B: The Statistics of Trauma. In The Biomechanics of Trauma, Nahulm AM Melvin J (ed), Appleton-Century-Crofts. Norwalk, Connecticut, 1985.

12. Waller Julian A: Injury as a Public Health Problem: Chapter 46. In Public Health and Preventative Medicine Last J EE 11th Edition, Appleton-Century-Croft. New York, NY, 1980.

13. Baker SP, O'Neil B, Karpf RS: The Injury Fact Book. Lexington Books, Toronto, Ontario, 1984.

14. Champion HJ, Copes Ws Buyer D et al.: Major Trauma in Geriatric Patients. *American J. Public Health*, 79:1278-1282, 1989.

15. Robertson LS: Injuries. Lexington Books, Toronto, Ontario, 1983.

16. Barancik JI, Chaterjee BF, Greene YC et al.: Northeastern Ohio Trauma Study: I. Magnitude of the Problem. *Am J. Public Health*, 73:746, 1983.

17. Haddon W Jr: Advances in the Epidemiology of Injuries as a Basis for Public Policy. *Public Health Rep*, 95:411-421, 1980.

18. Rutherford RB, Buerk CA: The Pathophysilogy of Trauma and Shock: Chapter 3. The Management of Trauma 3rd Edition: Zuidema GD, Rutherford RB, Ballinger WF. Wb Saunders, Toronto, Ontario, 1979.

19. Baker CC, Oppenheimer L, Stephens B et al.: Epidemiology of Trauma Deaths. *AM J Surg*, 140:144-150, 1980.

20. Dove D, Stalh WM, Del Guercio LRM: A Five-Year Review of Deaths Following Urban Trauma. *J of Trauma*, 20:760-766, 1980.

21. Gilroy D: Deaths From Blunt Trauma: A Review of 105 Cases. *Injury*, 15:304-308, 1984.

22. Hedges JR, Reero Stan, Moore B et al.: Comparison of prehospital trauma triage instruments in a semirural population J. Emerg Med 5:197-207 1987

23. Shires GT, Thal ER, Jones RC et al: Trauma Chapter 6 Principles of Surgery Shwartz SI, Shires GT, Spencer FC: Principles of Surgery, McGraw Hill. New York, NY 1989.

24. Advanced Trauma Life Support Course Instructor Manual American College of Physicians and Surgeons Committee on Trauma, 1984.

25. Yates DW: Scoring Systems for Trauma. *BMJ*, 301:1090-1094, 1990.

26. Baker CC, Degutis LC, DeSantis J, Baue AE: Impact of a Trauma Service on Trauma Care in a University Hospital. *Amer J Surg*, 149:453-458, 1985.

27. Shackford SR, Mackenzie RC, Hoyt DB, et al.: Impact of a Trauma System on Outcome of Severly Injured Patients. *Arch Surg*, 122:523-527, 1987.

28. McMurtry RY, Nelson WR: A Trauma Centre - What is Involved. *J Can Assoc of Radiologists*, 34:163-166, 1983.

29. Committee on Trauma of the American College of Surgeons: Field Categorization of Trauma Patients (Field Triage) - Appendix F to the *Hospital Resources Document American College of Surgeons Bulletin*, October 1986.

30. Hamilton S: Coalition on Trauma - Trauma Prevention and Trauma Care: Presidential Address. Trauma Associaton of Canada. *J Trauma*, 31:951-957, 1991.

31 Morris JA, Sanches AA, Bass SM and Mackenzie EJ: Trauma Patients Return to Productivity. *J Trauma*, 31:827-834, 1991.

