

**TRAUMA SCORING MODELS**  
**USING**  
**LOGISTIC REGRESSION**

**MD THESIS**

**Mr John Stephen Batchelor.**

**MB ChB. FRCS(I). FFAEM.**

**Leonard Cheshire Department of Conflict**

**Recovery, UCL, London.**

**2003**



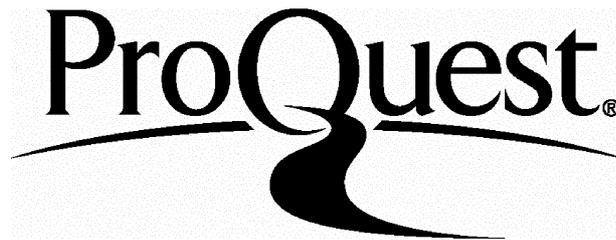
ProQuest Number: 10016140

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

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



ProQuest 10016140

Published by ProQuest LLC(2016). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code.  
Microform Edition © ProQuest LLC.

ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

## ABSTRACT

**Introduction.** Trauma scoring models form an important part in trauma audit. Limitations of the TRISS model however has led investigators to search for new models which more accurately predict trauma deaths. Logistic regression has played an integral role in the development of such models in view of the fact that most investigators in this field use a dichotomous dependent variable (death/survival). Limitations in the methods used to evaluate new models currently makes it difficult to assess their true worth.

**Aims.** The aim of this thesis was to evaluate current methods of model validation and also to evaluate alternative options.

**Methods and Results.** A data set of 7069 complete trauma cases from the Los Angeles Trauma Registry was used for the study. Five goodness of fit tests were studied and were compared to the Hosmer Lemeshow test. The Copas test was found to be the most viable alternative to the Hosmer Lemeshow test. The two main methods of validating a prognostic model (i.e. data splitting and cross-validation) were evaluated using a series of simulation studies. Both methods were found to be unreliable with regard to their ability to reduce over-fitting of a prognostic model. Bootstrapping was also evaluated as a means of generating confidence intervals for the Hosmer Lemeshow test and the Copas test. Two models were used to assess the accuracy of the percentile method. Accurate bootstrap confidence intervals were developed for the Copas test but not for the Hosmer Lemeshow test.

**Conclusions.** The Copas test with bootstrap confidence intervals provides a superior approach to validating a trauma model compared to the currently employed methods.

## **Acknowledgements**

I am indebted to Professor Demetriades for allowing me to use the Los Angeles trauma Registry data.

I am also grateful to Professor Ryan for his advice, supervision and encouragement.

# CONTENTS

	<b>Page Number</b>
<b>Chapter 1:</b> Introduction.	5
<b>Chapter 2:</b> Historical review.	11
<b>Chapter 3:</b> Data set preparation.	25
<b>Chapter 4:</b> The evaluation of trauma scoring models using the revised USC data set.	60
<b>Chapter 5:</b> A study to compare the three different methods of calculating the Hosmer Lemeshow chi-square statistic.	90
<b>Chapter 6:</b> A study to determine the effects of changing the covariate pattern on the Hosmer Lemeshow chi-square statistic.	118
<b>Chapter 7:</b> A comparison of six goodness of fit tests by sequential increase in data set size.	147
<b>Chapter 8:</b> A study to evaluate model validation using data splitting.	191
<b>Chapter 9:</b> A study to evaluate cross-validation using trauma scoring modelling.	210
<b>Chapter 10:</b> A study to determine the accuracy of bootstrap generated confidence intervals for two goodness of fit tests in logistic regression.	245
<b>Chapter 11:</b> A study to evaluate trauma models using the Copas goodness of fit test.	300
<b>Chapter 12:</b> Conclusions.	308
<b>Chapter 13:</b> References.	315

# **CHAPTER 1**

## **INTRODUCTION**

### **AN OVERVIEW OF LOGISTIC REGRESSION IN TRAUMA SCORING MODELLING**

Trauma is the third leading cause of death across all ages in the developed world (Bourbeau, 1993). A combination of accident prevention programs and optimal trauma care has helped to reduce the burden of this epidemic. Trauma scoring has become an integral part of many trauma institutions largely due to the pioneering work of Professor Howard Champion and colleagues (1980, 1981, 1983, 1989, 1990a). Trauma scoring models serve two main functions. The original idea behind a trauma scoring system was to provide a means of pre-hospital triage for paramedic crews. A critical value within a scoring system is used to determine whether a patient is taken to a designated trauma centre.

In the UK pre-hospital triage scoring systems have not gained great impetus in the pre-hospital setting largely due to the scarcity of designated trauma centers. Triage trauma scoring systems have in the UK been used as a means of trauma team activation. Triage trauma scoring models have largely been developed using 2 x 2 contingency tables and the calculation of sensitivity and specificity rates. Logistic regression has not played a large part in the development of these models and as such will not be discussed further in this chapter. A detailed review of trauma triage models has been published by Batchelor (2001).

The other main purpose of trauma scoring models is within the context of trauma audit. The TRISS model (Champion, 1990a) enables the survival probability of a trauma patient to be calculated. An institution's trauma survival rates can be compared with that predicted by TRISS. The  $W^*$  statistic measures the difference between the actual and predicted survival rates. The statistical

significance of this difference is measured by the  $z^{**}$  statistic. The  $M$  statistic is a measure of the similarity of the injury severity case mix of the institution compared to the injury severity case mix of the prediction database. Hollis et al (1995) pointed out that it is possible for two institutions with the same survival rate to have quite different  $W$  and  $Z$  scores despite having similar  $M$  scores. They proposed a new statistic  $W_s$  which is a standardized statistic with respect to injury severity case mix. An additional means of improving these statistics was suggested by Demetriades (2001). Each institution could calculate its own coefficients based upon its own trauma database of patients. This idea has already been explored by Lane et al (1997). These authors found that the predictive power of TRISS could be improved using institution derived coefficients.

$$W^* \text{ statistic} = \frac{100[(\text{Observed survivors}) - (\text{expected survivors})]}{\text{Total patients}}$$

$$Z^{**} \text{ statistic} = \frac{(\text{observed survivors}) - (\text{expected survivors})}{\sqrt{(\sum [P_s \times (1 - P_s)])}}$$

At present two main methods are used to assess the worth of a logistic trauma scoring model. These tests are normally performed on a separate data set from which the model or coefficients are derived in order to prevent over-fitting of the model. The Hosmer Lemeshow goodness of fit statistic (Hosmer and Lemeshow, 1989) is a measure of the model's calibration i.e. the precision of the model to predict survival (or death) over a range of injury severities. It is calculated using the formula:-

$$HL = \frac{\sum(\text{Observed} - \text{Expected Cases})^2}{\text{Expected Cases}}$$

A detailed description of how the statistic is calculated is provided in chapter 5. The smaller the HL value the better is the model calibration. The major problem with the HL statistic is that it is dependent upon the size of the data set. As the data set becomes smaller the HL value decreases and the significance value  $p$  also increases, thus erroneously indicating that a poorly fitting model fits better with a smaller data set. Because  $p_i$  is a biased estimator of  $p$  (i.e. is dependent on data set size) most investigators working in this field now place little importance on the  $p$  value.

To use the HL test effectively a moderately large sample size is required so that there are at least five expected events in each group (SPSS Regression Models, 1999). There are slight variations in the way that the statistical software packages calculate the HL statistic and this can add to further variability in the result (Hosmer, 1997).

The strength of the HL statistic is that it appears to be able to assess the worth of a model with a reasonable level of precision within the confines of a single data set. The major weakness of the HL statistic is that it does not allow comparisons to be made between results obtained from different sized data sets. The Receiver Operating Curve (ROC) is the other main statistical tool used to evaluate logistic trauma models. The ROC curve is a measure of the model's discrimination i.e. it measures how effective the model is at differentiating between survivors and non-

survivors. The ROC statistic is more robust than the HL tool in that it is less dependent on data set size.

Misclassification rates are no longer recognised as being an appropriate means of assessing the worth of a logistic model due to the arguments put forward by Harrell et al (1996). The argument against the use of misclassification rates can be explained by a simple example. If the model predicts that a group of patients have a 70% chance of survival ( $p=0.7$ ) then the model is also predicting that 30% of patients with this probability will also die. The deaths in this group are therefore not misclassified deaths.

The search for the optimal trauma scoring model still continues. A recent study by Meredith et al (2002) using the National US Trauma Data Bank of 76,871 patients found there was little difference between an ICD-9 based model and an injury severity based model (modified Anatomical Profile Score model; {Sacco, 1999}). ICISS had the best discrimination using ROC analysis. The modified Anatomical Profile Score model had the best calibration using the Hosmer Lemeshow test.

Although confidence intervals are routinely outputted in many statistically software packages that calculate ROC values the same is not true for the Hosmer Lemeshow statistic. The standard method for generating confidence intervals for many statistical parameters is by bootstrapping. This method generates additional data sets by random sampling with replacement from the original data set. If a large number of data sets are produced e.g. 1000 then confidence intervals can be calculated around the mean bootstrap

value. There have been no studies to date which have examined whether the bootstrap method can be used to generate confidence intervals for the Hosmer Lemeshow statistic. The use of confidence intervals would provide a greater degree of 'confidence' in the results rather than relying on a single set of values developed either from a single data split or a single external data set.

# **CHAPTER 2**

## **HISTORICAL DEVELOPMENT OF ADULT TRAUMA SCORING MODELS**

### **CONTENTS**

	Page Number
Section 1. Historical background of the AIS and ISS.	12
Section 2. Historical background of the Revised Trauma Score.	13
Section 3. TRISS (Revised Trauma Score/Injury Severity Score) and ASCOT models.	14
Section 4. Additional problems with the TRISS model.	17
Section 5. New Injury Severity Score (NISS).	19
Section 6. International Classification of Diseases Injury Severity Score (ICISS).	21
Section 7. The Anatomical Profile Score.	23
Section 8. Conclusions.	24

## **Section 1. Historical Background of the AIS and ISS.**

The abbreviated Injury Scale (AIS) was introduced by the Committee on Medical Aspects of Automotive Safety (1971) in order to quantify injuries sustained by road traffic accidents. The aim behind the AIS was to provide safety information to automobile design engineers. The original AIS was composed of nine categories of injury. Categories 1-6 are shown in table 1. AIS scores of 7 to 9 each defined fatal injuries at the scene or within 24 hours irrespective of injury severity. These categories were subsequently discarded. The Abbreviated Injury Scale was modified into the Injury Severity Score by Baker and colleagues from John Hopkins University (1974, 1976). Baker et al (1974) found that the mortality rate of patients with more than one injury was best represented by a quadratic equation of the form  $ISS = AIS^2 + AIS^2 + AIS^2$ . The three AIS scores representing the most severe injury from three separate body regions. Only one injury per body region can be represented in the ISS. Three AIS scores were chosen because the addition of a fourth AIS was not found to improve the model fit. The AIS underwent revisions in 1980 and 1985. In 1990 a third revision of the original AIS was published (American Association for Automotive Medicine). These revisions have resulted in significant improvements over the original AIS. The 75 injuries in the original AIS have been expanded to include more than 2000 injuries of both blunt and penetrating in type. A further revision of the AIS is currently in progress.

Table 1  
Abbreviated Injury Scale

<u>AIS Value</u>	<u>Description</u>
0	No injury.
1	Mild.
2	Moderate.
3	Severe (not life-threatening).
4	Severe (life-threatening, survival probable).
5	Critical (survival unclear).
6	Fatal injury.

## **Section 2. Historical Background of the Revised Trauma Score.**

Kirkpatrick and Youmans (1971) published one of the earliest pre-hospital triage scoring systems called the Trauma Index (TI). The Trauma Index had five coded variables; anatomical region, wound type, cardiovascular status and conscious level. Ogawa and Sugimoto (1974) evaluated a modified version of the Trauma Index. They found consistent correlation between ambulance crews' estimates of injury severity using the modified Trauma Index and the authors' outcome definition of major trauma, which was based upon hospitalization rates. The Trauma Index however was never fully evaluated at the time and failed to gain popularity.

Champion et al (1980) published a new injury severity score called the Triage Index. The Triage Index contained five physiological coded variables (respiratory expansion; capillary refill; eye

opening; verbal response and motor response) which formed an interval scale. The Triage Index was evaluated by Champion et al (1980) using a logistic regression model with a decision rule of 0.5 to determine the misclassification rate. The Triage Index was shortly amended to the Trauma Score (Champion, 1981). The Trauma Score (TS) also contained five physiological variables: respiratory rate; respiratory effort; systolic blood pressure; capillary refill and Glasgow Coma Scale. The Trauma Score was also evaluated using a logistic regression model. Misclassification rates and a less well known statistic called relative information gain\* were used to assess the model's goodness of fit. A further revision of the original Triage Index was published by Champion et al (1989) and the Trauma Score was amended to its current form:- the Revised Trauma Score.

\*information gain  $E = 2P(1-P) - \text{misclassification rate}$

\*relative information gain  $R = E/2P(1-P)$

### **Section 3. TRISS and ASCOT Models.**

The TRISS probability of survival trauma scoring model was developed by Champion et al (1981) using logistic regression. The model was developed originally using the Trauma Score (Champion, 1981). The Revised version of the Trauma Score was subsequently incorporated into the TRISS methodology (Champion, 1989).

One of the envisaged uses of the TRISS methodology was to identify unexpected trauma deaths and unexpected trauma survivors. An unexpected trauma death was defined as any patient who died whose probability of survival was  $> 0.5$  (TRISS fallout).

An unexpected survivor was any patient who survived whose probability of survival was  $< 0.5$ . Karmy-Jones et al (1992) showed that TRISS unexpected outcomes resulted in 54% of cases being misclassified compared to peer review. Hill et al (1992) also prospectively evaluated TRISS fallout groups and found that they tended to over estimate potentially avoidable deaths especially in patients with severe head injuries. More recently Norris et al (2002) reviewed 270 patients categorised by TRISS as unexpected survivors. Only 10.7% were found to be clinically unexpected survivors following peer review. Despite the limitations of TRISS with respect to fall-out cases TRISS methodology still remains an invaluable tool upon which to base trauma audit.

Cayten et al (1991) identified some further limitations of TRISS methodology, the most important being the inability to account for multiple injuries in one body region. This limitation of the TRISS model led Champion et al (1990b) to develop an alternative anatomically based injury severity scoring system call the Anatomical Profile Score. The Anatomical Profile defines four components; A (head and spinal cord, AIS  $> 2$ ), B (thorax and front of neck, AIS  $> 2$ ), C (all other serious injuries, AIS  $> 2$ ), D (all injuries with AIS  $\leq 2$ ). Component D although defined by the Anatomical Profile method is not utilised in any subsequent calculations. The component scores are calculated by taking the square root of the sum of all the AIS scores greater than 2 for that component. The square root value was used in the methodology because Champion et al (1990b) argue that this gives better representation to the most severe injury in each region, additional

injuries having a subliminal effect rather than a simple additive effect. The Anatomical Profile Score was subsequently incorporated into A Severity Characterisation of Trauma (abbreviated to:-ASCOT) model (Copes, 1990). The other variables in the logistic model being the individual coded values of the Revised Trauma Score:- Glasgow Coma Scale (G), systolic blood pressure (S), respiratory rate (R). The final variable in the model being age which is represented as a five component interval scale rather than a dichotomous variable as in TRISS.

The ASCOT model takes the form:-

$$K = K0 + K1(G) + K2(S) + K3(R) + K4(A) + K5(B) + K6(C) + K7(\text{Age}).$$

$$PS = 1/1+e^{-K}$$

It is important to note that the ASCOT method has four set-aside groups whose probability of survival is either extremely good or extremely poor; (1) AIS score = 6 and RTS = 0, (2) Maximum AIS score < 6 and RTS = 0, (3) AIS score 6 and RTS > 0, (4) Maximum AIS score = 1 or 2 and RTS > 0. These four set-aside groups were given a probability of survival based upon their survival rates obtained from the study data set and were not therefore used in further model development. The ASCOT and also new TRISS coefficients were developed from a data set 15,957 patients taken from the MTOS database. The two models (ASCOT and TRISS) were then evaluated on a second data set of 15,954 patients also derived from the MTOS database. The HL value for ASCOT and TRISS models were; ASCOT 24.8, TRISS (new coefficients) 43.9, TRISS (1986 MTOS coefficients) 45.6.

Markle et al (1992) undertook a comparative study of the TRISS model (using MTOS derived  $\beta$  coefficients) versus ASCOT (*m* coefficients were those derived by Copes et al {1990}) on a data set of 5,685 patients. Both models had p values for the HL statistic that were  $< 0.001$  i.e. both appeared to be a poor fit. Hannan et al (1995a) compared ASCOT to TRISS using coefficients derived from the ITEC database. They found that the ASCOT model performed acceptably but that the TRISS model did not. Hou et al (1996) performed a comparative study of TRISS versus ASCOT on a data set of 5,672 cases. Their results were inconclusive. A further evaluation study of the ASCOT model was performed by Champion et al (1996), ASCOT being compared to TRISS using the 1986 MTOS  $\beta$  coefficients. There was no significant difference between the ASCOT and TRISS models using ROC analysis in either the blunt or penetrating subgroup of patients. The HL values for the ASCOT model were superior to the TRISS model for both the blunt and penetrating group. Osterwalder et al (2000) compared the predictive survival rates of TRISS and ASCOT with the observed survival rates. The predictive survival rates were not significantly different between the TRISS and ASCOT models. These authors argued on the basis of their findings that the additional data collection effort required for ASCOT compared to TRISS may not be justified.

### **Section 5. Additional Problems with the TRISS Model.**

The original TRISS model used age as a dichotomous variable in the adult trauma group. Using a continuous variable in this way tends to result in either over prediction or under prediction of

survival. Since the inception of the TRISS model subsequent researchers have tended to use age as a continuous variable rather than as a dichotomous one. Demetriades et al (1998) found that using age as a dichotomous variable was an important cause for misclassifications in the TRISS fallout group. Al West et al (2000) used age as a quadratic spline (knots at 12, 25, and 65 years) in a logistic regression model which also included ICD-9 codes for anatomic injury, mechanism of injury and pre-existing disease. Two-way interaction terms for several combination of injuries were also included in the model. This model (HARM: Harborview Assessment for Risk of Mortality) was found to be superior to both TRISS and ASCOT using calibration and discrimination statistics.

Milzman et al (1992) found that pre-existing disease was an important independent risk factor for predicting death. The study involved a comparison of patients with and without pre-existing disease controlling for age and ISS. Sacco et al (1993) also found that pre-existing disease had a significant impact on survival. These authors however also showed that because the prevalence of pre-existing disease was relatively small in the trauma population (4.8%) institutional performance measures (Z and W values) were unlikely to be significantly affected. Other variables have been identified as potential independent risk factors for death in trauma patients e.g. falls (Hannan, 1995b; Demetriades, 1998), elevated blood lactate levels (Sauaia, 1994), blood pH (Milham, 1995) and a Systemic Inflammatory Response Syndrome\* Score of 2 on admission (Napolitano, 2000).

\*SIRS score ranges from 1-4. One point is scored for each component present; fever or hypothermia, tachycardia, tachypnea and leukocytosis.

It has been argued that death/survival is a crude outcome measure and this has led other investigators to use other outcome measures such as multiple organ failure (Roumen, 1993; Sauaia, 1994) and length of hospital stay (Clark, 1997). Multiple organ failure accounts for only a small proportion of trauma deaths (7%: Sauaia, 1995) and therefore is only applicable to a subgroup of trauma patients. Length of stay as an outcome variable has the drawback that it requires the exclusion of the subgroup of patients who die early (i.e. within 24 hours).

### **Section 5. New Injury Severity Score (NISS).**

Osler et al (1997) suggested a simple solution to overcome one of the main limitations of ISS i.e. its inability to account for multiple injuries in one anatomical region. The modification consisted of summing the squares of the three highest AIS scores irrespective of anatomical region. The modification was named NISS (New Injury Severity Score) by the authors. NISS was evaluated on two independent data sets (Albuquerque data set  $n = 3,136$  and Oregon data set  $n = 3,449$ ) and a comparison was made with ISS. The study appeared to demonstrate the superiority of NISS over ISS using the HL statistic and ROC analysis. A comparative study of NISS versus ISS was performed by Brenneman et al (1998) on a data set of 2,328 patients. The two scoring systems were discrepant in 68% of cases. In the group with discrepant scores NISS was superior to ISS using ROC curve analysis (0.852 versus 0.799  $p < 0.001$ ). Moini et al (2000) performed a comparative study of TRISS versus NISS + RTS + Age on a data set of 2,662 patients. Using TRISS derived Z and W scores they suggested that TRISS was superior to NISS + RTS + Age. Neither logistic regression nor

neural networks were used to evaluate the two models. Balogh et al (2000) performed a study comparing NISS to ISS on a data set of 295 patients using post-injury multiple organ failure (MOF) as the outcome variable. They found that NISS was superior to ISS. The main limitation of this study was that it was performed on a very small data set.