32. Kirkpatrick JR, Youmans RI: Trauma Index, An Aide in the Evaluation of Injury Victims. *J Trauma*, 11:711-714, 1971.

33. Williams RE and Schamadan JL: The 'SIMBOL' Rating and Evaluation System. *Arizona Medicine*, 886-887, 1969.

34. Kane G, Engelhardt R, Celentano J et al.: Emperical Development and Evaluation of Pre-hospital Trauma Triage Instruments, 25:482-488, 1985. J. Trauma 25:482-488 1985

35. Hawkins ML, Treat RC, Mansperger AR: Trauma Victims: Field Triage Guidelines. *Southern Medical Journal*, 80:562-565, 1987.

36. Shannon FL, Moore EE: Primary Repair of the Colon: When is it a Safe Alternative? *Surgery*, 98:851-859, 1985.

37. Champion HR, Sacco WJ, Long W: Indications for Early Hemodialysis in Multiple Trauma. *Lancet*, 1:1125-1127, 1974.

38. Trunkey DD (moderator), Siegel J, Baker SP et al.: Panel: Current Status of Trauma Severity Indices. *J Trauma*, 23:185-201, 1983.

39. Smith JS, Bartholomew MJ: Trauma Index Revisited: A Better Triage Tool *Critical Care Medicine*, 18:174-180, 1990.

40. Long WB, Bachulis BL, Hynes GD: Accuracy and Relationship of Mechanism of Injury, Trauma Score, and Injury Severiy Score in Identifying Major Trauma. *Amer J Surg*, 151:581-584, 1986.

41. Lowe DK, Oh GR, Neey KW and Peterson CG: Evaluation of Injury Mechanisms as a Criterion in Trauma Triage. *Amer J Surg*, 152:8-10, 1986.

42. Clemmer TP Orme JF, Thomas F, Brooks KA: Prospective Evaluation of the CRAMS Scale for Triaging Major Trauma. *J Trauma*, 25:188-191, 1985.

43. Champion HR, Sacco WJ, Copes WS et al.: A Revision of the Trauma Score. *J Trauma*, 29:623-629, 1989.

44. Baxt WG, Berry CC, Epperson MD et al.: The Failure of Prehospital Trauma Prediction Rules to Classify Trauma Patients Accurately. *Annals of Emergency Medicine*, 18:21-28, 1989.

45. Ornato J, Milnek E. Jr., Draren EJ, and Nelson N: Ineffectiveness of the Trauma Score and the Crams scale for Accurately Triaging Patients to Trauma Centers Ann Emerg Med 14:1061-1066 1985

46. McDowell I, Newell C: Measuring Health - A Guide to Rating Scales and Questionaires. Oxford University Press, New York, NY, 1987.

47. Streiner DL and Norman GR: Health Measurement Scales. Oxford Medical Publications, New York, NY, 1989.

48. Feinstein AR: Clinimetrics, Yale University Press, New Haven, 1987.

49. Apgar V, James LS: Further Observations on the Newborn Scoring System. *Amer J Dis Child*, 104:419-428, 1962.

50. Peel AAF, Semple T, Wang I et al: A Coronary Prognostic Index for Grading the Severity of Infarction. *British Hearth Journal*, 24:745-760, 1962.

51. Malatack JJ Schaid DJ Urbach AH et al.: Choosing a Pediatric Recipient for Orthoptic Liver Transplantation. *J Pediatr*, 111:479-489, 1987.

52. Cullen DJ, Ferrara LC, Gilbert J, Briggs BA, Walker MD: Indicators of Intensive Care in Critically Ill Patients. *Crit Care Med*, 5:173-179, 1977.

53. Krischer JP: Indexes of Severity: Underlying Concepts. *Health Services Research*, 11:143-157, 1976.

54. McNeil BJ, Hanley JA: Statistical Approaches to Clinical Predictions. *NEJM*, 304:1292-1294, 1981.

55. Rothman KJ: Modern Epidemiology. Little Brown and Company, Boston, MA, 1986.

56. Bravo G0 Potvin L: Estimating the Reliability of Continuous Measures with Cronbach's Alpha or the Intraclass Correlation Coefficient: Towards the Integration of Two Traditions. *J Clin Epidem*, 44:381-390, 1991.

57. Bland JM and Altman DG: Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet*, 1:307-310, 1986.

58. Aickin M: Maximum Likelihood Estimation of Agreement in Constant Predictive Probability Model and its Relation to Cohens Kappa. *Biometrics*, 46:293-302, 1990.

59. Koran LM: The Reliability of Clinical Methods, Data and Judgements (First of Two Parts). NEJM 293:642-646, 1975.

60. Koran LM: The Reliability of Clinical Methods, Data and Judgements (Second of Two Parts). NEJM 293:695-701, 1975.

61. Landy FJ: Stamp collecting vs Science. *Amer Psychologist*, 41:1183-92, 1986.

62. Lord FM, Novick MR: Statistical Theories of Mental Test Scores Addison-Wesley Reading Mass. 1968.

63. McNeil BJ, Keeler E, Adelstein SJ: Primer on Certain Elements of Medical Decision Making. *NEJM*, 293:211-215, 1975.

64. Koehler JJ, Malafa SA Hillesland J et al.: A Multicenter Validation of the Prehospital Index. *Annals of Internal Medicine*. 16:380-385, 1987.

65. Eichelberger MR, Gotschall CS, Saccos WJ et al.: Comparison of the Trauma Score, Revised Trauma Score and the Pediatric Trauma Score. *Annals of Emergency Medicine*, 18:1053-1058, 1989.

66. Sackett DL, Haynes RB, Tugwell P: Clinical Epidemiology - A Basic Science for Clinical Medicine. Little Brown and Company, Toronto, Ontario, 1985.

67. Radack JL, Rouan G, Hedges J: The Likelihood Ratio. *Arch Path Lab Med*, 110-693, 1986.

68. Fagan TJ: Nomogram for Bayes Theorem. *NEJM*, 293:257 1975.

69. Sacco WJ, Champion HR, Gainer PS, Morelli SA, Fallen S, Lawnicl MA: The Trauma Score as Applied to Penetrating Trauma. *Annals of Emergency Medicine*, 13:415-418, 1984.

70. Swets JA and Pickett RM: Evaluations of Diagnostic Tests: Methods From Signal Detection. Academic Press, Paris, 1982.

71. Miettinen OS: Personal Communication.

72. Miettinen O: Estimability and Estimation in Case-referent Studies. *Am J Epidemiol*, 103:226-235, 1976.

73. Beck JR, Shultz EK: The Use of Relative Operating Characteristic (ROC) Curves in Test Performance Evaluation. *Arch Path Lab Med*, 110:13-20, 1986.

74. Swets JS, Pickett RM, Whitehead SF, Getty DJ, Schnur JA Swets JB, Freeman BA: Assessment of Diagnostic Technologies. *Science*, 205:753-759, 1979.

75. Metz CE: Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine* VIII:283-298, 1978.

76. Metz C: Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC Studies. *Radiology* 24:234-245, 1989.

77. Hanley JA, McNeil BJ: A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived From the Same Cases. *Radiology*, 148:839-843, 1983.

78. Delong ER, Delong DM and Clarke-Peterson DL: Comparing the Area Under or More Correlated Receiver Operating Curves: A Non-parametric Approach. *Biometrics*, 44:837-845, 1988.

79. Swets JA: Indices of Discrimination or Diagnostic Accuracy: Their ROC's and Implied Models. *Psychological Bulletin*, 99:100-117, 1986.

80. Begg CB and Greenes RA: Assessment of Diagnostic Tests When Disease Verification is Subject to Selection Bias. *Biometrics*, 39:207-215, 1983.

81. Champion HR, Sacco WJ, Hannan DS, et al.: Assessment of Injury Severity: The Triage Index. *Crit Care Med*, 8:201-208, 1980.