Al West (2000a) performed a comparative study of four models (ISS, NISS, TRISS, NTRISS {NISS in lieu of ISS}) using the National Trauma Data bank of 52,566 cases. ISS was found to outperform NISS using measures of calibration (ISS HL: 134.4; NISS HL: 166.8) and discrimination (ISS ROC: 0.888; NISS ROC: 0.882;  $p = 0.0001$ ). There was no significant difference in the ROC values ( $p = 0.039$ ) between TRISS (0.956) and NTRISS (0.958). The HL values however showed a slight superiority of TRISS with new coefficients (HL: 186.2) over NTRISS (207.7). Grisoni et al (2001) evaluated ISS and NISS using ROC and HL analysis on 9,151 paediatric patients. They found that the predictive ability of NISS (using ROC) was not significantly superior to ISS, neither was the HL value. The majority of patients in the data set had mild injuries with an  $ISS < 9$ .

More recently Husum (2002) performed a comparative study of NISS versus ISS on a data set of 1,787 patients with penetrating war injuries. NISS was significantly better than ISS at predicting post-injury complications, however the accuracy of both tests was only moderate. The authors of this study emphasised that the mortality in this study group was only 2.7%. At the present time the issue as to whether NISS is superior to ISS remains unresolved.

The differences between the two models do however appear to be partly dependent upon the composition of the data set.

## **Section 6. International Classification of Diseases**

### **Injury Severity Score (ICISS).**

The foundations of the ICD based scoring models are attributable to the work by Levy and co-workers (Levy, 1978; Goldberg, 1980; Levy, 1982; Goldberg, 1984). They conceptualised that the probability of survival could be calculated by multiplying the probability of survival for each injury using ICD based codes. They developed an Estimated Survival Probability Index (RESP) (Levy, 1978; Levy, 1982; Goldberg, 1984). This index failed to gain popularity at the time and was shortly superseded by the TRISS model.

Rutledge et al (1993) developed a data driven injury severity scoring model based upon the ICD-9 coding system. A mortality risk ratio (MRR) was developed for each ICD-9 code using the North Carolina database. The MRR model developed by Rutledge et al (1993) was based upon the premise that a given injury in isolation had the same value as a given injury in combination. This was a major weakness in the model design which probably led to it being abandoned.

The ICISS (International Classification of Disease-9 based Injury Severity Score) was developed by Osler et al (1996). The ICISS injury severity scoring model was based upon the ICD-9 discharge codes (800-959.9). A survival risk ratio (SRR) was determined for

each of the ICD-9 trauma injury codes using the North Carolina trauma database of 314,402. The SRR was calculated by dividing the number of times an ICD-9 code occurred in a surviving patient by the total number of times the ICD-9 code occurred in the data set. The ICISS was calculated for a given patient by the product of all their SRR's. Using ROC analysis, outcome prediction was maximal when the five worst injuries (SRR's) were used. However the outcome prediction was not reduced by using more injuries and so to avoid problems in injury selection all the SRR's are used to calculate the ICISS. The ICISS was validated on the New Mexico database of 3,142 patients and the results were compared to ISS. ICISS was found to outperform ISS using ROC curve analysis and misclassification rates.

Two follow-up studies were subsequently published by Rutledge et al (1997, 1998). The first follow-up study (1997) showed that calibration of the ICISS model was superior to ISS using the HL test although the discrimination between the two models using ROC analysis was not significant. The results of the second follow-up study (Rutledge, 1998) also showed that the ICISS model was superior to the TRISS model. Hannan et al (1999) performed a comparative study comparing ICISS (ICD-9) with TRISS using the New York State trauma registry database. Although the ICISS outperformed both TRISS models the differences in results were less impressive than those obtained by Rutledge et al (1998).

## **Section 7. The Anatomical Profile Score.**

Sacco et al (1999) performed a prospective study comparing five anatomical injury severity scales; ISS, ICISS, NISS, Anatomical Profile Score (APS{Copes et al, 1990}) and a modified version of the Anatomical Profile Score (mAP). The modified Anatomical Profile Score is composed of the three components (A,B,C) of the APS as previously described in section four of this chapter plus the maximum AIS value across all body regions. Their ICD mapped counterparts (ICD/ISS, ICD/NISS, ICD/APS, ICD/mAP) were also evaluated. The models were assessed using logistic regression on the PTOS database (derived from Pennsylvania's 26 Level I trauma centres) using a data set of 60,574 patients. Coefficients were developed from half of the data set and model validation was performed on the other half. Using ROC analysis there was little difference between the models (ISS: 0.86, NISS: 0.86, ICISS: 0.88, APS: 0.87, mAP: 0.87). The HL statistic did however demonstrate large differences between the models (ISS: 384, NISS: 40, mAP: 29, APS: 70, ICISS: 370). The ICD mapped models showed no improvement in calibration or discrimination compared to their respective counterparts.

Stephenson et al (2002) more recently performed a comparative study of five anatomic injury severity models: mapped (i.e. ICDMAP-90) modified-APS, mapped APS, mapped ISS, mapped NISS and ICISS. The study was performed on 349,409 cases derived from the New Zealand National Minimum Data Set. The ICISS model was found to have the largest concordance value (ROC=0.901). The mapped modified-APS had the best calibration using the HL statistic. Meredith et al (2002) also performed a

comparative study of nine scoring algorithms using a data set of 76,871 cases taken from the National Trauma Data Bank. The models used were modified-APS, ISS, NISS, maxAIS, mapped modified-APS, mapped ISS, mapped NISS, mapped maxAIS (i.e. the largest AIS score across all body regions) and ICISS. Model fit was assessed using ROC and HL statistics. Model validation was performed by 10-fold cross-validation, the final ROC and HL statistics being the mean of the cross validated results. The results from this study showed that the ICISS had the greatest discrimination (ROC=0.893 c.f. modified-APS ROC=0.887 {the second best score}). The modified-APS showed the best calibration (HL=12.07 c.f. maxAIS HL=19.47 {the second best score}). Interestingly ISS was found to outperform NISS with respect to calibration (HL:- NISS=48.34 HL:- ISS=33.25) and discrimination (ROC:- NISS=0.871 ISS=0.876).

## **Section 8. Conclusions**

At the current time the issue of which is the 'best' anatomic injury model remains unresolved. Based upon recent studies by Meredith et al (2002) and Stephenson et al (2002) ICISS appears to have the better discrimination, whilst modified-APS appears to have the better calibration. Despite numerous studies it remains unclear as to whether NISS is superior to ISS. Differences between these two anatomic scoring models may be due more to the composition of the training and test data sets than any inherent superiority of one model over the other.

# CHAPTER 3

## PREPARATION OF THE LOS ANGELES TRAUMA REGISTRY DATA

### CONTENTS

	Page Number
<b>Section 1.</b> Composition of the USC data set.	26
<b>Section 2.</b> Preparation of the USC revised data set.	31
Step 1. Placing data onto the SPSS spread sheet.	
Step 2. Selection of variables for model building.	
Step 3. Deletion of paediatric cases.	
Step 4. Deletion of missing variables.	
Step 5. Further refining of the data set.	
Step 6. Deletion of cases with inaccuracies by variable.	
Step 7. Identifying and deleting cases with group inaccuracies.	
Step 8. Preparation of the Mechanism of Injury variable.	
<b>Section 3.</b> Composition of the revised USC data set.	47
<b>Section 4.</b> Discussion.	56

## **Section 1: THE DATA SET**

The USC data set was obtained from the department of Traumatology, Los Angeles County Hospital. This is the Level 1 trauma centre serving the whole of Los Angeles. The hospital is affiliated to the University of Southern California. The data set will be referred to as the USC data set.

### **Section 1.1 Composition of the USC data set**

All cases entered into the trauma registry from the 1st January 1995 to the 31 December 1998 are included in the data set.

The following variables were included in the data set:-

**Age;** (years) The value is rounded off to a whole year (except when less than 12 months of age).

**Mechanism of Injury Variable.** There are two mechanism of injury variables: First mechanism of injury (MOI1) and second mechanism of injury (MOI2). The mechanism of injury codes contain a combination of epidemiology codes (e.g. work related:- WR) and actual mechanisms of injury codes (e.g. stab wound:- SW). These two types of codes may occur in either the first or second MOI group. MOI1 is the main category. MOI2 is used when additional mechanism of injury data is available.

The codes for the mechanism of injury variable categories (MOI1 and MOI2) are given below:-

VV = Versus Vehicle, subcategory of motorcycle /moped

MM = Motorcycle/moped

OT = Other mechanism of injury

PS = Passenger space intrusion

SA = Self inflicted wound, accidental

SI = Self inflicted wound, intentional

WR = Work related

WB = With blunt instrument, subcategory of assault

SF = Survivor of fatal accident, subcategory of enclosed vehicle

SP = Sports injury

GS = Gunshot

ST = Stabbing

AS = Assault

FA = Fall

TB = Thermal burn

ES = Electric shock

UN = Unknown

EV, EX, EJ, PB, TR, although often result in blunt trauma, penetrating injuries cannot be excluded.

EV = Enclosed vehicle

EX = Extrication required

EJ = Ejected from vehicle

PB = Pedestrian/bike versus vehicle

TR = Trauma arrest

SB = Seat belt. This is not strictly a mechanism of injury but is recorded in the MOI2 category when possible, in accidents when the victim is inside a vehicle.

HL = Helmet. Similarly this is not a mechanism of injury but is recorded in the MOI2 category when possible, in accidents when the victim is riding a motorbike or moped.

There are no mechanism of injuries which are restricted purely to MOI1 or MOI2 although certain combinations are commonly seen e.g. TR (MOI1) and GS (MOI2). Approximately 50% of patients had a second mechanism of injury recorded.

The number of penetrating injuries in the first mechanism of injury category was 3030 (SI = 38, SA = 24, ST = 1667, GS = 1301). The number of cases with a penetrating mechanism of injury in the second mechanism of injury category was 2110 (SI = 125, SA = 20, ST = 126, GS = 1839).

In 120 cases a penetrating injury (SI, SA, ST, GS) was recorded in both first and second mechanism of injury categories. There were only two cases where a patient had both a stabbing injury (ST) and also a gun shot wound (GS). There were no cases where the same mechanism of injury code was recorded in both the first and second category. The total number of patients with a penetrating mechanism of injury recorded was therefore 5020 [(3030 +2110) - 120].

There is good evidence from the work by Champion et al (1990a) that patients with penetrating injuries have a lower predicted probability using TRISS because of the limitations of the ISS with regard to its constraints of including only one AIS per body region. Mechanism of injury was not used for modelling because it can only be represented as a dichotomous variable i.e. penetrating versus blunt. Dichotomous variables generally perform poorly in logistic regression models.

### **Other Variables**

Abbreviated Injury Score; (abbreviated notation;- AIS).

Admission Systolic blood pressure (abbreviated notation:- SBP).

Admission Glasgow Coma Scale (abbreviated notation:- GCS).

Admission Respiratory Rate (abbreviated notation:- RR).

ISS:- Injury Severity Score calculated using ICD-9 conversion software.

HCISS:- Injury Severity Score. A hand calculated Injury Severity Score summing the square of the three highest Abbreviated Injury Scores (AIS) with the proviso that any individual body region can only be represented once.

Outcome (death or survival); coded as death = 0, survival = 1.

Body region. The region was coded using AIS 85 classification:-

Head/Neck=1, Face=2, Chest=3, Abdomen/Pelvis=4,

Extremities=5, External=6.

A total of 13447 cases are in the data set for the period:-

01/01/1995 - 31/12/1998.

Number of cases recorded per year: 1995: - 3397, 1996:- 2901,  
1997:- 3167, 1998:- 3972.

## **Section 1.2: Individual Variables**

**Age:** Patients of all ages are included in the USC data set.

Number of missing cases =  $13447 - 13443 = 4$

**AIS:** Number of missing cases =  $13447 - 13073 = 374$

**RR:** Number of missing cases  $13447 - 12796 = 651$

**SBP:** Number of missing cases  $13447 - 12993 = 454$

**GCS:** Number of missing cases  $13447 - 12747 = 700$

**HCISS:** Number of missing cases  $13447 - 13283 = 164$

**ISS:** Number of missing cases  $13447 - 13069 = 378$

**Outcome:** Number of missing cases  $13447 - 13381 = 66$

## **Section 2: Preparation of the Revised Data Set.**

### **Step 1.**

The data set was transferred from an *Excel* spreadsheet onto an SPSS spreadsheet for analysis.

### **Step 2. Selection of variables for the final data set.**

The following variables were selected; HCISS, SBP, RR, GCS, coded Outcome and Age. HCISS was selected in preference to ICD-9 mapped ISS due to the inaccuracies which may occur during the conversion. The mechanism of injury variables (MOI1 and MOI2) were also included in the revised data set so that the percentage of penetrating injuries and blunt injuries could be calculated. Blunt injury was coded 1, penetrating injury was coded 2, cases classified only by using epidemiology codes e.g. SP (sports injury) were coded 3.

#### **Step 2.1 Coding of mechanism of injury variables.**

Blunt injuries were coded 1.

Penetrating injuries were coded 2.

Epidemiological coding e.g. WR (work related), SP (sports injury).

AS (assaults) were coded 3.

Cases classified solely as UN (unknown) were coded 1.

TR (trauma arrests) were coded 4.

### **Step 3. Deletion of Paediatric Cases.**

Cases where the age of the patient was less than 16 years were deleted from the revised data set in view of the fact that paediatric

patients have different normal values for systolic blood pressure and respiratory rate depending upon age.

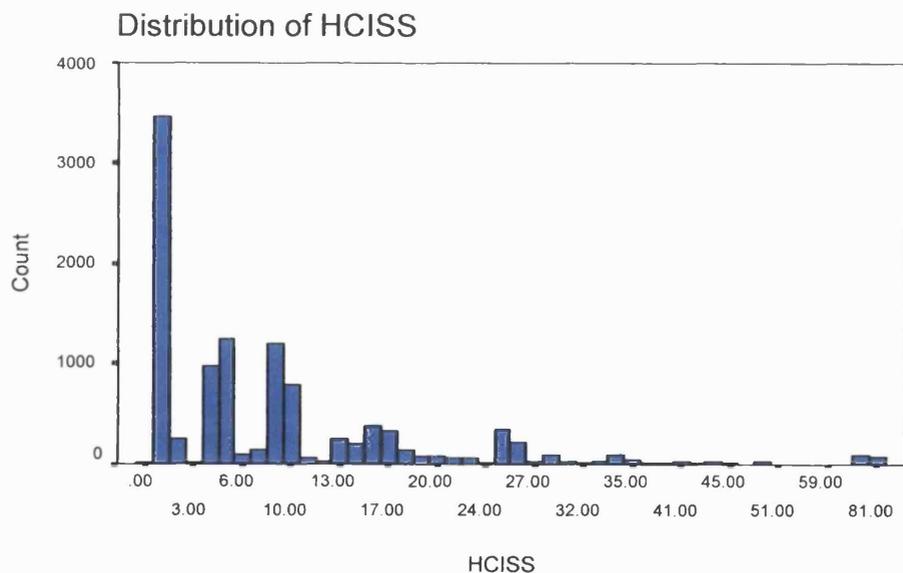
A total of 1100 cases were present where the patient's age was found to be less than 16 years. No further analysis of these cases was performed. Deletion of cases with an age of 15 years or less left the revised data set with 12,347 cases.

#### **Step 4. Deletion of Missing Variables.**

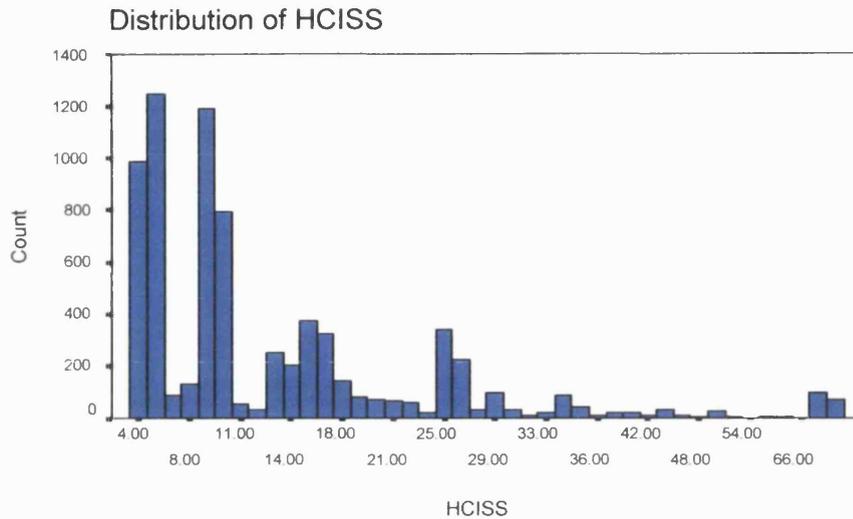
The Hosmer Lemeshow statistic is one of the main goodness of fit tests used in logistic regression modelling. This test will be evaluated further in this thesis. The value of the HL test like several other goodness of fit tests (e.g. Copas, Deviance and Pearson) increases as the data set size increases. In order to evaluate these goodness of fit tests with different predictor variables it is a pre-requisite that there are no missing values from any of the variables which are to be used for model building. Deletion of cases with missing values using the variables previously mentioned resulted in the revised data set having 11,108 cases. The cases which had missing values from the mechanism of injury variables (MOI1 and MOI2) were not deleted from the data set in view of the fact that it was not intended to use these two variables for model building. The deletion of missing variables was performed by *pasting* the data set into SAS. A short SAS program was then written using the statement; *if variable = . then delete.* The data set was then exported in excel format and transposed back into SPSS.

### Step 5.1 Refining the Data Set

All patients admitted under the Department of Traumatology have their injuries, demographic details and epidemiology recorded in the trauma registry. Many patients are admitted for observation on the basis of mechanism of injury criteria alone and therefore will turn out to have minor injuries and an ISS of 1, 2 or 3. Inclusion of such patients results in a large left hand skew to the distribution of ISS values. Graph 1 shows this large left hand skew due to the large number of minor injuries. Cases with an ISS of 1, 2 or 3 represent such minor injuries that any deaths that occur in this group can only be explained by co-morbidity or by errors in calculating or recording the ISS. For the latter two reasons all HCISS cases less than 4 were deleted from the revised data set. The deletion of all HCISS cases  $< 4$  resulted in a revised data set of 7,364 cases.



**Graph 1:** The distribution of all HCISS cases demonstrates a strong left hand skew.



**Graph 2:** The distribution following deletion of all HCISS cases < 4 still leaves a strong left hand skew.

### Step 5.2

A HCISS value of 81 represents an injury which is not classifiable, these cases were deleted from the data set. 72 cases were identified with an HCISS=81. The refined data set = 7,292 cases.

### Step 5.3

Ten cases where HCISS= 75 and dl=1 were found (table 1). All of these cases were deleted. The refined data set = 7,282.

**Case Summaries**

	ISS	HCISS	GCS	SBP	RR	AGE	DL	NUM
1	75.00	75.00	14.00	72.00	28.00	39.00	1.00	1031.00
2	4.00	75.00	15.00	136.00	20.00	35.00	1.00	1034.00
3	1.00	75.00	15.00	136.00	24.00	46.00	1.00	1124.00
4	75.00	75.00	3.00	.00	.00	16.00	1.00	3823.00
5	75.00	75.00	4.00	.00	1.00	55.00	1.00	6814.00
6	75.00	75.00	15.00	137.00	38.00	18.00	1.00	7356.00
7	75.00	75.00	15.00	120.00	20.00	22.00	1.00	8122.00
8	75.00	75.00	3.00	121.00	1.00	19.00	1.00	8360.00
9	25.00	75.00	3.00	.00	1.00	20.00	1.00	9052.00
10	10.00	75.00	15.00	136.00	28.00	41.00	1.00	10492.00
Total N	10	10	10	10	10	10	10	10

a. Limited to first 100 cases.

**Table 1**

### Step 5.4

There have been no reported cases of patients surviving with an ISS of greater than 50 although the possibility of this remains as Trauma Units become more skilled at treating patients with severe injuries. Only one patient was found to have an ISS > 50 with an outcome = 1 (table 2). The values of the associated variables (i.e. all other variables within normal range) suggest that there was probably an error in recording the outcome variable. This case was deleted. The revised data set size = 7,281.

**Case Summaries**

	GCS	SBP	RR	HCISS	AGE	DL	NUM	ISS
1	15.00	136.00	20.00	59.00	20.00	1.00	1110.00	59.00
Total N	1	1	1	1	1	1	1	1

a. Limited to first 100 cases.