82. Cohen J: A Coefficient of Agreement For Nominal Scales. *Education and Psychological MeasurEment*, 20:37-46, 1960.

83. Hanley JA: Personal Communication 1992.

84. Smith EJ, Ward AJ, Smith D: Trauma Scoring Methods. *British J of Hosp Med*, 44:114-118, 1990.

85. Ogawa M, Sugimoto T: Rating Severity of the injured by Ambulance Attendants: Field Research of Trauma Index. *J Trauma*, 14:934, 1974.

86. Champion HR, Sacco WJ, Carnazzo AJ et al.: Trauma Score. *Crit Care Med*, 9:672-676, 1981.

87. Morris JA, Auerbach PS, Marshall GA et al.: The Trauma Score as a Triage Tool in the Prehospital Setting. *JAMA*, 256:1319-1325, 1986.

88. Boyd CR, Tolson MA, Copes WS: Evaluating Trauma Care: The TRISS Method. *J Trauma*, 27:370-378, 1987.

89. Moreau M, Gainer PS, Champion H, Sacco WJ: Applicaion of the Trauma Score in the Prehospital Setting. *Annals of Emergency Medicine*, 14:1049-1054, 1985.

90. Gormican SP: Crams Scale: Field Triage of Trauma Victims. *Annals of Emergency Medicine*, 11:132-135, 1982.

91. Knudson P, Frecceri CA and DeLateur SA: Improving the Field Triage of Major Trauma Victims. *J Trauma*, 28:602-606, 1988.

92. Champion HR and Sacco WJ: Triage of Trauma Victims in Current Therapy of Trauma, Trunkey DD and Lewis FR EE., B.C. Decker Inc, Toronto, Ontario, 1986.

93. Koehler JJ, Jalafa SA, Hillesland J et al.: Prehospital Index: A Multicenter Trial. *Annals of Emerg Med*, 15:625, 1986.

94. Tepas JJ, Mollitt DL, Talbert JL and Bryant M: The Pediatric Trauma Score as a Predictor of Injury Severity in the Injured Child. *J Ped Surg* 22:14-18 1987

95. Kaufmann CR, Maier RV, Rivara FP and Carrico J: Evaluation of the Pediatric Trauma Score. *JAMA* 263:69-72, 1990.

96. Knaus WA, Zimmerman JE, Wagner DP et al.: APACHE-acute Physiology and Chronic Health Evaluation: A Physiologically-Based Classification System. *Crit Care Med*, 9:591-597, 1981.

97. Knaus WA, Draper EA, Waagner DP, and Zimmerman JE: APACHE II: A Severity of Disease Classification System. *Crit Care Med*, 13:818-829, 1985.

98. Rhee KJ, Baxt WG, Mackenzie JR et al: APACHE II Scoring in the Injured Patient. *Crit Care Med*, 18:827-830, 1990.

99. Moreau R, Soupison T, Vauquelin P, et al.: Comparison of Two Simplified Severity Scores (SAPS and APACHE II) for Patients with Acute Myocardial Infarction. *Crit Care Med*, 17:409-413, 1989.

100. Le-Gall J-G, Loirat P, Alperovitch A et al.: A Simplified Acute Physiology Score for ICU Patients. *Crit Care Med*, 12:975-977, 1984.

101. CivilID, Schwab CW: The Abbreviated Injury Scale: A Condenced Chart for Clinial Use. *J Trauma*, 28:87-90, 1988.

102. Association for the Advancement of Automotive Medicine: The Abbreviated Injury Scale - 1990 Revision.

103. Mackenzie EJ, Shapiro S, Eastham JN: The Abbreviated Injury Scale and Injury Severity Score. *Medical Care*, 23:823-835, 1985.

104. Baker SP, O'Neil B, Haddon W & Long WB: The Injury Severity Score: A Method for Describing Patients With Multiple Injuries and Evaluating Emergency Care. *J Trauma*, 14:187-196, 1974.