**Table 2**

**Step 6: Deletion of cases with inaccuracies by variable.**

Inaccuracies in the variables can be determined in two ways. Firstly by identifying values which are outside the anticipated range for the physiological variable (i.e. systolic blood pressure, respiratory rate) or parameter (ISS or GCS). The second method of identifying inaccuracies is to compare the value of all the variables together and then to look for physiological inconsistencies.

**Identification Of Values Outside The Normal Range.**

**Step 6.1 The GCS Variable.**

Examination of the distribution of GCS values showed that some typographical errors had occurred with GCS values of greater than 15 being recorded. All GCS values greater than 15 were deleted from the data set (n=18). The revised data set = 7,263 after deletion of GCS > 15. A single case was identified (table 3) where the GCS was recorded as less than 3. This case was deleted. The revised data set after deletion = 7,262.

**Case Summaries**

	ISS	GCS	SBP	RR	HCISS	AGE	DL	NUM
1	9.00	2.00	146.00	18.00	4.00	25.00	1.00	757.00
Total N	1	1	1	1	1	1	1	1

<sup>a</sup>Limited to first 100 cases.

Table 3

### **Step 6.2 Respiratory Rate Variable :- RR**

A precise upper limit for respiratory rate is not clearly defined compared to the GCS. It is unlikely however that an adult patient could maintain ventilation with a rate above 60 breaths per minute. Examination of the distribution of all patients with a respiratory rate greater than 40 per minute is shown in table 4. This demonstrated that only two cases had a RR which was just over (i.e. defined as being less than 65) the cut-off value of 60 :- (number 44, RR=64 and number 45, RR=62). Using a cut-off value of sixty would not result in loss of a large number of borderline cases. All cases with a RR over 60 per minute were deleted (n=32). The revised data set = 7,230.

Case Summaries <sup>a</sup>

	ISS	HCISS	GCS	SBP	RR	AGE	DL	NUM
1	16.00	16.00	15.00	68.00	99.00	18.00	1.00	69.00
2	17.00	17.00	7.00	999.00	999.00	55.00	1.00	136.00
3	35.00	25.00	3.00	999.00	999.00	20.00	.00	161.00
4	25.00	25.00	10.00	999.00	999.00	20.00	1.00	211.00
5	14.00	9.00	15.00	157.00	96.00	20.00	1.00	444.00
6	17.00	12.00	7.00	131.00	44.00	27.00	1.00	647.00
7	10.00	5.00	15.00	128.00	52.00	24.00	1.00	1226.00
8	10.00	10.00	15.00	177.00	44.00	22.00	1.00	1812.00
9	25.00	29.00	6.00	122.00	45.00	93.00	.00	2201.00
10	16.00	16.00	15.00	108.00	42.00	18.00	1.00	2448.00
11	18.00	18.00	15.00	138.00	48.00	28.00	1.00	3046.00
12	9.00	9.00	15.00	144.00	44.00	18.00	1.00	3145.00
13	18.00	18.00	3.00	32.00	58.00	17.00	1.00	3148.00
14	18.00	25.00	15.00	121.00	91.00	19.00	1.00	4150.00
15	9.00	9.00	15.00	18.00	66.00	58.00	1.00	4152.00
16	43.00	43.00	11.00	146.00	44.00	65.00	.00	4338.00
17	25.00	25.00	15.00	132.00	48.00	22.00	1.00	4670.00
18	5.00	5.00	15.00	142.00	98.00	22.00	1.00	4877.00
19	6.00	6.00	15.00	131.00	86.00	19.00	1.00	4808.00
20	19.00	19.00	15.00	40.00	44.00	22.00	1.00	5051.00
21	16.00	25.00	12.00	175.00	60.00	21.00	1.00	5178.00
22	5.00	5.00	13.00	129.00	89.00	61.00	1.00	5479.00
23	29.00	29.00	13.00	144.00	44.00	45.00	.00	5765.00
24	34.00	34.00	8.00	128.00	60.00	47.00	1.00	5807.00
25	29.00	29.00	15.00	156.00	52.00	27.00	1.00	5813.00
26	4.00	4.00	13.00	139.00	91.00	43.00	1.00	6753.00
27	25.00	25.00	15.00	170.00	52.00	31.00	1.00	7091.00
28	16.00	16.00	15.00	82.00	80.00	37.00	1.00	7091.00
29	10.00	10.00	15.00	110.00	48.00	26.00	1.00	7243.00
30	29.00	38.00	5.00	152.00	69.00	50.00	.00	7389.00
31	25.00	25.00	15.00	184.00	48.00	43.00	1.00	7532.00
32	10.00	10.00	13.00	149.00	100.00	55.00	1.00	7734.00
33	5.00	5.00	15.00	99.00	86.00	34.00	1.00	7953.00
34	25.00	25.00	15.00	109.00	60.00	37.00	1.00	8062.00
35	26.00	17.00	13.00	131.00	44.00	21.00	1.00	8642.00
36	16.00	16.00	12.00	92.00	48.00	67.00	1.00	8983.00
37	17.00	26.00	7.00	220.00	44.00	30.00	1.00	9043.00
38	9.00	9.00	15.00	98.00	92.00	34.00	1.00	9046.00
39	9.00	9.00	15.00	204.00	44.00	50.00	1.00	9095.00
40	25.00	9.00	13.00	170.00	92.00	22.00	1.00	9175.00
41	13.00	13.00	15.00	170.00	70.00	62.00	1.00	9781.00
42	6.00	11.00	14.00	159.00	78.00	46.00	1.00	9784.00
43	10.00	10.00	15.00	132.00	77.00	25.00	1.00	9817.00
44	5.00	5.00	15.00	134.00	64.00	64.00	1.00	9846.00
45	10.00	5.00	15.00	101.00	62.00	24.00	1.00	10069.00
46	4.00	4.00	15.00	136.00	70.00	68.00	1.00	10135.00
47	20.00	20.00	11.00	120.00	48.00	40.00	1.00	10184.00
48	14.00	14.00	9.00	120.00	48.00	21.00	1.00	10189.00
49	11.00	11.00	15.00	117.00	70.00	31.00	1.00	10330.00
50	35.00	35.00	15.00	82.00	70.00	34.00	1.00	10567.00
51	16.00	16.00	9.00	155.00	72.00	43.00	1.00	10840.00
52	9.00	16.00	14.00	124.00	66.00	42.00	1.00	10876.00
53	9.00	9.00	15.00	131.00	77.00	16.00	1.00	11121.00
54	17.00	17.00	15.00	144.00	44.00	77.00	1.00	11444.00
55	5.00	5.00	15.00	131.00	95.00	28.00	1.00	11529.00
56	19.00	19.00	15.00	156.00	70.00	40.00	1.00	11666.00
57	9.00	9.00	15.00	122.00	54.00	40.00	1.00	11685.00
58	22.00	27.00	15.00	110.00	44.00	32.00	1.00	12189.00
59	14.00	14.00	11.00	134.00	60.00	17.00	1.00	12295.00
60	9.00	9.00	15.00	130.00	70.00	20.00	1.00	12574.00
61	9.00	9.00	15.00	144.00	48.00	28.00	1.00	12681.00
62	19.00	26.00	15.00	180.00	48.00	32.00	1.00	12742.00
63	18.00	18.00	15.00	107.00	56.00	24.00	1.00	12747.00
64	17.00	17.00	15.00	155.00	78.00	79.00	1.00	13062.00
65	9.00	9.00	15.00	170.00	46.00	51.00	1.00	13329.00
Total	65	65	65	65	65	65	65	65

a. Limited to first 100 cases.

Table 4

### Step 6.3 Systolic Blood Pressure Variable :- SBP

The upper limit of normal for systolic blood increases with age. The majority of patients attending emergency departments will have some elevation in their blood pressure due to 'stress'. The combination of these two makes an easily identifiable cut-off level between the upper limit of normal and an error in recording the SBP problematic. The cut-off point for systolic blood pressure was made at 250mmHg. Five patients had a systolic blood pressure over 250mmHg. The revised data set following the deletion of these five cases was 7,225.

Case Summaries

	HCISS	GCS	SBP	RR	AGE	DL	NUM
1	26.00	3.00	999.00	1.00	19.00	.00	326.00
2	8.00	14.00	254.00	26.00	71.00	1.00	1624.00
3	9.00	15.00	256.00	32.00	52.00	1.00	5532.00
4	5.00	15.00	253.00	24.00	30.00	1.00	5745.00
5	30.00	7.00	252.00	1.00	46.00	1.00	1787.00
Total N	5	5	5	5	5	5	5

a. Limited to first 100 cases.

Table 5

### Step 6.4: Age Variable.

No inaccuracies were noted with the recorded age values. All patients had an age of less than 100 years.

### Step 7. Group Inaccuracies.

Inaccuracies in the data were also identified by examining the three physiological variables as a group.

### **Step 7.1**

A case with a respiratory rate recorded as zero but with a GCS > 3 would be inconsistent and these cases were identified (table 6) and deleted. A patient could still have a relatively normal blood pressure for a brief period of time before progressing to cardio-respiratory arrest. Cases with a recordable systolic blood pressure were therefore not deleted. 76 cases were deleted in total and the revised data set = 7,149.

### **Step 7.2**

Examination of cases with a respiratory rate of less than 10 revealed some inconsistencies. A patient with a respiratory rate of between 1 and 5 would be expected to be unconscious; consistent with a GCS of 8 or less. All patients with a GCS of greater than 8 and a respiratory rate between 1 and 5 were identified (table 7) and deleted (n=37). The revised data set = 7,112.

Case Summaries <sup>a</sup>

	ISS	HCISS	GCS	SBP	RR	AGE	DL	NUM
1	17.00	26.00	5.00	148.00	.00	19.00	.00	15.00
2	20.00	29.00	4.00	.00	.00	25.00	.00	523.00
3	17.00	17.00	5.00	155.00	.00	20.00	1.00	552.00
4	38.00	75.00	15.00	.00	.00	25.00	.00	766.00
5	50.00	43.00	7.00	82.00	.00	25.00	.00	778.00
6	26.00	26.00	7.00	116.00	.00	17.00	.00	938.00
7	16.00	25.00	5.00	142.00	.00	18.00	.00	1002.00
8	43.00	43.00	7.00	70.00	.00	58.00	.00	1063.00
9	57.00	48.00	4.00	50.00	.00	36.00	.00	1077.00
10	10.00	10.00	7.00	100.00	.00	25.00	1.00	1114.00
11	16.00	25.00	5.00	96.00	.00	20.00	.00	1196.00
12	13.00	13.00	7.00	134.00	.00	26.00	1.00	1218.00
13	17.00	16.00	8.00	86.00	.00	36.00	.00	1675.00
14	20.00	24.00	4.00	116.00	.00	30.00	1.00	1890.00
15	16.00	9.00	9.00	211.00	.00	20.00	1.00	1891.00
16	25.00	16.00	15.00	120.00	.00	27.00	1.00	1906.00
17	5.00	41.00	6.00	90.00	.00	16.00	.00	1971.00
18	16.00	25.00	6.00	124.00	.00	17.00	.00	2100.00
19	16.00	16.00	7.00	132.00	.00	35.00	1.00	2132.00
20	16.00	16.00	6.00	166.00	.00	20.00	1.00	2153.00
21	26.00	21.00	10.00	70.00	.00	35.00	.00	2214.00
22	16.00	25.00	4.00	130.00	.00	20.00	.00	2246.00
23	26.00	26.00	4.00	64.00	.00	73.00	.00	2282.00
24	41.00	50.00	9.00	80.00	.00	30.00	.00	2319.00
25	38.00	38.00	7.00	116.00	.00	22.00	1.00	2393.00
26	25.00	25.00	5.00	94.00	.00	26.00	.00	2434.00
27	16.00	26.00	5.00	193.00	.00	34.00	.00	2690.00
28	35.00	35.00	7.00	181.00	.00	71.00	.00	2812.00
29	22.00	17.00	6.00	146.00	.00	23.00	1.00	2879.00
30	16.00	16.00	15.00	165.00	.00	50.00	1.00	2935.00
31	26.00	17.00	6.00	95.00	.00	19.00	.00	2997.00
32	9.00	9.00	6.00	105.00	.00	40.00	1.00	3201.00
33	26.00	26.00	5.00	102.00	.00	19.00	.00	3261.00
34	16.00	25.00	11.00	213.00	.00	23.00	.00	3476.00
35	16.00	16.00	4.00	138.00	.00	53.00	1.00	3525.00
36	25.00	25.00	5.00	181.00	.00	21.00	1.00	3631.00
37	26.00	25.00	6.00	142.00	.00	47.00	1.00	3884.00
38	9.00	9.00	9.00	114.00	.00	18.00	1.00	3959.00
39	29.00	29.00	5.00	88.00	.00	30.00	1.00	4122.00
40	22.00	22.00	8.00	120.00	.00	21.00	1.00	4128.00
41	11.00	11.00	14.00	160.00	.00	70.00	1.00	4216.00
42	29.00	25.00	7.00	73.00	.00	34.00	.00	4299.00
43	4.00	4.00	8.00	179.00	.00	16.00	1.00	4357.00
44	9.00	9.00	11.00	82.00	.00	30.00	1.00	4486.00
45	14.00	14.00	8.00	145.00	.00	43.00	1.00	4745.00
46	29.00	24.00	7.00	92.00	.00	38.00	1.00	4754.00
47	26.00	35.00	8.00	137.00	.00	20.00	1.00	4852.00
48	10.00	10.00	12.00	216.00	.00	45.00	1.00	4904.00
49	28.00	26.00	5.00	106.00	.00	16.00	.00	4970.00
50	21.00	21.00	10.00	156.00	.00	34.00	1.00	5009.00
51	30.00	30.00	6.00	101.00	.00	41.00	.00	5066.00
52	4.00	4.00	15.00	150.00	.00	36.00	1.00	5108.00
53	26.00	5.00	4.00	112.00	.00	23.00	1.00	5138.00
54	26.00	26.00	6.00	112.00	.00	26.00	.00	5238.00
55	17.00	16.00	15.00	104.00	.00	20.00	.00	5402.00
56	13.00	8.00	5.00	117.00	.00	28.00	1.00	5421.00
57	9.00	9.00	4.00	76.00	.00	20.00	1.00	5513.00
58	75.00	75.00	6.00	.00	.00	25.00	.00	5629.00
59	34.00	43.00	6.00	138.00	.00	19.00	1.00	5729.00
60	75.00	75.00	6.00	.00	.00	29.00	.00	5734.00
61	41.00	50.00	13.00	70.00	.00	67.00	.00	5927.00
62	17.00	17.00	6.00	60.00	.00	36.00	1.00	6043.00
63	17.00	16.00	4.00	157.00	.00	30.00	1.00	6052.00
64	25.00	34.00	6.00	193.00	.00	27.00	1.00	6053.00
65	20.00	20.00	15.00	120.00	.00	52.00	1.00	6097.00
66	26.00	26.00	7.00	112.00	.00	36.00	.00	6120.00
67	26.00	26.00	6.00	108.00	.00	16.00	.00	6165.00
68	29.00	29.00	10.00	66.00	.00	82.00	.00	6179.00
69	17.00	12.00	6.00	.00	.00	35.00	1.00	6190.00
70	16.00	16.00	15.00	147.00	.00	21.00	1.00	6726.00
71	34.00	34.00	15.00	.00	.00	39.00	1.00	6750.00
72	16.00	16.00	6.00	.00	.00	20.00	.00	7084.00
73	4.00	4.00	15.00	122.00	.00	35.00	1.00	9782.00
74	26.00	17.00	14.00	192.00	.00	65.00	.00	11105.00
75	9.00	9.00	15.00	144.00	.00	20.00	1.00	12801.00
76	34.00	34.00	9.00	58.00	.00	36.00	.00	13193.00
Total	N	76	76	76	76	76	76	76

a. Limited to first 100 cases.

Table 6

**Case Summaries**

	HCISS	GCS	SBP	RR	AGE	NUM	DS
1	26.00	12.00	172.00	1.00	19.00	.00	454.00
2	10.00	11.00	.00	4.00	17.00	.00	3551.00
3	9.00	15.00	118.00	2.00	31.00	1.00	4134.00
4	9.00	15.00	85.00	1.00	31.00	1.00	4281.00
5	16.00	10.00	97.00	1.00	47.00	.00	5770.00
6	41.00	15.00	72.00	1.00	30.00	.00	5876.00
7	29.00	10.00	105.00	1.00	35.00	1.00	6017.00
8	20.00	10.00	130.00	1.00	24.00	1.00	6090.00
9	18.00	13.00	104.00	1.00	31.00	1.00	6198.00
10	25.00	12.00	54.00	1.00	46.00	1.00	6270.00
11	5.00	15.00	142.00	1.00	22.00	1.00	6487.00
12	33.00	9.00	71.00	1.00	59.00	.00	6696.00
13	9.00	9.00	85.00	1.00	38.00	1.00	6850.00
14	29.00	10.00	156.00	1.00	55.00	1.00	7133.00
15	20.00	11.00	110.00	1.00	28.00	1.00	7264.00
16	14.00	15.00	136.00	1.00	30.00	1.00	7287.00
17	33.00	15.00	80.00	1.00	19.00	.00	7364.00
18	33.00	11.00	121.00	1.00	40.00	1.00	7579.00
19	29.00	12.00	.00	1.00	35.00	.00	8045.00
20	21.00	9.00	153.00	1.00	21.00	.00	8149.00
21	9.00	15.00	85.00	1.00	48.00	1.00	8173.00
22	22.00	11.00	.00	1.00	65.00	.00	8332.00
23	14.00	14.00	128.00	1.00	30.00	1.00	9453.00
24	18.00	15.00	.00	1.00	17.00	1.00	9831.00
25	41.00	15.00	152.00	1.00	24.00	.00	9880.00
26	30.00	11.00	.00	1.00	54.00	.00	10202.00
27	36.00	12.00	134.00	1.00	50.00	.00	10206.00
28	8.00	9.00	169.00	1.00	18.00	1.00	10831.00
29	75.00	15.00	.00	1.00	38.00	.00	11162.00
30	16.00	15.00	.00	4.00	25.00	.00	11202.00
31	17.00	10.00	164.00	1.00	35.00	1.00	11455.00
32	25.00	15.00	.00	1.00	20.00	.00	11660.00
33	17.00	15.00	.00	1.00	29.00	.00	11697.00
34	26.00	10.00	185.00	1.00	75.00	.00	11744.00
35	50.00	14.00	.00	1.00	35.00	.00	12225.00
36	22.00	13.00	108.00	1.00	18.00	.00	12341.00
37	25.00	15.00	.00	1.00	19.00	.00	12958.00
Total N	37	37	37	37	37	37	37

a. Limited to first 100 cases.

Table 7

### Step 7.3

Patients with a systolic blood pressure equal to zero are clinically dead and therefore would have a GCS of 3. Cases with a SBP = 0 and a GCS > 3 were identified and deleted (table 8). The revised data set = 7,077.

Case Summary<sup>a</sup>

	ISS	HCISS	GCS	SBP	RR	AGE	DL	NUM
1	25.00	26.00	11.00	.00	16.00	32.00	.00	1.00
2	16.00	25.00	4.00	.00	1.00	25.00	.00	4.00
3	34.00	41.00	4.00	.00	1.00	30.00	.00	201.00
4	25.00	25.00	12.00	.00	16.00	39.00	1.00	338.00
5	32.00	32.00	14.00	.00	30.00	21.00	1.00	1262.00
6	16.00	16.00	13.00	.00	32.00	21.00	.00	2741.00
7	18.00	18.00	12.00	.00	16.00	27.00	1.00	2795.00
8	4.00	9.00	14.00	.00	32.00	28.00	1.00	2906.00
9	9.00	9.00	9.00	.00	24.00	16.00	1.00	2968.00
10	25.00	18.00	14.00	.00	24.00	18.00	.00	3130.00
11	26.00	26.00	11.00	.00	10.00	25.00	.00	4500.00
12	17.00	26.00	14.00	.00	22.00	44.00	.00	4643.00
13	14.00	17.00	15.00	.00	16.00	43.00	1.00	4764.00
14	43.00	43.00	13.00	.00	28.00	20.00	.00	6440.00
15	59.00	50.00	12.00	.00	24.00	32.00	.00	6914.00
16	10.00	10.00	15.00	.00	20.00	18.00	1.00	6996.00
17	9.00	9.00	15.00	.00	22.00	17.00	1.00	7094.00
18	17.00	17.00	10.00	.00	24.00	20.00	1.00	7134.00
19	17.00	17.00	15.00	.00	32.00	30.00	1.00	8046.00
20	17.00	17.00	14.00	.00	18.00	24.00	.00	8630.00
21	17.00	17.00	7.00	.00	1.00	31.00	.00	8661.00
22	16.00	16.00	8.00	.00	14.00	22.00	1.00	8709.00
23	51.00	51.00	15.00	.00	24.00	17.00	.00	9047.00
24	51.00	51.00	6.00	.00	1.00	71.00	.00	9381.00
25	11.00	11.00	15.00	.00	16.00	64.00	1.00	9773.00
26	14.00	14.00	15.00	.00	20.00	17.00	1.00	9810.00
27	26.00	26.00	7.00	.00	1.00	17.00	.00	10369.00
28	13.00	13.00	15.00	.00	20.00	20.00	1.00	10646.00
29	25.00	25.00	7.00	.00	1.00	18.00	.00	11554.00
30	30.00	30.00	8.00	.00	20.00	24.00	1.00	11785.00
31	16.00	9.00	15.00	.00	24.00	25.00	1.00	12290.00
32	9.00	9.00	15.00	.00	24.00	19.00	.00	12507.00
33	18.00	18.00	7.00	.00	24.00	22.00	1.00	12564.00
34	14.00	14.00	15.00	.00	24.00	25.00	1.00	13014.00
35	16.00	9.00	15.00	.00	24.00	27.00	1.00	13258.00
Total N	35	35	35	35	35	35	35	35

a. Limited to first 100 cases.

Table 8

### Step 7.4

Patients with a systolic blood pressure less than 50mmHg would be expected to have a depressed level of consciousness with a GCS of < 15. Cases with a SBP < 50 and a GCS = 15 were identified (table 9) and deleted. The revised data set = 7,069.

Case Summary<sup>a</sup>

	HCISS	GCS	SBP	RR	AGE	NUM	DS
1	10.00	15.00	28.00	28.00	31.00	1.00	358.00
2	25.00	15.00	20.00	34.00	30.00	1.00	1363.00
3	19.00	15.00	40.00	28.00	17.00	.00	2485.00
4	19.00	15.00	40.00	44.00	22.00	1.00	5051.00
5	35.00	15.00	40.00	36.00	18.00	.00	5395.00
6	35.00	15.00	48.00	36.00	17.00	.00	5655.00
7	20.00	15.00	41.00	36.00	40.00	1.00	7240.00
8	25.00	15.00	48.00	40.00	45.00	1.00	11483.00
Total N	8	8	8	8	8	8	8

a. Limited to first 100 cases.

Table 9

### Step 8.

#### Further Preparation of the Mechanism of Injury Variables

#### Step 8.1

1,010 cases were recoded as value 4 (traumatic arrest) in the MOI1 variable group. 1,008 of these cases had a corresponding recoded value of 2 in the MOI2 variable group and the final two cases in the MOI2 cell were blank. All of the cases except for the 2 blank cells (case numbers 8820 and 9776) were recoded 2 in a revised mechanism of injury variable group (MOI3). The cases with the blank second cell were recoded as 3 in the MOI3 group.

**Step 8.2**

Fifteen cases were found to have a value of 1 (blunt trauma) in the MOI1 group and a value of 2 (penetrating trauma) in the MOI2 group (table 10). No changes were made and the MOI3 variable was coded as 1.

**Case Summaries**

	HCISS	GCS	SBP	RR	AGE	DL	NUM	MOI1	MOI2	MOI1REV	moi1 =1 and moi2 =2 (FILTER)
1	9.00	15.00	134.00	18.00	34.00	1.00	121.00	1.00	2.00	1.00	Selected
2	4.00	15.00	116.00	18.00	34.00	1.00	2078.00	1.00	2.00	1.00	Selected
3	5.00	3.00	162.00	1.00	19.00	1.00	5922.00	1.00	2.00	1.00	Selected
4	16.00	3.00	136.00	1.00	40.00	1.00	6550.00	1.00	2.00	1.00	Selected
5	5.00	10.00	131.00	20.00	33.00	1.00	6705.00	1.00	2.00	1.00	Selected
6	26.00	6.00	108.00	14.00	25.00	.00	7013.00	1.00	2.00	1.00	Selected
7	24.00	15.00	124.00	12.00	23.00	1.00	8021.00	1.00	2.00	1.00	Selected
8	4.00	15.00	128.00	20.00	31.00	1.00	8387.00	1.00	2.00	1.00	Selected
9	5.00	14.00	168.00	22.00	33.00	1.00	9315.00	1.00	2.00	1.00	Selected
10	8.00	15.00	155.00	26.00	40.00	1.00	10117.00	1.00	2.00	1.00	Selected
11	10.00	7.00	141.00	1.00	33.00	1.00	10892.00	1.00	2.00	1.00	Selected
12	4.00	15.00	103.00	18.00	40.00	1.00	11249.00	1.00	2.00	1.00	Selected
13	10.00	15.00	112.00	20.00	19.00	1.00	12807.00	1.00	2.00	1.00	Selected
14	5.00	6.00	133.00	24.00	38.00	1.00	12808.00	1.00	2.00	1.00	Selected
15	5.00	15.00	133.00	18.00	21.00	1.00	12819.00	1.00	2.00	1.00	Selected
Total N	15	15	15	15	15	15	15	15	15	15	15

<sup>a</sup>. Limited to first 100 cases.

Table 10

**Step 8.3**

Twenty cases were found to have a value of 2 in the MOI1 group and a value 1 in the MOI2 group. These results are shown in table 11 (page 46). No changes were made and the MOI3 variable was coded as 2.

Case Summaries

	HCISS	GCS	SBP	RR	AGE	DL	NUM	MOI1	MOI2	MOHREV	moi1 =2 and moi2 = 1 (FILTER)
1	9.00	15.00	106.00	18.00	50.00	1.00	687.00	2.00	1.00	2.00	Selected
2	4.00	15.00	102.00	20.00	19.00	1.00	789.00	2.00	1.00	2.00	Selected
3	14.00	15.00	142.00	16.00	50.00	1.00	1193.00	2.00	1.00	2.00	Selected
4	6.00	15.00	152.00	28.00	18.00	1.00	1212.00	2.00	1.00	2.00	Selected
5	50.00	11.00	68.00	24.00	42.00	.00	1334.00	2.00	1.00	2.00	Selected
6	20.00	3.00	120.00	6.00	19.00	1.00	1359.00	2.00	1.00	2.00	Selected
7	5.00	15.00	138.00	20.00	18.00	1.00	2504.00	2.00	1.00	2.00	Selected
8	9.00	15.00	131.00	20.00	23.00	1.00	2740.00	2.00	1.00	2.00	Selected
9	9.00	15.00	145.00	20.00	28.00	1.00	4613.00	2.00	1.00	2.00	Selected
10	75.00	3.00	.00	.00	46.00	.00	6004.00	2.00	1.00	2.00	Selected
11	19.00	14.00	121.00	10.00	46.00	.00	7202.00	2.00	1.00	2.00	Selected
12	5.00	4.00	142.00	18.00	28.00	1.00	9429.00	2.00	1.00	2.00	Selected
13	36.00	14.00	140.00	24.00	30.00	1.00	9553.00	2.00	1.00	2.00	Selected
14	25.00	3.00	164.00	20.00	19.00	.00	10655.00	2.00	1.00	2.00	Selected
15	4.00	15.00	158.00	28.00	29.00	1.00	10899.00	2.00	1.00	2.00	Selected
16	5.00	13.00	159.00	24.00	30.00	1.00	11168.00	2.00	1.00	2.00	Selected
17	21.00	3.00	.00	1.00	36.00	.00	12105.00	2.00	1.00	2.00	Selected
18	8.00	15.00	169.00	24.00	34.00	1.00	12333.00	2.00	1.00	2.00	Selected
19	4.00	15.00	133.00	22.00	36.00	1.00	12397.00	2.00	1.00	2.00	Selected
20	12.00	15.00	108.00	22.00	38.00	1.00	12491.00	2.00	1.00	2.00	Selected
Total N	20	20	20	20	20	20	20	20	20	20	20

a. Limited to first 100 cases.

Table 11

### Step 8.4

95 cases were found to have a value of 3 in the MOI1 group and a value of 1 in the MOI2 group. All of these cases were recoded as 1 in the MOI3 variable group.

The MOI3 variable was created by combining data from the MOI1 and MOI2 variables. The recoding was performed using a SAS program (shown below) after pasting the SPSS data set onto the SAS window. The data set was then exported and converted back into an SPSS data file.

SAS program used for creating the MOI3 variable.

```
DATA USCFINAL;
INPUT ISS HCISS GCS SBP RR AGE NUM DS moi1 moi2 moi3;

IF moi1 = 3 and moi2 = 1 then moi3 = 1;
IF moi1 = 3 and moi2 = 2 then moi3 = 2;
IF moi1 = 3 and moi2 = . then moi3 = 3;
IF moi1 = 3 and moi2 = 4 then moi3 = 3;
IF moi1 = 1 and moi2 = 2 then moi3 = 1;
IF moi1 = 1 and moi2 = . then moi3 = 1;
```

```

IF moi1 = 1 and moi2 = 3 then moi3 = 1;
IF moi1 = 1 and moi2 = 4 then moi3 = 1;
IF moi1 = 2 and moi2 = 1 then moi3 = 2;
IF moi1 = 2 and moi2 = . then moi3 = 2;
IF moi1 = 2 and moi2 = 3 then moi3 = 2;
IF moi1 = 2 and moi2 = 4 then moi3 = 2;
IF moi1 = 4 and moi2 = 2 then moi3 = 2;
IF moi1 = 4 and moi2 = . then moi3 = 3;
DATALINES;
;
RUN;

```

## Section 2: Composition of the Revised Data Set.

Final number of cases in the revised data set = 7,069.

### 1. The Revised Mechanism of Injury Variable (MOI3)

Number of blunt trauma cases = 4,172

(fifteen of these cases also had a penetrating injury).

Number of deaths in the blunt trauma group = 398

Survival rate in the blunt group = 90.5%

Number of penetrating trauma cases= 2,483

(20 of these cases also had a blunt injury).

Number of deaths in the penetrating trauma group = 481

Survival rate in the penetrating group = 80.6%

Number of cases unclassifiable = 396

Number of deaths in the unclassifiable group = 13

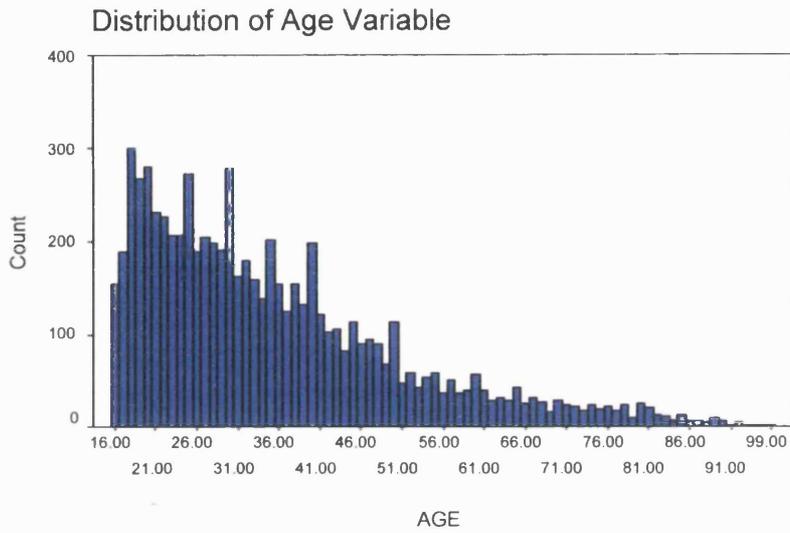
Survival rate in the unclassifiable group = 96.7%.

Number of cases where data was not recorded in either the MOI1 or MOI2 cell = 18. There was one death in this subgroup.

Survival rate = 94.4%.

## The Age Variable

The Age distribution in the revised data set is shown in Graph 3 below.



**Graph 3**

The mean Age and standard deviation is shown in Table 12 below.

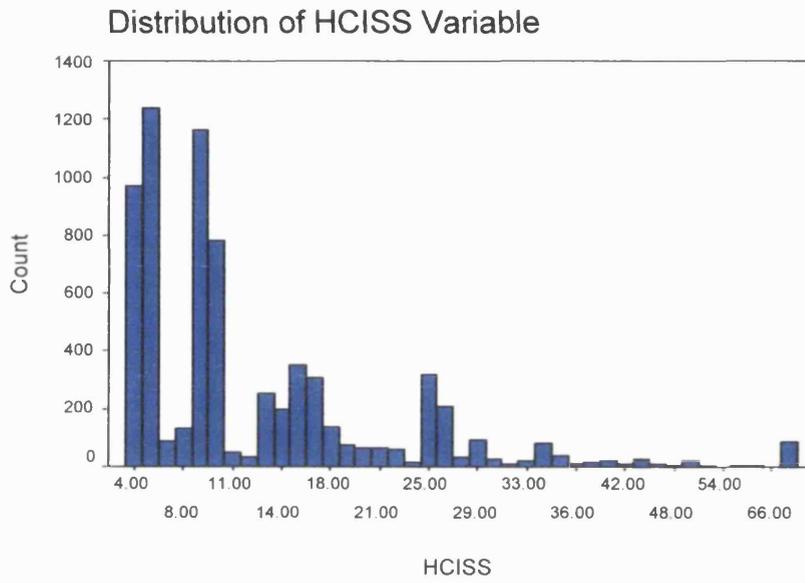
**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	7069	16.00	99.00	35.3334	15.78480
Valid N (listwise)	7069				

Table 12

## The HCISS Variable

The HCISS distribution in the revised data set is shown in Graph 4 below.



**Graph 4**

The mean HCISS and standard deviation is shown in Table 13 below.

**Descriptive Statistics**

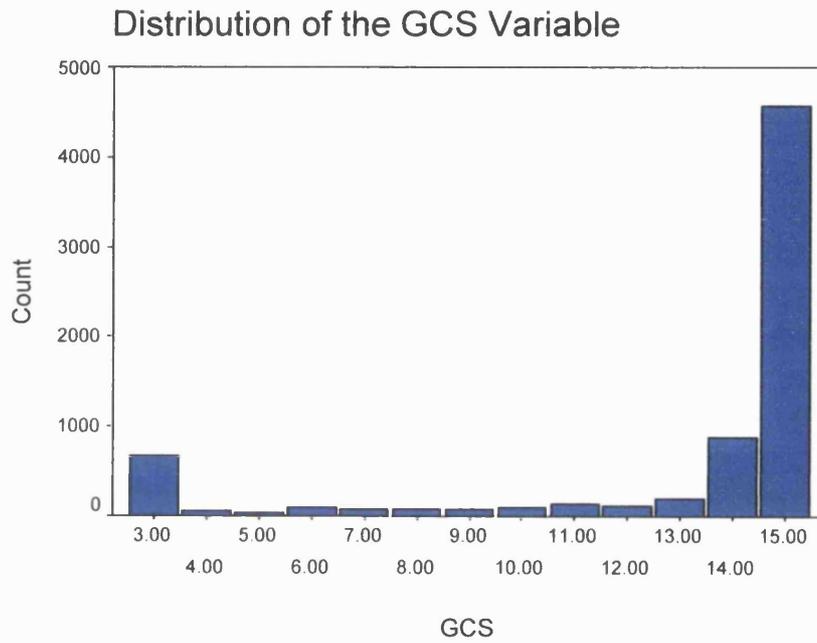
	N	Minimum	Maximum	Mean	Std. Deviation
HCISS	7069	4.00	75.00	13.0648	11.19991
Valid N (listwise)	7069				

Table 13



## The GCS Variable

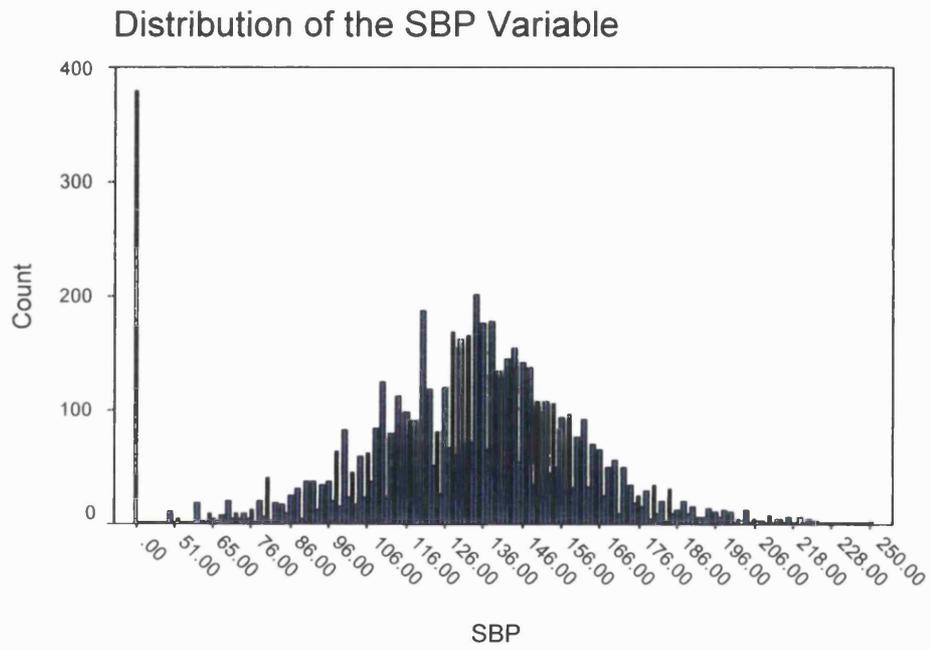
The GCS distribution in the revised data set is shown in Graph 5.



**Graph 5**

## The Systolic Blood Pressure Variable

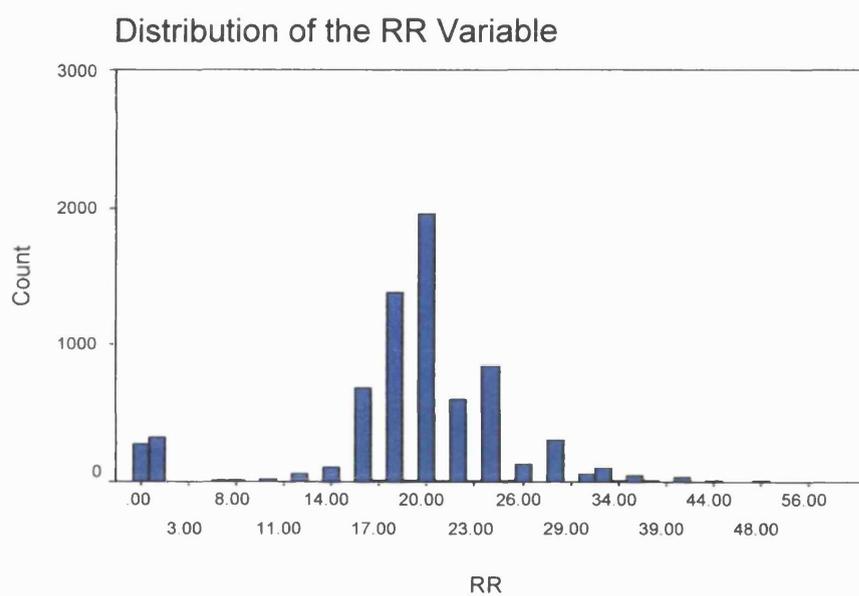
The SBP (systolic blood pressure) distribution in the revised data set is shown in graph 6.



**Graph 6**

## The Respiratory Rate Variable

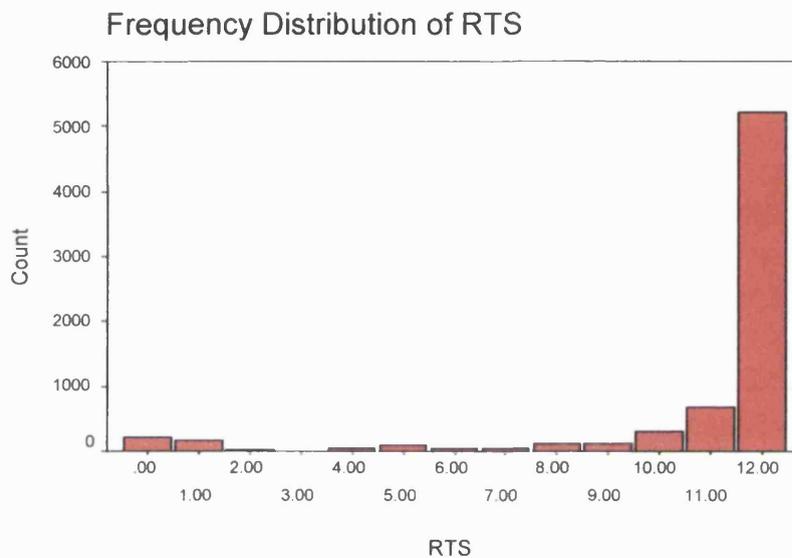
The RR variable (respiratory rate) distribution in the revised data set is shown in Graph 7.