105. Baker SP, O'Neil B: The Injury Severity Score: An Update. *J Trauma*, 16:882-885, 1976.

106. Copes WS, Champion HR, Sacco WJ et al.: The Injury Severity Score Revisited. *J Trauma*, 28:69-77, 1988.

107. Beverland DE, and Rutherford WH: An Assessment of the Validity of the Injury Severity Score When Applied to Gun Shot Wounds. *Injury*, 15:19-22, 1983.

108. Bull JP: The Injury Severity Score of Road Traffic Casualties in Relation to Mortality, Time of Death, Hospital Treatment Time and Disability. *Accid Anal and Prev*, 7:249-255, 1975.

109. Boyd CR, Tolson MA, Copes WS:Evaluating Trauma Care: The TRISS Method *J.Trauma* 27:370-378 1987

110. Champion HR, Copes WS, Sacco WJ: A New Characterization of Injury Severity. *J Trauma*, 30:539-546, 1990.

111. Cayten CG, Stahl WM, Murphy JG et al.: Limitations of the TRISS Method for Interhospital Comparisons: A Multihospital Study. *J Trauma*, 31:471-482, 1991.

112. Contington EM, Shufflebarger Cm, Townsend R: The Power of the Z Statistic : Implications for Trauma Research and Quality Assurance Review. *J Trauma*, 29:1500-1509, 1989.

113. Levy PS, Goldberg J, Rothrock J: The Revised Estimated Survival Probability Index of Trauma Severity. *Public Health Reports*, 97:452-459, 1982.

114. Milholland AV, Cowley RA & Sacco WJ: Development and Propective Study of an Anatomical Index and Acute Trauma Index. *The American Surgeon*, 45:246-254, 1979 .

115. Champion HR, Sacco WJ, Lepper RL et al.: An Anatomic Index of Injury Severity. *J Trauma*, 20:197-202, 1980.

116. Sacco WJ, Jameson JW, Copes WS et al.: Progress Towards a New Injury Severity Characterization: Severity Profiles. *Comput Biol Med*, 18:419-429, 1988.

117. West J.G.: An autopsy method for evaluating trauma care. *J Trauma* 21:332-34 1981

118. Cales RH and Trunkey DD: Preventable Trauma Deaths. *JAMA*, 254:1059-1063, 1985.

119. West JG: Validation of Autopsy Method for Evaluation of Trauma Care. *Arch Surg*, 117:1033-1035, 1982.

120. Young WW, Young JC, Smith JS Jr. et al.: Defining the Major Trauma Patient and Trauma Severity. *J Trauma*, 31:1125-1135, 1991.

121. Champion HR: Editorial. *J Trauma*, 31:1049-1050, 1991.

122. Cheadle WG, Wilson M, Hershman MJ et al.: Comparison of Trauma Assessment Scores and Their Use in Prediction of Infection and Death. *Ann Surg*, 209:541-546, 1989.

123. Kleinbaulm DG, Kupper LL and Morgenstern H: Epidemiologic Research, Principles and Quantitative Methods. Van Nostrand Reinhold, New York, NY, 1982.

124. SAS Institute Inc.: SAS/STAT Users's Guide, Version 6 Fourth Edition Volume 2, Cary, NC:SAS Institute Inc.' 1989.

125. Bartko JJ: The Intraclass Correlation Coeffiecient as a Measure of Reliablity. *Psych Reports*, 19:3, 1966.

126. Fleiss JL: The Design and Analysis of Clinical Experiments. John Wiley and Sons, New York, NY, 1986.

127. Hosmer DW and Lemeshow S: Applied Logistic Regression. John Wiley and Sons, New York, NY, 1989.

128. McAnena OJ, Moore EE, and Marx JA: Initial Evaluation of the Patient with Blunt Abdominal Trauma. *Surgical Clinics of North America*, 70:495-515, 1990.

129. Oreskovich MR, Howard JD, Copass MK et al.: Geriatric Trauma: Injury Patterns and Outcome. 24:565-572, J. Trauma 24:565-572 1984