**Graph 7**

## The Revised Trauma Score Variable

The frequency distribution of the triage Revised Trauma Score (RTS) is shown in graph 8.



**Graph 8**

The mean RTS and standard deviation is shown in Table 14 below.

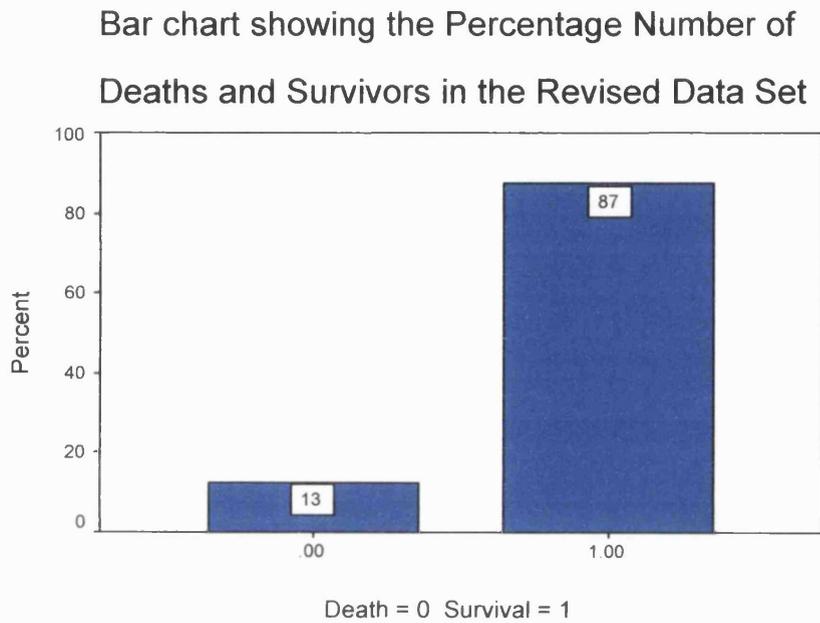
**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
RTS	7069	.00	12.00	10.8290	2.87945
Valid N (listwise)	7069				

Table 14

## Outcome Variable

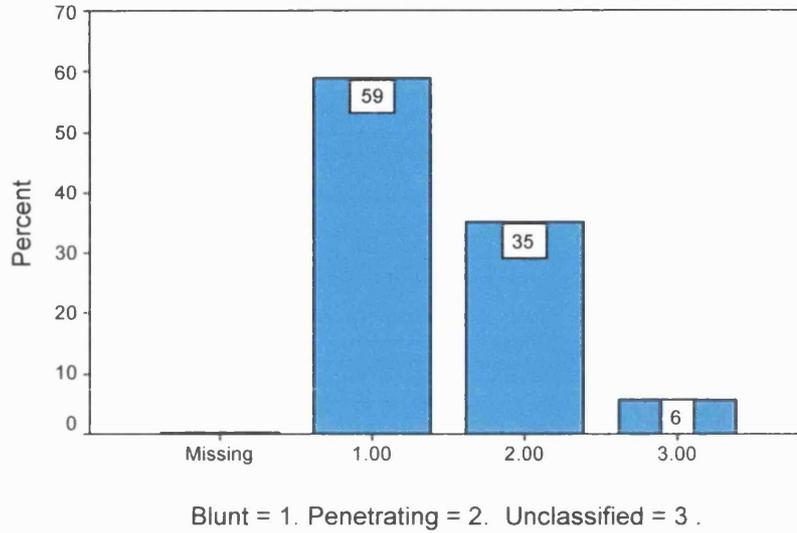
The Survival Rate in the revised data set is shown in graph 9.



**Graph 9**

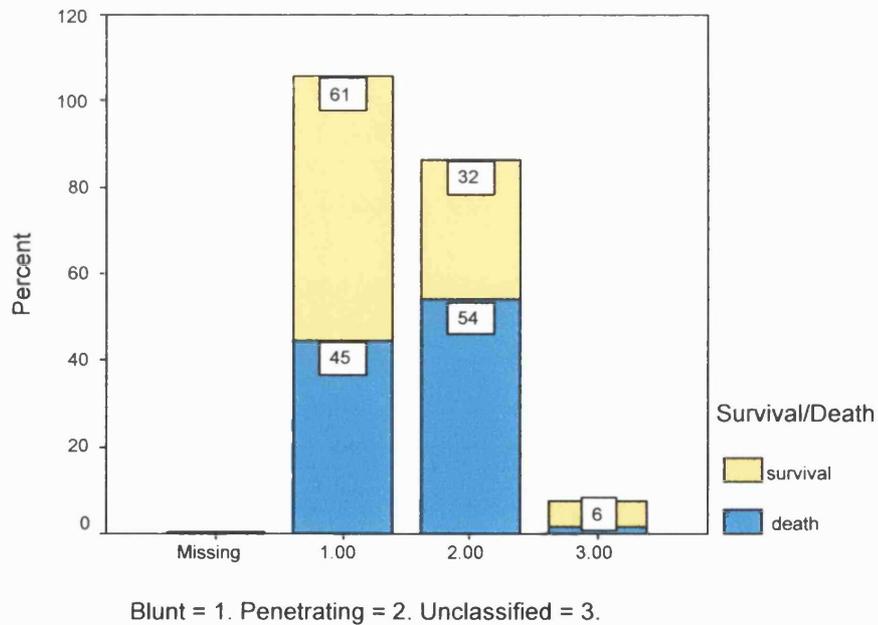
## Blunt and Penetrating Subgroups

Percentage of Blunt and Penetrating Cases in the Revised Data Set



Graph 10

## Survival for Blunt and Penetrating Subgroups



Graph 11

## **Section 4: Discussion**

A rigorous approach was applied to the deletion of cases with inaccurate or missing values. This was due to the impracticality of reviewing all the case notes with incomplete or inaccurate data. The use of missing variable analysis methods to the data set was not appropriate for several reasons. Firstly, each case stands independently of the cases which precede and follow it. Secondly, all of the variables are represented by complete integers which excludes any method for which a mean value is calculated. The third reason is that all of the variables excluding the age variable have some degree of multicollinearity with each other. The combination of physiological values needs to be clinically valid. For example, as discussed in the methodology section, a patient with an extremely low blood pressure or respiratory rate would not be able to maintain a normal conscious level and therefore a Glasgow Coma Score of fifteen would be an inconsistent combination of values.

The reason for deleting all cases with missing variables was as explained in the methodology section, to enable an unbiased comparison between different models. The Hosmer Lemeshow statistic like several other goodness of fit tests has the inherent weakness in that it is affected by the size of the data set. In the case of the Hosmer Lemeshow statistic reducing the size of the data set can change a model which appears to have a poor fit into a model which fits well, particularly if reliance is placed upon the p value.

All cases with an ISS < 4 were deleted from the data set prior to performing Logistic Regression because the probability of death correlates poorly with an AIS < 3 (American Association for Automotive Medicine, 1990; page 2). It is unlikely that death in a patient with an ISS of 3 or less (i.e. three minor injuries [AIS = 1]) would be attributable to the injury(s) itself. Death in a patient with an ISS of 4 (moderate injury [AIS = 2]) however could be attributable to the injury itself particularly in the elderly or in a patient with pre-existing disease. The overall mortality rate in the 'raw' Los Angeles data set was 9.1% (number of deaths = 1,213; number of survivors = 13,381; number of missing cases = 66). The overall mortality rate in the revised data set was 13% (graph 9). The 1990 MTOS study used a database of 80,544 patients (years 1982 to 1987). The data set had an overall mortality of 9% (Champion, 1990a). The MTOS database like the USC data set also included all patients who were clinically 'dead on arrival.' This subgroup of patients accounts for 23% of all the trauma deaths in the revised data set (207/893). Exclusion of patients with an ISS < 4 has not resulted in a major increase in the overall mortality rate in the revised data set. The mortality rate in the revised data set is comparable to the 1990 MTOS database.

The number of recorded penetrating injuries in the revised data set was found to be 35%. This is some what higher than the 21% penetrating injury rate of the MTOS trauma database reported by Champion et al (1990a). Unclassifiable cases represent a subgroup of cases where the mechanism of injury is either not known or of such an unusual type that it does not easily fit into any of the mechanism of injury categories. One could surmise that the

majority of these cases would have been non-penetrating in nature. If the unclassifiable cases were assumed to be blunt, the penetrating injury rate becomes 29%. The latter penetrating injury rate is still substantially higher than the 1990 MTOS rate of 21%. The high penetrating injury rate seen in the USC data set may well be explained by the increased use of firearms over the last decade. The number of cases with missing values in the final mechanism of injury group accounted for a very small proportion of the total number of cases (.0025%: i.e. 18 /7069). As a consequence of this and in order to prevent deletion of any further cases these cases were not deleted from the revised data set (survival rate in the missing group was 5% {one death}). A further reason for not deleting cases with missing values for the mechanism of injury variable was that it was anticipated that this variable would not be used for modeling for the reason given early in this chapter.

Significant variations in the proportion of penetrating injuries within trauma data sets have been reported. Osler et al (1997) used two data sets in the development of NISS (New Injury Severity Score). The Albuquerque Database contained 3,136 cases of which 25% were penetrating injuries. In contrast the trauma database in Portland, Oregon contained 3,449 cases of which only 13% were penetrating injuries. The survival rates for the two data sets were Albuquerque 91%; Portland 93%. Rutledge et al (1998) in the validation of the ICISS model used a data set from the North Carolina trauma registry. The data set contained 7,276 patients after exclusion of incomplete data and outliers. The survival rate in this data set was 96.3%. Patients who were declared dead in the emergency department were included in the registry. The

percentage of patients with blunt and penetrating trauma was not quoted. The mean age and the mean revised trauma score values however were similar to the revised USC data set (North Carolina Data Set, mean Age:-  $42.1 \pm 23.7$ , mean RTS:-  $11.4 \pm 1.8$ ; USC Revised Data Set, mean Age:-  $35.3 \pm 15.8$ , mean RTS:-  $10.8 \pm 2.9$ ).

### **Summary**

The revised USC data set contains 7,069 cases with complete data. Characteristics of the data set, such as mean age and mean RTS are comparable to the North Carolina data set. The data set contains both blunt and penetrating injuries. Deaths include both in-hospital deaths and those that are 'dead on arrival.'

# CHAPTER 4

## THE EVALUATION OF TRAUMA SCORING MODELS USING THE REVISED USC DATA SET

### CONTENTS

	<b>Page Number</b>
<b>Section 1: Introduction</b>	<b>61</b>
<b>Section 2: Aims</b>	<b>62</b>
<b>Section 3: Methodology</b>	<b>63</b>
<b>Section 4: Results</b>	<b>66</b>
<b>Section 5: Discussion</b>	<b>84</b>

## **Section 1: Introduction**

The TRISS model developed by Champion et al (1989, 1990a) utilised three predictor variables; ISS (Injury Severity Score), RTS (Revised Trauma Score) and Age. The RTS in the TRISS model was calculated by multiplying the coded variables by their priori coefficients and then summing all three 'weighted' coded variables. The rationale for this was to correct for differences in the three RTS variables as predictors of death or survival. Limitations of the TRISS model which have been previously discussed in Chapter 2 led Champion et al (1990b) to develop a new model named **A Severity Characterisation Of Trauma** which is usually abbreviated to ASCOT. This model like the TRISS model had three main components; anatomical injury variables, age (a five component interval scale) and lastly three physiological variables. The three physiological variables were those used in the RTS but in the ASCOT model they were used as separate coded variables rather than as a single RTS composite variable. A possible inference from this is that the three coded variables may be superior compared to the composite RTS. Hannan et al (1999) performed a study comparing TRISS and ICISS on a data set of 20,883 patients with blunt injuries. The authors found as an incidental finding that both TRISS and ICISS models could be improved using the three elements of the RTS as separate independent predictor variables. The full characteristics of ISS based models combined with coded RTS variables have not been fully explored.

## **Section 2: Aims**

The aim of this chapter was to evaluate several TRISS type models using the revised USC data set.

Hypotheses to be tested.

1. That a model containing the three components of the RTS + the Injury Severity Score is superior to a model containing one or two components of the RTS + the Injury Severity Score.
2. That a model containing the three components of the RTS is superior than a model containing the composite RTS when both are combined with the Injury Severity Score variable.
3. That the unweighted RTS model is superior to the RTS with MTOS weights model when both are combined with the Injury Severity Score variable.
4. That the addition of age as a continuous variable improves model fit when combined with a model containing HCISS and either composite RTS or component RTS.

## Section 3: Methodology

### Section 3.1 Statistical software.

SPSS version 9.0 was used to perform the logistic regression modelling.

**Section 3.2 Data Set:** The revised USC data set was used.

**Section 3.3** The following models were evaluated:-

<u>Dependent Variable</u>	<u>Predictor Variables</u>
1. D/S*	HCISS
2. D/S	HCISS + coded GCS**
3. D/S	HCISS + coded GCS + coded SBP
4. D/S	HCISS + coded GCS + coded SBP + coded RR
5. D/S	HCISS + coded GCS + coded SBP + coded RR + Age
6. D/S	HCISS + RTS (unweighted)
7. D/S	HCISS + RTS (unweighted) + Age
8. D/S	HCISS + RTS (MTOS weights)
9. D/S	HCISS + RTS (MTOS weights) + Age

\*D/S is the abbreviated form for the death/survival variable.

\*\*The coded values for GCS, SBP and RR were those used in the Revised Trauma Score (Table 1). The triage Revised Trauma Score (abbreviated form:- RTS) is the sum of the three coded variables. The physiological variables are recorded on admission and the RTS is often used as a tool for trauma team activation ('trauma triage'). Coding of the RR (variable) is required in particular because values above and below the normal RR range both indicate physiological derangement.

**Table 1: Coded Values for the RTS**

<u>GCS</u>	<u>SBP</u>	<u>RR</u>	<u>Coded Value</u>
13-15	>89	10-29	4
9-12	76-89	>29	3
6-8	50-75	6-9	2
4-5	1-49	1-5	1
3	0	0	0

### **Section 3.4 Method of Model Selection**

1. Backward LR test (likelihood ratio).

(The calculated probabilities were for survival rather than death).

### **Section 3.5**

#### **Method to Assess for Linearity of the Predictor Variables:**

Plot of Log Odds against predictor Variables.

(a) Plotting the log odds against HCISS.

(b) Plotting the log odds against RTS and controlling for HCISS.

### **Section 3.6 Methods used to assess the model fit.**

(1) Global Goodness of fit test:-

(a) -2 Log Likelihood (abbreviated form:  $-2LL$ )

(2) Tests to assess the degree of explained variation:-

(a) Cox and Snell test.

(b) Nagelkerke test.

(3) Goodness of Fit Tests for Internal Validation\* of the Model.

**Calibration:-** Hosmer Lemeshow chi-squared goodness of fit test.

**Discrimination:-** Area under the ROC curve. Confidence intervals are calculated automatically by the SPSS program, using the *non-parametric assumption* option.

\*Internal validation being validation of the model on the data set upon which it was derived.

## Section 4: Results

The results given below are a modified version of the SPSS output.

### Section 4.1

Model: - Dependent variable: D/S. Predictor variable: HCISS

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.1.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Only the constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.1.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1.HCISS Estimation terminated at iteration number 5 because

Log Likelihood decreased by less than .01 percent.

### Section 4.1.4

-2 Log Likelihood 3238.602

Cox & Snell - R<sup>2</sup> .260

Nagelkerke - R<sup>2</sup> .488

### Section 4.1.5

	Chi-Square	df	Significance
Model	2124.547	1	.0000
Block	2124.547	1	.0000
Step	2124.547	1	.0000

### Section 4.1.6 Hosmer Lemeshow Goodness of Fit Test

HL goodness of fit test = 52.0798 df: 7 p = .0000

### Section 4.1.7 Area under the ROC

ROC = .908 (95% CI: .898 - .919)

### Section 4.1.8

**Table 4.1.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1762	.0053
Constant	5.0111	.1161

### Section 4.2.1

Model:- Dependent variable: D/S.

Predictor variables: HCISS + coded GCS

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.2.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.2.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1.HCISS REGCS Estimation terminated at iteration number 6 because Log Likelihood decreased by less than .01 percent.

### Section 4.2.4

-2 Log Likelihood	2212.865
Cox & Snell - R <sup>2</sup>	.360
Nagelkerke - R <sup>2</sup>	.676

### Section 4.2.5

	Chi-Square	df	Significance
Model	3150.284	2	.0000
Block	3150.284	2	.0000
Step	3150.284	2	.0000

### Section 4.2.6 Hosmer and Lemeshow Goodness of Fit Test

HL goodness of fit test = 21.5294 df: 7 p = .0031

### Section 4.2.7 Area under the ROC

ROC = .957 (95% CI: .950 - .964)

### Section 4.2.8

**Table 3.2.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1355	.0059
REGCS	.9940	.0344
Constant	1.5809	.1549

### Section 4.3.1

Model:- Dependent variable: D/S.

Predictor variable: HCISS + coded GCS + coded SBP

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.3.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Only the constant is included in the model.

Estimation terminated at iteration number 4 because Log Likelihood decreased by less than .01 percent.

### Section 4.3.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1.HCISS REGCS RESBP

Estimation terminated at iteration number 6 because Log Likelihood decreased by less than .01 percent.

### Section 4.3.4

-2 Log Likelihood	2002.377
Cox & Snell - R <sup>2</sup>	.378
Nagelkerke - R <sup>2</sup>	.712

### Section 4.3.6

	Chi-Square	df	Significance
Model	3360.772	3	.0000
Block	3360.772	3	.0000
Step	3360.772	3	.0000

### Section 4.3.6 Hosmer and Lemeshow Goodness of Fit Test

HL goodness of fit test = 21.3366 df: 7 p= .0033

### Section 4.3.7 Area under the ROC

ROC = .963 (95% CI: .957 - .970)

### Section 4.3.8

**Table 4.3.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1392	.0064
REGCS	.7570	.0385
RESBP	.8955	.0757
Constant	-.9187	.3124

No more variables can be deleted or added.

### Section 4.4.1

Model:- Dependent variable: D/S.

Predictor variables:

HCISS + coded GCS + coded SBP + coded RR.

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.4.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Only the constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.4.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1. HCISS REGCS RESBP RERR

Estimation terminated at iteration number 6 because

Log Likelihood decreased by less than .01 percent.

### Section 4.4.4

-2 Log Likelihood	1988.961
Cox & Snell - R <sup>2</sup>	.380
Nagelkerke - R <sup>2</sup>	.714

### Section 4.4.5

	Chi-Square	df	Significance
Model	3374.188	4	.0000
Block	3374.188	4	.0000
Step	3374.188	4	.0000

### Section 4.4.6 Hosmer and Lemeshow Goodness of Fit Test

HL goodness of fit test = 21.5060 df: 7 p = .0031

### Section 4.4.7 Area under the ROC

ROC = .964 (95% CI: .957 - .970)

### Section 4.4.8

**Table 3.4.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1379	.0064
REGCS	.6472	.0490
RESBP	.8144	.0813
RERR	.2482	.0682
Constant	-1.1898	.3345

No more variables can be deleted or added.

### Section 4.5.1

Model:- Dependent variable: D/S.

Predictor variables: -

HCISS + coded GCS + coded SBP + coded RR + Age

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.5.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.5.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1.HCISS REGCS RESBP RERR AGE

Estimation terminated at iteration number 6 because

Log Likelihood decreased by less than .01 percent.

### Section 4.5.4

-2 Log Likelihood 1911.581

Cox & Snell - R<sup>2</sup> .386

Nagelkerke - R<sup>2</sup> .727

### Section 4.5.6

	Chi-Square	df	Significance
Model	3451.568	5	.0000
Block	3451.568	5	.0000
Step	3451.568	5	.0000

### Section 4.5.6 Hosmer and Lemeshow Goodness of Fit

HL goodness of fit test = 20.6868 df: 8 p = .0080

### Section 4.5.7 Area under the ROC

ROC = .968 (95% CI: .962 - .973)

### Section 4.5.8

**Table 3.5.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1390	.0065
REGCS	.6332	.0502
RESBP	.8541	.0812
RERR	.2832	.0691
AGE	-.0295	.0034
Constant	-.2320	.3476

No more variables can be deleted or added.

### Section 4.6.1

Model:- Dependent variable: D/S.

Predictor variable: HCISS + RTS

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.6.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Only the constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.6.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1. HCISS RTS

Estimation terminated at iteration number 6 because

Log Likelihood decreased by less than .01 percent.

### Section 4.6.4

-2 Log Likelihood	2012.657
Cox & Snell - R <sup>2</sup>	.377
Nagelkerke - R <sup>2</sup>	.710

### Section 4.6.5

	Chi-Square	df	Significance
Model	3350.492	2	.0000
Block	3350.492	2	.0000
Step	3350.492	2	.0000

### Section 4.6.6 Hosmer and Lemeshow Goodness of Fit Test

HL goodness of fit test = 32.9552 df: 7 p = .0000

### Section 4.6.7 Area under the ROC

ROC = .964 (95% CI: .957 - .970)

### Section 4.6.8

**Table 3.6.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1374	.0063
RTS	.5349	.0211
Constant	-.8354	.2536

No more variables can be deleted or added.

### Section 4.7.1

Model: Dependent variable: D/S.

Predictor variables: HCISS + RTS + Age

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.7.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Only the constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.7.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1.HCISS RTS AGE

Estimation terminated at iteration number 6 because

Log Likelihood decreased by less than .01 percent.

### Section 4.7.4

-2 Log Likelihood 1934.834

Cox & Snell - R<sup>2</sup> .384

Nagelkerke - R<sup>2</sup> .723

### Section 4.7.5

	Chi-Square	df	Significance
Model	3428.315	3	.0000
Block	3428.315	3	.0000
Step	3428.315	3	.0000

### Section 4.7.6 Hosmer and Lemeshow Goodness of Fit Test

HL goodness of fit test = 28.2686 df: 8 p = .0004

### Section 4.7.7 Area under the ROC

ROC = .967 (95% CI: .961 - .973)

### Section 4.7.8

**Table 4.7.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1377	.0064
RTS	.5480	.0212
AGE	-.0295	.0033
Constant	.2028	.2749

No more variables can be deleted or added.

### Section 4.8.1

Model:- Dependent variable: D/S.

Predictor variables: HCISS + RTS (weighted)

The weights used were those developed from the MTOS trauma data base (Boyd et al, 1987).

Coded GCS weight (coefficient) = 0.9368

Coded SBP weight (coefficient) = 0.7326

Coded RR weight (coefficient) = 0.8724

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

### Section 4.8.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149\*

\* Only the constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.8.3

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1. HCISS WRTS (weighted RTS)

Estimation terminated at iteration number 6 because

Log Likelihood decreased by less than .01 percent.

### Section 4.8.4

-2 Log Likelihood 2001.908

Cox & Snell - R<sup>2</sup> .378

Nagelkerke - R<sup>2</sup> .712

### Section 4.8.5

	Chi-Square	df	Significance
Model	3361.241	2	.0000
Block	3361.241	2	.0000
Step	3361.241	2	.0000

### Section 4.8.6 Hosmer and Lemeshow Goodness of Fit Test

HL goodness of fit test = 22.5708 df: 7 p = .0020

### Section 4.8.7 Area under the ROC

ROC = .963 (95% CI: .957 - .970)

## Section 4.8.8

**Table 3.8.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1363	.0063
WRTS	.7850	.0295
Constant	-.5077	.2302

No more variables can be deleted or added.

## Section 4.9.1

Model:- Dependent variable: D/S

Predictor variables: HCISS + RTS (weighted) + Age

The weights used were those developed from the MTOS trauma data base (Boyd et al, 1987).

Coded GCS weight (coefficient) = 0.9368

Coded SBP weight (coefficient) = 0.7326

Coded RR weight (coefficient) = 0.8724

Total number of cases: 7069 (unweighted). Number of selected cases: 7069. Number of unselected cases: 0

## Section 4.9.2

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 5363.149

\* Constant is included in the model.

Estimation terminated at iteration number 4 because

Log Likelihood decreased by less than .01 percent.

### Section 4.9.2

Beginning Block Number 1. Method: Backward Stepwise (LR)

Variable(s) Entered on Step Number

1.HCISS WRTS AGE

Estimation terminated at iteration number 6 because

Log Likelihood decreased by less than .01 percent.

### Section 4.9.3

-2 Log Likelihood	1929.593
Cox & Snell - R <sup>2</sup>	.385
Nagelkerke - R <sup>2</sup>	.724

### Section 4.9.5

	Chi-Square	df	Significance
Model	3433.556	3	.0000
Block	3433.556	3	.0000
Step	3433.556	3	.0000

### Section 4.9.6 Hosmer and Lemeshow Goodness of Fit Test

HL goodness of fit test = 18.9775 df: 8 p = .0150

### Section 4.9.7 Area under the ROC

ROC = .967 (95% CI: .961 - .973)

### Section 4.9.8

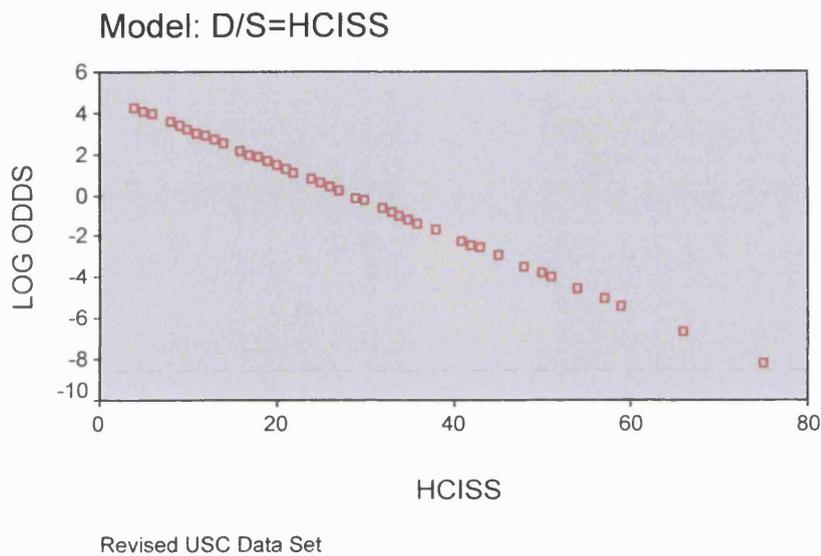
**Table 3.9.8 Variables included in Model Building**

Variable	Coefficient	Standard Error
HCISS	-.1368	.0064
WRTS	.8034	.0297
Age	-.0285	.0034
Constant	-.5067	.2557

No more variables can be deleted or added.

## Section 4.1.9

Log Odds Plotted Against HCISS.  
(odds for survival)

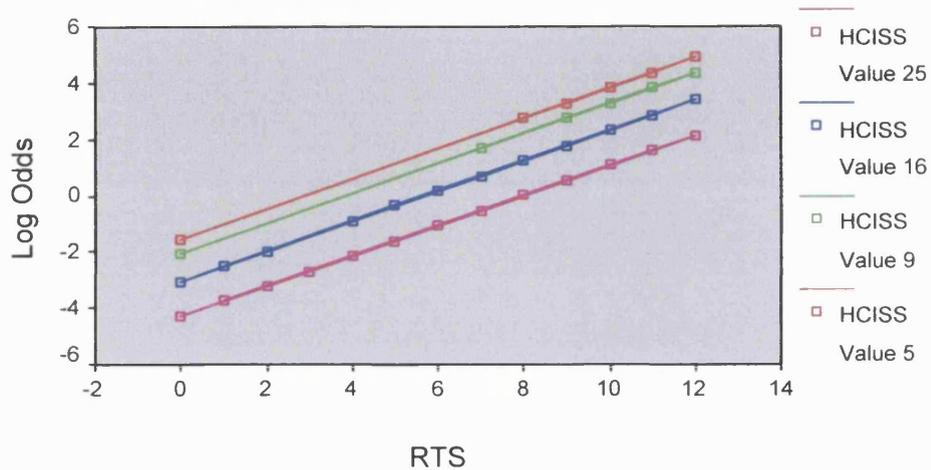


A plot of the log odds against HCISS demonstrates a linear relationship indicating that the logit model is the correct model.

## Log Odds Plotted Against RTS (unweighted)

Model:  $D/S = \text{HClSS} + \text{RTS (unweighted)}$

HClSS values:- 5 ,9, 16, 25, 36



USC Revised Data Set

A plot of the log odds against RTS controlling for HClSS demonstrates a linear relationship indicating that the logit model is the correct model.

## Summary Statistics: Table 1

Model**	H L Value*
1.HCISS	52.0798
2.HCISS + coded GCS	21.5294
3.HCISS + coded GCS + coded SBP	21.3366
4.HCISS + coded GCS + coded SBP + coded RR	21.5060
5.HCISS + coded GCS + coded SBP + coded RR + Age	20.6868
6.HCISS + RTS	32.9552
7.HCISS + RTS + Age	28.2686
8.HCISS + RTS (MTOS weights)	22.5708
9.HCISS + RTS (MTOS weights) + Age	18.9775

\*\* Modelling method:- Backward LR (Likelihood Ratio)

\*Hosmer-Lemeshow chi-squared statistic

## Summary Statistics: Table 2

Model**	ROC*	LCI	UCI
1.HCISS	.908	.898	.919
2.HCISS + coded GCS	.957	.950	.964
3.HCISS + coded GCS + coded SBP	.963	.957	.970
4.HCISS + codedGCS + codedSBP + codedRR	.964	.957	.970
5.HCISS+codedGCS+coded SBP+coded RR+Age	.968	.962	.973
6.HCISS + RTS	.964	.957	.970
7.HCISS + RTS + Age	.967	.961	.973
8.HCISS + RTS (MTOS weights)	.963	.957	.970
9.HCISS + RTS(MTOS weights) + Age	.967	.961	.973

\*\* Modelling method:- Backward LR (Likelihood Ratio)

\* Area under the ROC curve.

UCI: 95% Upper confidence value

LCI: 95% Lower confidence value

### Summary Statistics: Table 3

Model**	C&S <sup>1</sup>	Nag <sup>2</sup>
1.HCISS	.260	.488
2.HCISS + coded GCS	.360	.676
3.HCISS + coded GCS + coded SBP	.378	.712
4.HCISS + coded GCS + coded SBP + coded RR	.380	.714
5.HCISS + coded GCS + coded SBP + coded RR + Age	.386	.727
6.HCISS + RTS	.377	.710
7.HCISS + RTS + Age	.384	.723
8.HCISS + RTS (MTOS weights)	.378	.712
9.HCISS +RTS(MTOS weights) + Age	.385	.724

\*\* Modelling method:- Backward LR (Likelihood Ratio)

1. Cox and Snell  $R^2$       $R^2 = 1 - [L(0)/L(B)]^{2/N}$

Where L(0) is the likelihood for the model with only the constant, L(B) is the likelihood for the model with the predictor variables and N is the sample size.

2. Nagelkerke  $R^2$       $R^2 = \frac{R^2}{R^2_{MAX}}$      where  $R^2_{MAX} = 1 - [L(0)]^{2/N}$

## Section 5: Discussion

### Results Sections 4.1-4.8 (subsections: .1)

All cases in the revised data set were accepted for analysis. The dependent variable was encoded such that death had a value of zero and survival had a value of one. The probabilities generated were for survival rather than death.

### Results Sections 4.1-4.8 (subsections: .2 - .5)

The backward likelihood ratio method was used as the means for selecting the regressors. The likelihood is the probability of the observed results given the parameter estimates. Since this is a small number less than 1 the likelihood is usually expressed as  $-2$  times the log of the likelihood. A good model therefore has a high likelihood and therefore a small  $-2LL$ . The  $-2LL$  for the constant in all the cases was 5363.149 (subsections: .2 ) The  $-2LL$  for the whole model is given in subsections .4 and the model chi-square which is the difference between the  $-2LL$  for the constant and the whole model is reported in subsections: .5. The degrees of freedom for the model chi-square is the difference between the number of parameters in the two models. The chi-squared value for the entry labelled *Block* (subsections: .4) is the change in the  $-2LL$  between successive entry blocks during model building. Because the variables were entered as a single block for each of the nine models developed, the block chi-square is the same as the model chi-square. The entry labelled *Step* in subsections: .4 is the change in the  $-2LL$  between successive steps in model building. In this study only two models were considered at each stage (constant and a single block) therefore the *Step* chi-square is the same as the *block* chi-square. The SPSS program has two other methods of model

selection; the Wald statistic and the Conditional statistic. The conditional statistic option also uses the likelihood ratio test but is computationally less intensive than the likelihood ratio option in SPSS. The Wald statistic in SPSS is calculated by dividing the coefficient by its standard error and has an approximate chi-squared distribution with one degree of freedom. Unfortunately when the coefficient becomes large the estimated standard error becomes too large. This produces a Wald statistic which is too small resulting in failure to reject a predictor (i.e. failure to reject the null hypothesis that the value of the predictor coefficient is zero). Harrell (2002) considers the likelihood ratio test to be the preferred method for selecting predictor variables.

#### **Results Sections 4.1-4.9 (subsections .4)**

The worth of a model in linear regression is determined using  $R^2$  (see footnote\*). This statistic is a function of the Y residuals i.e. it measures the difference between the predicted value ( $\hat{Y}$ ) and the actual value.  $R^2$  is therefore a measure of the *explained variation* of Y. In logistic regression  $R^2$  should not be used when there are only two possible values for Y. The statistic may under predict the worth of the model even if the model fits nearly perfectly (Ryan, 1997). The Cox and Snell test (1989) is a measure of explained variation in the dependent variable which utilises the ratio of the likelihood for the null model by the likelihood for the model with the regressors. Unfortunately the maximum value that  $R^2$  can achieve is .75 using the Cox and Snell method. Nagelkerke (1991) proposed a modification so that the value of 1 could be achieved.

$$* R^2 = \frac{\sum(\hat{Y} - \check{Y})}{\sum(\hat{Y} - Y)}$$

$\bar{Y}$  = mean value of the dependent variable.

$\hat{Y}$  = predicted value for the dependent variable using the model coefficient.

$Y$  = actual value for the dependent variable for a given value of the independent variable.

From table 3 the model with the greatest proportion of explained variation is model 5 (HCISS + coded GCS + coded SBP + coded RR + Age). The model with the smallest proportion of explained variation is model 1 (HCISS only). All models with a physiological variable were found to have a greater proportion of explained variation than model 1 (HCISS only).

#### **Results Sections 4.1-4.9 (subsections .6 and .7)**

Model validation was performed using internal validation of the fitted model rather than by external validation on another data set or cross-validation. Although this is a less stringent test than external validation it still provides some useful information. Subsequent chapters of this thesis will address in detail the different methods of validating a prognostic model.

The Hosmer Lemeshow statistic in SPSS is calculated using an algorithm approach with variable cell numbers rather than using a fixed cell number method. The SPSS algorithm method divides the cases into roughly 10 approximately equal groups based upon their predicted probabilities. Cases with the same combination of values for the predictor variables are kept in the same group. The different methods of calculating the Hosmer Lemeshow statistic and the drawbacks of each method will be discussed in chapters 5 and 6.

### **Results Sections 4.1.9**

The plot of log odds against HCISS (model 1) and RTS controlling for HCISS (model 6) both generated straight lines demonstrating that for this particular data set the logit model is correct for both models. This is in contrast to a recent study by Osler et al (2002) who found that the log odds for death was not linear when plotted against ISS but was better approximated when a squared term was added to the ISS variable. The study was performed on three data sets; a paediatric data set of 53,113 cases, a New Mexico data set of 3,142 and the Portland data set of 2,916 cases. The reason why the ISS model was not linear for the two adult data sets appears to be due to the fact that the odds (and therefore log odds) was calculated from the actual number of deaths and survivors for a given ISS value. The correct way of calculating the log odds is to use the probabilities generated by the fitted model (Harrell, 2002; Chapter 10). The logistic model should be linear when the log odds are plotted against  $X\beta$  ( $X$  is the predictor{s} variable). For a model with only one predictor variable:  $\beta = \beta_0 + \beta_1 X_1$

Therefore  $\text{Log Odds} = \beta_0 + \beta_1 X_1$

$\beta_0$  is the intercept and  $\beta_1$  defines the slope of the line.

If this line is not linear then mathematical adjustments such as quadratic terms may need to be considered to transform the model into its linear form.

The results for the Hosmer Lemeshow goodness of fit statistic and the ROC analysis (tables 1 and 2) demonstrates several points. Firstly, it confirms the well established observation (Champion, 1983) that the addition of the Revised Trauma Score variable (weighted or unweighted) to the Injury Severity Score variable

results in a substantial improvement in the calibration and discrimination of both models (model 6 and 8) when compared to the model with only the ISS variable (model 1). In this study separation of the Revised Trauma Score variable into its three components resulted in better model calibration but not discrimination when compared to the composite Revised Trauma Score variable without weights (model 4 c.f. to model 6). The addition of MTOS weights to the RTS variable substantially improved model calibration but not discrimination (model 8 c.f. model 6). The addition of age as a continuous variable to the three previous models (models 4, 6 and 8) resulted in an improvement in model calibration for all three models. Discrimination was not however significantly improved with the addition of age to the previous three models (4, 6 and 8). The former results are in broad agreement with the findings by Stephenson et al (2002) i.e. that the addition of age as a continuous variable improves model calibration. Interestingly enough model 9 (HCISS + RTS {MTOS weights} + Age) had the best calibration of all nine models. Model 5 had the best discrimination although this was not statistically significant when compared to models 3–9.

The results of this study are in part agreement with the results of Hannan et al (1999) who also found that the RTS variable was superior when utilised in its component form rather than in its weighted composite form. The magnitude of this difference was not mentioned in the monograph although the inference was that the difference was significant. This is in contrast to this study where the difference in calibration between the component RTS and the weighted RTS was minimal. There are several important

differences between the study by Hannan et al (1999) and the author's study. Firstly, in this study model validation was performed using internal validation in contrast to the study by Hannan et al who used external validation on a separate data set. Both external validation and cross-validation provide a more stringent test of model fit (Harrell, 1996). Secondly, in the study by Hannah et al the test data set contained only patients with blunt injuries in comparison to the data set used by this author which contained both blunt and penetrating cases. Thirdly, Hannan et al used the 'deciles of risk' method (i.e. ten equal sized groups of predicted probabilities) to calculate the HL value. In contrast the SPSS algorithm method (see chapter 5 for a detailed description of this method) was used in this study. The various methods of calculating the Hosmer Lemeshow statistic can result in different values for the test statistic as was pointed out by Hannan et al (1995).

This study also showed that the addition of coded SBP and coded RR to model 2 (HCISS + coded GCS) resulted in a marginal reduction in the HL value. These findings support the work of Becalick et al (2001) who found that SBP and RR were relatively unimportant when compared to the motor and verbal components of the GCS. Their results were based upon a 16 predictor variable model developed using a Neural Network method.

In summary the results of this study are in broad agreement with other authors (Hannan, 1999; Becalick, 2001; Stephenson, 2002). Of interest was that the model with the best calibration was in fact a TRISS type model using age as a continuous variable.

# **CHAPTER 5**

## **A STUDY TO COMPARE THE THREE DIFFERENT METHODS OF CALCULATING THE HOSMER LEMESHOW CHI-SQUARE STATISTIC**

### **CONTENTS**

	<b>Page Number</b>
<b>Section 1: Introduction</b>	<b>91</b>
<b>Section 2: Aims</b>	<b>93</b>
<b>Section 3: Methodology</b>	<b>94</b>
<b>Section 4: Results</b>	<b>96</b>
<b>Section 5: Discussion</b>	<b>104</b>
<b>Appendix 1:</b>	<b>108</b>
<b>Appendix 2:</b>	<b>112</b>
<b>Appendix 3:</b>	<b>116</b>

## Section 1: Introduction

The Hosmer Lemeshow goodness of fit statistic (Hosmer DW and Lemeshow S, 1989) is a measure of the model's calibration i.e. the precision of the model to predict survival (or death) over a range of injury severities. It is calculated using the formula:-

$$HL = \frac{\sum(\text{Observed} - \text{Expected Cases})^2}{\text{Expected Cases}}$$

The HL statistic divides the data into roughly ten, approximately equal sized groups (cells) based on their 'deciles' of probability i.e. 0–0.10, 0.11–0.20, etc. The ten cells are then divided into two subgroups based upon the outcome variable; death or survival. The difference between the observed number of deaths minus the expected number of deaths is determined. The expected number of deaths is calculated from the sum of the predicted probabilities. The observed number of deaths is the number of deaths which actually occur in that group. The values for each cell are then summed to give a number. The same procedure is calculated for the survival subgroup. The expected number of survivors is calculated by subtracting the expected number of deaths from the total number (deaths plus survival) in that group. The statistical significance of the final result (HL  $\chi^2$  value) can be determined by looking up the p value in a chi-square table with n-2 degrees of freedom (n = number of cells). The smaller the HL value the better is the model calibration. One of the problems with the HL statistic, as mentioned previously in the introductory chapter is that it is dependent upon the size of the data set. As the data set becomes smaller the HL value decreases and the significance value p also

increases, thus erroneously indicating that a poorly fitting model fits better with a smaller data set.

There are three ways in which the HL statistic can be calculated. Firstly, the predicted probabilities can be divided into ten equal sized groups. Secondly, the predicted probabilities can be divided into ten groups using fixed percentile cut off points (deciles of risk). The test statistic was developed on ten groupings so that the cut-off points were probabilities of 0.1, 0.2, 0.3 through to 0.9. Many statistical software packages use a modified equal groups method (algorithm method). The predicted probabilities are divided into roughly ten groups, although in practice the number is often less than ten. Predicted probabilities with the same covariate pattern are placed into the same group. This method is used by SPSS and also SAS, hence both software packages produce very similar groupings and therefore similar results.

Hosmer et al (1997) found that the various commercially available software packages have different algorithm methods. In a simulation study using six software packages, but not including SPSS and SAS, Hosmer et al (1997) found that the p value ranged from 0.02 to 0.16 depending upon which algorithm method was used. The two other methods for calculating the HL (fixed percentile and fixed group) statistic can also result in substantial differences in the test statistic result. A fact highlighted in a paper by Bertolini et al (2000). This problem was also apparent in a recent comparative study of five anatomic injury severity models by Stephenson et al (2002). The work was performed on a data set of 349,409 patients. The HL statistic (equal sized group method)

gave a value of 3,774 for the mapped ISS model. The same model using the fixed percentile method gave a HL value of 17,634.

## **Section 2: Aims**

The aim of the study was to compare the fixed percentile and fixed group (equal sized groups) method with the HL algorithm method used by SAS.

Hypotheses to be tested.

- (1) That the fixed percentile and fixed group methods would either under predict or over predict model fit compared to the SAS algorithm method for models with one, two, three and four predictor variables.
- (2) That the results would be consistent over a range of data set values.

## **Section 3: Methodology**

**Section 3.1** Software: SAS version 8 was used.

**Section 3.2** Data Set: the revised USC data set was used.

**Section 3.3** Programs for the HL statistic using the Equal Sized Groups (fixed group) and Fixed Percentile Method were written by this author using the SAS programming language. The programs are given at the end of this chapter (appendix 1-2). SAS *proc logistic* program (*Lackfit* statement) was used to generate the HL algorithm method.

**Section 3.4** Method of model selection using the SAS program *proc logistic* was by the backward likelihood ratio test. (The calculated probabilities were for survival rather than death).

### **Section 3.5**

Four models were chosen: -

#### **Model 1: Single predictor variable**

Dependent variable: death/survival

Predictor variable: HCISS

#### **Model 2: Single predictor variable**

Dependent variable: death/survival

Predictor variables: GCS

#### **Model 3: Two predictor variables**

Dependent variable: death/survival

Predictor variables: HCISS + RTS

#### **Model 4: Three predictor variables**

Dependent variable: death/survival

Predictor variables: coded GCS + coded SBP + coded RR

(coding for all three variables was performed using the triage RTS cut-off values)

#### **Model 5: Four predictor variables**

Dependent variable: death/survival

Predictor variables: HCISS + coded GCS + coded SBP + coded RR

(coding for all three variables was performed using the triage RTS cut-off values)

### **Section 3.6**

The HL goodness of fit tests were applied to the five models by internal validation. Seven HL values for each model were obtained by sequentially increasing the size of the data set.

Data set values used were:-

1. First 1000 cases.
2. First 2000 cases.
3. First 3000 cases.
4. First 4000 cases.
5. First 5000 cases.
6. First 6000 cases.
7. First 7000 cases.

## Section 4: Results

**Table 1a. HL Values: SAS Algorithm Method**

Data Set Size	Model HClSS	Model HClSS+RTS	Model HClSS+cGCS+cSBP+cRR
1000.00	6.84	7.30	6.60
2000.00	10.64	10.27	10.64
3000.00	12.97	15.08	10.27
4000.00	65.03	15.29	14.81
5000.00	71.04	21.57	18.08
6000.00	41.62	27.66	22.81
7000.00	50.09	32.49	21.57

**Table 1b. HL Values: SAS Algorithm Method**

Data Set Size	Model GCS	Model cGCS+cSBP+cRR
1000.00	7.52	12.87
2000.00	10.17	15.58
3000.00	24.19	28.07
4000.00	31.17	41.73
5000.00	38.21	37.14
6000.00	35.52	46.00
7000.00	39.30	39.84

**Table 2a. HL Values: Equal Sized Groups Method**

Data Set Size	Model HClSS	Model HClSS+RTS	Model HClSS+cGCS+cSBP+cRR
1000.00	10.33	8.06	8.28
2000.00	10.21	11.69	10.66
3000.00	27.86	10.79	8.58
4000.00	53.90	15.62	14.96
5000.00	45.47	25.18	22.29
6000.00	36.15	36.80	27.27
7000.00	45.12	40.39	29.07

**Table 2b. HL Values: Equal Sized Groups Methods**

Data Set Size	Model GCS	Model cGCS+cSBP+cRR
1000.00	8.38	11.61
2000.00	12.33	19.87
3000.00	30.15	38.70
4000.00	54.67	42.12
5000.00	74.15	49.48
6000.00	72.53	60.25
7000.00	83.46	73.42

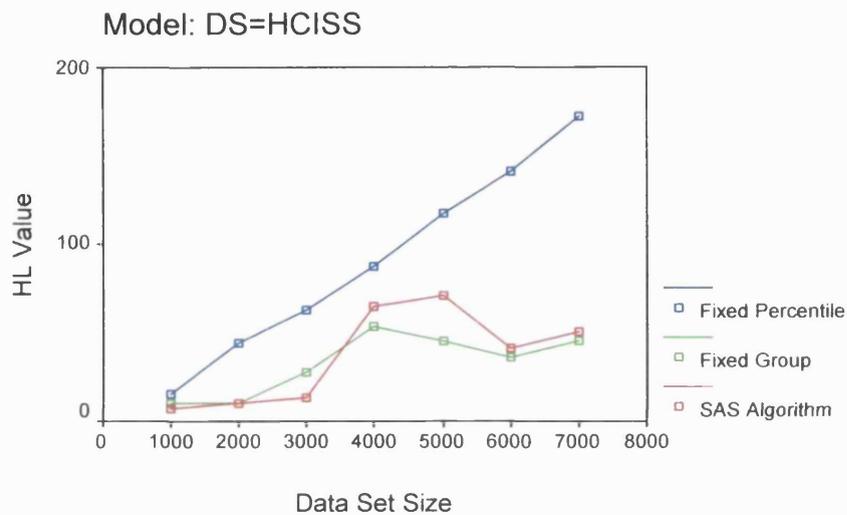
**Table 3a. HL Values: Fixed Percentile Method**

Data Set Size	Model HClSS	Model HClSS+RTS	Model HClSS+cGCS+cSBP+cRR
1000.00	15.28	6.28	6.80
2000.00	44.27	11.39	15.39
3000.00	62.38	6.43	8.27
4000.00	87.65	9.20	9.52
5000.00	117.68	13.34	18.46
6000.00	140.96	19.72	18.83
7000.00	172.23	28.11	29.96

**Table 3b. HL Values: Fixed Percentile Methods**

Data Set Size	Model GCS	Model cGCS+cSBP+cRR
1000.00	20.24	5.19
2000.00	27.71	21.80
3000.00	45.16	24.91
4000.00	60.92	24.18
5000.00	67.51	28.66
6000.00	44.56	27.07
7000.00	62.27	20.80

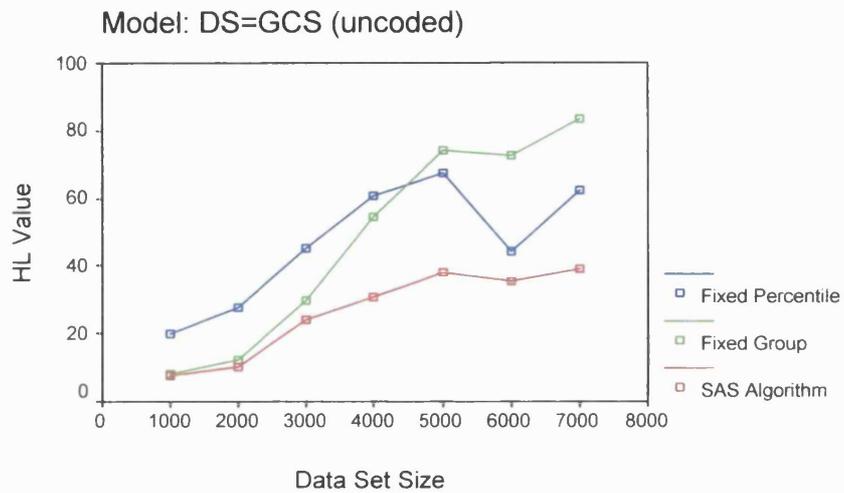
## The Three Different HL Methods Plotted Against Sequential Increase In Data Set Size



**Graph 1**

Graph 1 shows that the fixed percentile method consistently under predicts the goodness of fit compared to the algorithm method. The fixed group method produced HL values which were similar to the algorithm method.

## The Three Different HL Methods Plotted Against Sequential Increase In Data Set Size

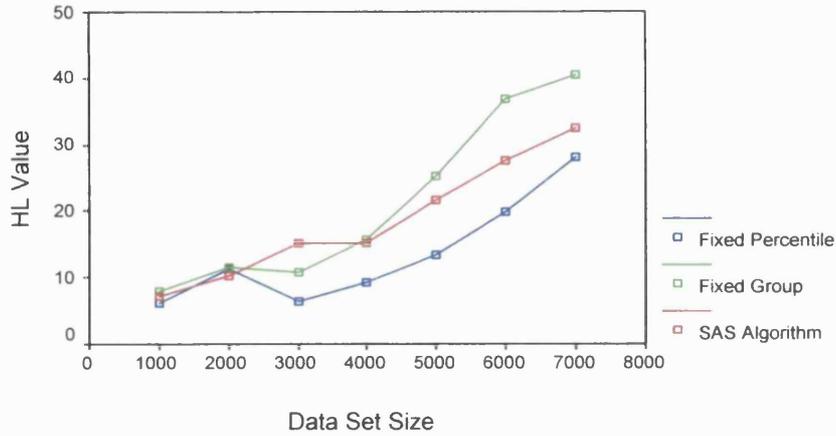


**Graph 2**

Graph 2 shows under prediction for both the fixed percentile method and the fixed group method compared to the algorithm method over the range of data set values.

## The Three Different HL Methods Plotted Against Sequential Increase In Data Set Size

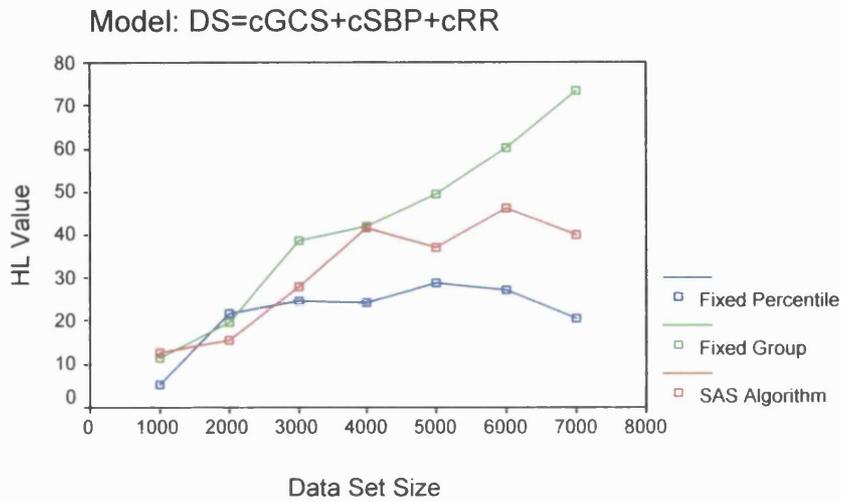
Model: DS=HClSS+RTS



**Graph 3**

Graph 3 shows some under prediction and one over prediction by the fixed group method. The fixed percentile method shows over prediction for the majority of values.

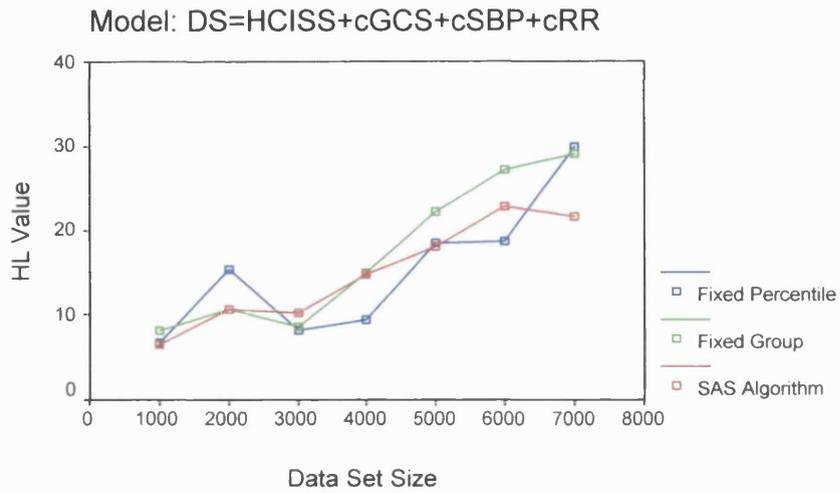
## The Three Different HL Methods Plotted Against Sequential Increase In Data Set Size



**Graph 4**

Graph 4 shows predominantly under prediction for the fixed group method and predominantly over prediction for the fixed percentile method compared to the algorithm method.

## The Three Different HL Methods Plotted Against Sequential Increase In Data Set Size



**Graph 5**

Graph 5 shows under and over prediction of the fixed percentile and fixed group method compared to the algorithm method.

## Section 5: Discussion

The results of this study demonstrate the variability of the three different methods for calculating the HL test. No clear trends were identified. The fixed percentile method under predicted with respect to the algorithm method for the single HCISS and GCS models. The fixed percentile method however over predicted for nearly all values for the HCISS + RTS model and also the three predictor variable model. The fixed group method showed a trend towards under prediction compared to the algorithm method. This effect was most pronounced for the single GCS predictor variable model.

The fixed percentile method is intuitively the best method to assess calibration of the model over a range of percentiles but has the disadvantage of requiring that the predicted probabilities are represented over the entire range. If not cells are created of varying size and some cells may be empty resulting in erratic performance of the test statistic as pointed out by Pigeon et al (1999a). The equal sized group (fixed group) method overcomes the latter problem but has the theoretical disadvantage that cells may contain predicted probabilities with widely differing values. Le Cessie et al (1991, 1995) argue that the latter problem results in a lack of power in the 'x' space. The algorithm method used in SPSS and SAS has advantages of both methods. It avoids grouping predicted probabilities with widely differing values into the same cell. It also ensures that the cells are similar in size. The main disadvantage with the SPSS algorithm method generated by the logistic program is that it can only be used for internal validation procedures. This problem also occurs with the *proc logistic* program in SAS. A SAS

program is however available from the sas.com website (provided in the appendix 3 section of this chapter) which enables the HL algorithm method to be applied to a test data set.

Pigeon (1999b) proposed a modified version of the Hosmer Lemeshow\*\* statistic based upon simulation studies using a modified Pearson chi-squared test.

$$**J^2 = \sum_{G=1}^G \sum_{k=1}^2 \frac{(O_{gk} - E_{gk})^2}{\Phi E_{gk}}$$

Expected number of events in group  $E_{g1} = \sum_{i=1}^{n_g} \pi_{i1}$

Expected number of non-events in group  $E_{g2} = n_g - E_{g1}$

Observed number of events in group g is  $O_{g1} = \sum_{i=1}^{n_g} y_{i1}$

Observed number of non-events is  $O_{g2} = n_g - O_{g1}$

$$\text{Where } \Phi_g = \frac{\sum \pi_{i1}(1 - \pi_{i1})}{n_g \bar{y}_{g1}(1 - \bar{y})}$$

The expected number of events in group g =  $\pi_{i1}$

$\bar{y}_{g1}$  is the average of  $\pi_{i1}$  for the  $n_g$  subjects in the gth group

$J^2$  is approximately  $\chi^2$  distributed with G-1 degrees of freedom

Using the low birth weight data given in the now classic book *Applied Logistic Regression* (Hosmer and Lemeshow, 1989), Pigeon (1999b) found that his new statistic produced a similar result to the HL statistic using the equal group method: ('deciles of risk') HL = 5.23 df=8  $J^2$ =5.26 df=9.

Pigeon (1999b) points out that the grouping strategy of the HL statistic tends to correct for model over-fitting (underdispersion), a problem commonly seen with the Pearson chi-squared statistic. The correction factor (i.e.  $\Phi_g$ ) performs the same function for the  $\mathcal{J}^2$  as the grouping strategy does for the HL statistic. Further simulation studies are required apart from those performed by Pigeon (1999b) in order to determine whether the  $\mathcal{J}^2$  statistic is superior to the HL statistic. Predicted probabilities for the HL statistic can be grouped by factors other than group size or percentile e.g. age. Meredith et al (2002) in a comparative study of several injury severity models used a modified HL test where the predicted probabilities were grouped on their index score. Pigeon et al (1999a) cautions against the use of groupings other than by deciles of risks as the distribution for the HL test has only been validated using the deciles of risk strategy.

## **Conclusions**

The results of this study failed to demonstrate any clear trends in performance between the three methods of calculating the HL statistic. The variability in the test result between methods makes accurate comparisons between studies which use differing methods of calculating the HL test problematic. Hosmer et al (1988) performed a series of simulation studies comparing the fixed percentile method to the fixed group method. They concluded that both methods produced erratic results, usually under prediction particularly when the cell sizes were less than five. They recommended that one should avoid using the fixed percentile

group when the expected cell frequencies are likely to be small. If the range of predicted probabilities is narrow then the fixed percentile method should be used in preference to the fixed group method.

# APPENDIX 1

## Program For Equal Sized Groups

```
proc logistic data=sasuser.Hcfinal2 (firstobs=1 obs=4000) DES;
model ds = hciiss /
      selection = backward;
      output out=lout p=pred;
run;
```

```
proc sort data=lout;
by pred;
run;
```

```
data res;
merge work.lout sasuser.num4000;
run;
```

```
data group1a group2a group3a group4a group5a
      group6a group7a group8a group9a group10a;
set res;
if number <= 400 then output group1a;
if (number <= 800 and number > 400) then output group2a;
if (number <= 1200 and number > 800) then output group3a;
if (number <= 1600 and number > 1200) then output group4a;
if (number <= 2000 and number > 1600) then output group5a;
if (number <= 2400 and number > 2000) then output group6a;
if (number <= 2800 and number > 2400) then output group7a;
if (number <= 3200 and number > 2800) then output group8a;
if (number <= 3600 and number > 3200) then output group9a;
if (number <= 4000 and number > 3600) then output group10a;
run;
```

```
proc means data=group1a noprint;
var ds pred;
output out=gpla sum=;
run;
```

```
data gplaresult;
set work.gpla;
chil = (ds - pred)**2;
chi = chil/pred;
run;
```

```
data gplbresult;
set work.gplaresult;
death = (_freq_ - ds);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;
```

```
proc means data=group2a noprint;
  var ds pred;
  output out=gpla sum=;
run;
```

```
data gp2aresult;
  set work.gpla;
  chil = (ds - pred)**2;
  chi = chil/pred;
run;
```

```
data gp2bresult;
  set work.gp2aresult;
  death = (_freq_ - ds);
  ppd = _freq_ - pred;
  chil = (death - ppd)**2;
  chi = chil/ppd;
run;
```

```
proc means data=group3a noprint;
  var ds pred;
  output out=gp3a sum=;
run;
```

```
data gp3aresult;
  set work.gp3a;
  chil = (ds - pred)**2;
  chi = chil/pred;
run;
```

```
data gp3bresult;
  set work.gp3aresult;
  death = (_freq_ - ds);
  ppd = _freq_ - pred;
  chil = (death - ppd)**2;
  chi = chil/ppd;
run;
```

```
proc means data=group4a noprint;
  var ds pred;
  output out=gp4a sum=;
run;
```

```
data gp4aresult;
  set work.gp4a;
  chil = (ds - pred)**2;
  chi = chil/pred;
run;
```

```
data gp4bresult;
  set work.gp4aresult;
  death = (_freq_ - ds);
  ppd = _freq_ - pred;
  chil = (death - ppd)**2;
  chi = chil/ppd;
run;
```

```
proc means data=group5a noprint;
  var ds pred;
output out=gp5a sum=;
run;
```

```
data gp5aresult;
set work.gp1a;
chil = (ds - pred)**2;
chi = chil/pred;
run;
```

```
data gp5bresult;
set work.gp5aresult;
death = (_freq_ - ds);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;
```

```
proc means data=group6a noprint;
  var ds pred;
output out=gp6a sum=;
run;
```

```
data gp6aresult;
set work.gp6a;
chil = (ds - pred)**2;
chi = chil/pred;
run;
```

```
data gp6bresult;
set work.gp6aresult;
death = (_freq_ - ds);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;
```

```
proc means data=group7a noprint;
  var ds pred;
output out=gp7a sum=;
run;
```

```
data gp7aresult;
set work.gp7a;
chil = (ds - pred)**2;
chi = chil/pred;
run;
```

```
data gp7bresult;
set work.gp7aresult;
death = (_freq_ - ds);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;
```

```

proc means data=group8a noprint;
  var ds pred;
output out=gp8a sum=;
run;

  data gp8aresult;
  set work.gp8a;
  chil = (ds - pred)**2;
  chi = chil/pred;
run;

data gp8bresult;
set work.gp8aresult;
death = (_freq_ - ds);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group9a noprint;
  var ds pred;
output out=gp9a sum=;
run;

  data gp9aresult;
  set work.gp9a;
  chil = (ds - pred)**2;
  chi = chil/pred;
run;

data gp9bresult;
set work.gp9aresult;
death = (_freq_ - ds);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group10a noprint;
  var ds pred;
output out=gp10a sum=;
run;

  data gp10aresult;
  set work.gp10a;
  chil = (ds - pred)**2;
  chi = chil/pred;
run;

data gp10bresult;
set work.gp10aresult;
death = (_freq_ - ds);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

```

```

data finala;
set gp1aresult gp2aresult gp3aresult gp4aresult gp5aresult
  gp6aresult gp7aresult gp8aresult gp9aresult gp10aresult
  gp1bresult gp2bresult gp3bresult gp4bresult gp5bresult
  gp6bresult gp7bresult gp8bresult gp9bresult gp10bresult;
  run;

  data finalb;
set work.finala;
  ds = dt;
  run;

proc means data=finalb noprint;
  var chi;
output out=finalb sum=;
run;

```

## APPENDIX 2

### Program For Fixed Percentiles

```

proc logistic data=sasuser.hcfinal (FIRSTOBS=1 OBS=1000) DES;
model dl= hciss regcs resbp rerr / selection = backward lackfit;
Output OUT=res(KEEP= dl pred)P=pred;
run;

```

```

data group1a group2a group3a group4a group5a
  group6a group7a group8a group9a group10a;
set work.res;
if pred <= 0.1 then output group1a;
if (pred<= 0.2 and pred > 0.1) then output group2a;
if (pred <= 0.3 and pred > 0.2) then output group3a;
  if (pred <= 0.4 and pred > 0.3)then output group4a;
if (pred <= 0.5 and pred > 0.4) then output group5a;
if (pred <= 0.6 and pred > 0.5) then output group6a;
if (pred <= 0.7 and pred > 0.6) then output group7a;
if (pred <= 0.8 and pred > 0.7) then output group8a;
if (pred <= 0.9 and pred > 0.8) then output group9a;
if (pred <= 1.0 and pred > 0.9) then output group10a;
run;

```

```

proc means data=group1a noprint;
  var dl pred;
output out=gpla sum=;
run;

```

```

data gp1aresult;
set work.gpla;
chil = (dl - pred)**2;
chi = chil/pred;
run;

```

```

data gp1bresult;
set work.gp1aresult;

```

```

death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group2a noprint;
  var dl pred;
output out=gpla sum=;
run;

data gp2aresult;
  set work.gpla;
  chil = (dl - pred)**2;
  chi = chil/pred;
run;

data gp2bresult;
set work.gp2aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group3a noprint;
  var dl pred;
output out=gp3a sum=;
run;

  data gp3aresult;
  set work.gp3a;
  chil = (dl - pred)**2;
  chi = chil/pred;
  run;

data gp3bresult;
set work.gp3aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group4a noprint;
  var dl pred;
output out=gp4a sum=;
run;

  data gp4aresult;
  set work.gp4a;
  chil = (dl - pred)**2;
  chi = chil/pred;
  run;

data gp4bresult;
set work.gp4aresult;

```

```

death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group5a noprint;
  var dl pred;
output out=gp5a sum=;
run;

  data gp5aresult;
  set work.gp1a;
  chil = (dl - pred)**2;
  chi = chil/pred;
run;

data gp5bresult;
set work.gp5aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group6a noprint;
  var dl pred;
output out=gp6a sum=;
run;

  data gp6aresult;
  set work.gp6a;
  chil = (dl - pred)**2;
  chi = chil/pred;
run;

data gp6bresult;
set work.gp6aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group7a noprint;
  var dl pred;
output out=gp7a sum=;
run;

  data gp7aresult;
  set work.gp7a;
  chil = (dl - pred)**2;
  chi = chil/pred;
run;

data gp7bresult;
set work.gp7aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;

```

```

chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group8a noprint;
  var dl pred;
output out=gp8a sum=;
run;

  data gp8aresult;
  set work.gp8a;
  chil = (dl - pred)**2;
  chi = chil/pred;
  run;

data gp8bresult;
set work.gp8aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group9a noprint;
  var dl pred;
output out=gp9a sum=;
run;

  data gp9aresult;
  set work.gp9a;
  chil = (dl - pred)**2;
  chi = chil/pred;
  run;

data gp9bresult;
set work.gp9aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

proc means data=group10a noprint;
  var dl pred;
output out=gp10a sum=;
run;

  data gp10aresult;
  set work.gp10a;
  chil = (dl - pred)**2;
  chi = chil/pred;
  run;

data gp10bresult;
set work.gp10aresult;
death = (_freq_ - dl);
ppd = _freq_ - pred;
chil = (death - ppd)**2;
chi = chil/ppd;
run;

```

```

data finala;
set gp1aresult gp2aresult gp3aresult gp4aresult gp5aresult
    gp6aresult gp7aresult gp8aresult gp9aresult gp10aresult
    gp1bresult gp2bresult gp3bresult gp4bresult gp5bresult
    gp6bresult gp7bresult gp8bresult gp9bresult gp10bresult;
run;

proc means data=finala noprint;
var chi;
output out=finalb sum=;
run;

```

## APPENDIX 3

Program for using the HL Algorithm Method (SAS) to externally validate a model.

```

/* Example 1: Binary response
=====
=====*/
data train; /* Training data set */
input dl hciss;
cards;
.00 26.00
1.00 1.00
.00 26.00
.00 25.00
1.00 5.00
1.00 13.00
1.00 1.00
1.00 1.00
.00 25.00
1.00 1.00
1.00 16.00
1.00 1.00
1.00 10.00
.00 43.00
.00 26.00
1.00 1.00
;

/* Fit the model to the training data set.
=====
=====*/
ods trace on;
ods listing close;
ods output lackfitchisq=lack1;

```

```

proc logistic data=train (FirstObs=2001 Obs=3000) outest=parms
descending;
  model dl = hciss / lackfit;
  output out=out1 p=p;
  run;

proc print data=out1;
  var hciss _level_ p;
  run;

ods listing;
proc print data=lack1;

data test; /* Validation data set */
  input dl hciss;
cards;
1.00 9.00
1.00 4.00
1.00 1.00
.00 75.00
1.00 10.00
1.00 1.00
1.00 1.00
1.00 1.00
1.00 10.00
1.00 4.00
1.00 1.00
1.00 18.00
1.00 9.00
1.00 29.00
1.00 10.00
1.00 4.00
1.00 14.00
;
/* Score the validation data set.

=====
=====*/
ods trace on;
ods listing close;
ods output lackfitchisq=lack2;

proc logistic data=test (FirstObs=12001 Obs=13000) inest=parms
descending;
  model dl = hciss / lackfit maxiter=0;
  output out=out2 p=p;
  run;

proc print data=out2;
  var hciss _level_ p;
  run;

ods listing;

proc print data=lack2;

```

# CHAPTER 6

## A STUDY TO DETERMINE THE EFFECTS OF CHANGING THE COVARIATE PATTERN ON THE HOSMER LEMESHOW CHI- SQUARED STATISTIC

### CONTENTS

	<b>Page Number</b>
<b>Section 1: Introduction</b>	<b>119</b>
<b>Section 2: Aims</b>	<b>119</b>
<b>Section 3: Methodology</b>	<b>120</b>
<b>Section 4: Results</b>	<b>131</b>
<b>Section 5: Discussion</b>	<b>145</b>

## **Section 1: Introduction**

Several authors have highlighted the deficiencies of the HL goodness of fit test (Hosmer, 1997; Stephenson, 2002). One of the most notably being the differing methods of calculating the statistic. Relatively little work has been undertaken to determine the effect of changing the covariate pattern on the HL statistic. The impact of changing the covariate pattern was noted as an incidental finding by Stephenson et al (2002). In their comparative study, the addition of age as a continuous variable to five anatomic injury severity models resulted in an apparent over fitting of four of the models. For example, they found that the HL (equal group method) value for the ICISS model changed from 3,173 to 81 simply by the addition of the age variable. This dramatic change in the HL value can only be explained by an erratic change in the covariate pattern rather than by a massive improvement in the model fit due to the addition of the age variable. A more impressive example of this problem is given by Meredith et al (2002). These authors divided the National Trauma Data Bank of 76,871 patients into four covariate groups using date of birth (odd or even year) as the sole predictor variable. The resulting model produced a HL result of 1.88. Lack of awareness of these potential problems can lead to errors when evaluating the goodness of fit of a logistic model using the HL statistic.

## **Section 2: Aims**

The aim of the study was to determine the effect of changing the covariate pattern of a predictor variable on the Hosmer Lemeshow chi-square statistic.

## **Section 3: Methodology**

### **Study 1**

**Section 3.1a** Software: SAS version 8 was used.

**Section 3.2a** Data Set: the revised USC data set was used.

**Section 3.3a** The hypotheses to be tested.

1. Reducing the number of HCISS covariate patterns by random recoding into 12 groups would reduce the HL value compared to the original HCISS model.
2. Random recoding of the HCISS variable into 6 groups would produce a smaller HL value compared to the model with 12 groups.
3. Recoding of the HCISS using clinical cut-off points into 6 groups would produce a smaller HL value compared to the above two models.

### **Section 3.4a Models**

Method of model selection for the logistic model was by the backward likelihood ratio test. Four models were chosen: -

#### **The 'Control' Model**

Dependent variable: death/survival

Predictor variable: HCISS (hand calculated Injury Severity Score variable)

#### **Model 1**

Dependent variable: death/survival

Predictor variable: coded Injury Severity Score variable (HCISS)

The first model was developed by reducing the HCISS predictor variable into six covariate groups:-

HCISS values 4 – 8: recoded value: - 1

HCISS values 9 – 15: recoded value: - 2

HCISS values 16 – 24: recoded value: - 3

HCISS values 25 – 35: recoded value: - 4

HCISS values 36 – 50: recoded value: - 5

HCISS values 51 – 75: recoded value: - 6

The cut-off point 15/16 was chosen as Champion et al (1989) has shown that these values can be used to separate patients into major and minor trauma categories. The cut-off point of 8/9 was based on the fact that an ISS of 9 includes patients with an AIS of 3 as well as an AIS combination of 2, 2 and 1. Champion et al (1996) in the development of the ASCOT model found that the strongest predictor variable was the maximum AIS (i.e. the worst injury). One would anticipate therefore that a patient with an ISS of 9 (AIS of 3) would have a worse outcome than a patient with an ISS of 8 (AIS of 2 and 2). The cut-off point 50/51 was chosen on the basis that patients rarely survive with an ISS of greater than 50. This assumption was confirmed by an analysis of the USC data which showed that there were no survivors with an ISS over 50. The remaining points were arbitrarily chosen to produce similar coverage values for the coded variables.

## **Model 2**

Dependent variable: death/survival

Predictor variable: coded Injury Severity Score variable (HCISS)

The second model was also developed by reducing the HCISS predictor variable into six covariate groups. The coverage range was similar to model 1 except that the cut-off points were randomly chosen with less clinical correlation except for coded value 6.

HCISS values 4 – 10: recoded value: - 1

HCISS values 11 – 20: recoded value: - 2

HCISS values 21 – 30: recoded value: - 3

HCISS values 31 – 40: recoded value: - 4

HCISS values 41 – 50: recoded value: - 5

HCISS values 51 – 75: recoded value: - 6

### **Model 3**

Dependent variable: death/survival

Predictor variable: coded Injury Severity Score variable (HCISS)

The third model was developed by reducing the HCISS predictor variable into 12 covariate groups. The cut-off points were chosen to produce similar coverage values for each coded value.

HCISS values 4 – 10: recoded value: - 1

HCISS values 11 – 15: recoded value: - 2

HCISS values 16 – 20: recoded value: - 3

HCISS values 21 – 25: recoded value: - 4

HCISS values 26 – 30: recoded value: - 5

HCISS values 31 – 35: recoded value: - 6

HCISS values 36 – 40: recoded value: - 7

HCISS values 41 – 45: recoded value: - 8

HCISS values 46 – 50: recoded value: - 9

HCISS values 51 – 55: recoded value: - 10

HCISS values 56 – 70: recoded value: - 11

HCISS values 71 – 75: recoded value: - 12

### **Section 3.5a**

The three methods of calculating the Hosmer Lemeshow chi-squared statistic previously described in chapter 5 were used. Programs for the HL statistic using the Equal Sized Groups and Fixed Percentile methods are given at the end of chapter 5 (appendix 1 and 2). SAS version 8 was used to generate the HL algorithm method.

The three HL goodness of fit tests were applied to the four models by internal validation. Seven HL values for each model were obtained by sequentially increasing the size of the data set by the method previously described in chapter 5. The first data set point was the first 1000 cases, the second data set point was the first 2000 cases, the third data set point was the first 3000 cases, the fourth data set point was the first 4000 cases, the fifth data set point was the first 5000 cases, the sixth data set point was the first 6000 cases and the seventh data set point was the first 7000 cases.

## **Study 2**

**Section 3.1b** The hypotheses to be tested.

1. Reducing the GCS covariate pattern would reduce the HL value.
2. This effect would hold for all three methods of calculating the HL value.
3. The effect would hold for a range of data set values.

**Section 3.2b Software:** SAS version 8 was used.

**Section 3.3b Data Set:** the revised USC data set was used.

### **Section 3.4b Models**

Method of model selection for the logistic model was by the backward likelihood ratio test. The models were: -

#### **The 'Control' Model**

Dependent variable: Death/ survival

Predictor variable: GCS

#### **Model 1**

Dependent variable: Death/ survival

Predictor variable: RTS coded GCS

The three HL goodness of fit tests were applied to the two models by internal validation. Seven HL values for each model were obtained by sequentially increasing the size of the data set by the method previously described in study 1, section 3.5a.

## **Study 3**

**Section 3.1c** The hypotheses to be tested.

1. Reducing the SBP covariate pattern would reduce the HL value.
2. This effect would hold for all three methods of calculating the HL value.
3. The effect would hold for a range of data set values.

**Section 3.2c Software:** SAS version 8 was used.

**Section 3.3c Data Set:** the revised USC data set was used.

### **Section 3.4c Models**

Method of model selection for the logistic model was by the backward likelihood ratio test. The models chosen were: -

#### **The 'Control' Model**

Dependent variable: Death/survival

Predictor variable SBP

#### **Model 1**

Dependent variable: Death/ survival

Predictor variable: RTS coded SBP

The three HL goodness of fit tests were applied to the two models by internal validation. Seven HL values for each model were obtained by sequentially increasing the size of the data set by the method previously described in study 1, section 3.5a.

### **Study 4**

**Section 3.1d** The hypotheses to be tested.

1. Reducing the Age covariate pattern would reduce the HL value.
2. This effect would hold for all three methods of calculating the HL value.
3. The effect would hold for a range of data set values.

### **Section 3.2d Models**

Method of model selection for the logistic model was by the backward likelihood ratio test. Four models were chosen: -

#### **The 'Control' Model**

Dependent variable: Death/survival

Predictor variable Age > 15 (continuous variable)

#### **Model 1**

Dependent variable: Death/ survival

Predictor variable: coded Age

16 - 20 years coded 1

21- 30 years coded 2

31 - 40 years coded 3

41 – 50 years coded 4

51 – 60 years coded 5

61 -70 years coded 6

71 – 80 years coded 7

81 – 90 years coded 8

91 – 100 years coded 9

The cut-off points were designed to produce 9 covariate groups.

The coverage values are equal except for coded value 1.

#### **Model 2**

16– 20 years coded 1

21– 40 years coded 2

41 – 60 years coded 3

61 – 80 years coded 4

81 – 100 years coded 5

The cut-off points were designed to produce 5 covariate groups.

The cut-off points are equal except for coded value 1.

### **Model 3**

16 – 40 years coded 1

41 – 70 years coded 2

71 – 80 years coded 3

The cut-off points were randomly chosen to produce 3 covariate groups. Coverage values are similar for the three groups.

The three HL goodness of fit tests were applied to the four models by internal validation. Seven HL values for each model were obtained by sequentially increasing the size of the data set by the method previously described in study 1, section 3.5a.

### **Study 5**

**Section 3.1e** The hypotheses to be tested.

1. Reducing the age covariate grouping would have little effect on the HL value of a model containing the HCISS variable and the age variable due to the large number of covariate patterns already present in the HCISS variable.
2. This effect would hold for all three methods of calculating the HL value.
3. The effect would hold for a range of data set values.

### **Section 3.2e Models**

Method of model selection for the logistic model was by the backward likelihood ratio test. Four models were chosen: -

#### **The 'Control' Model**

Dependent variable: death/survival

Predictor variables: HCISS + Age (continuous variable)

#### **Model 1**

Dependent variable: death/survival

Predictor variables : HCISS + Age (coded variable as for model 1 in previous section)

#### **Model 2**

Dependent variable: death/survival

Predictor variables: HCISS + Age (coded variable as for model 2 in previous section)

#### **Model 3**

Dependent variable: death/survival

Predictor variables: HCISS + Age (coded variable as for model 3 in previous section)

The three HL goodness of fit tests were applied to the four models by internal validation. Seven HL values for each model were obtained by sequentially increasing the size of the data set by the method previously described in study 1, section 3.5a.

## **Study 6**

### **Section 3.1f**

The hypotheses to be tested.

1. Reducing the age covariate pattern would have little effect on the HL value of a model containing the HCISS variable, the RTS variable and also the age variable due to the large number of covariate patterns already present in the model.
2. This effect would hold for all three methods of calculating the HL value.
3. The effect would hold for a range of data set values.

### **Section 3.2f Models**

Method of model selection for the logistic model was by the backward likelihood ratio test. Four models were chosen: -

#### **The 'Control' Model**

Dependent variable: death/survival

Predictor variables: HCISS + RTS + Age (continuous variable)

#### **Model 1**

Dependent variable: death/survival

Predictor variables: HCISS + RTS + Age (coded variable as for model 1 in previous section)

#### **Model 2**

Dependent variable: death/survival

Predictor variables: HCISS + RTS + Age (coded variable as for model 2 in previous section)

### **Model 3**

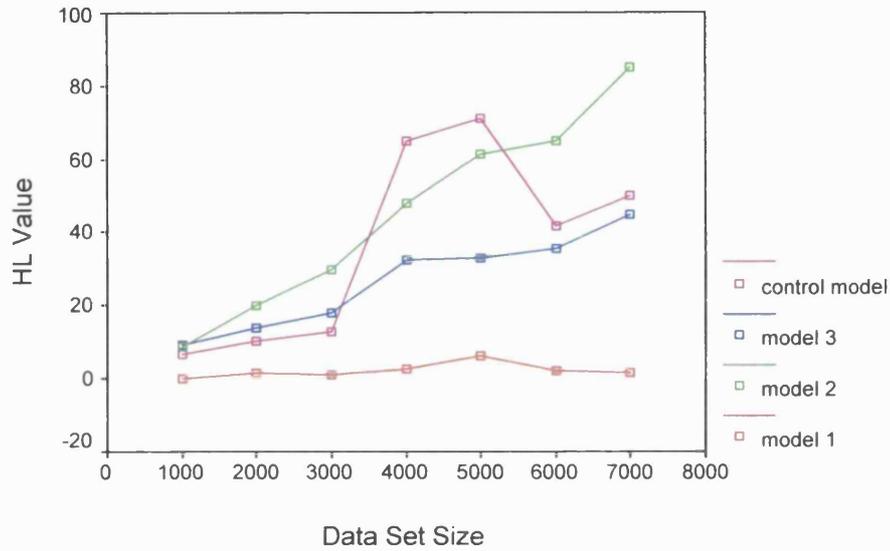
Dependent variable: death/survival

Predictor variables: HCISS + RTS + Age (coded variable as for model 3 in previous section)

The three HL goodness of fit tests were applied to the four models by internal validation. Seven HL values for each model were obtained by sequentially increasing the size of the data set by the method previously described in study 1, section 3.5a.

## RESULTS: Study 1

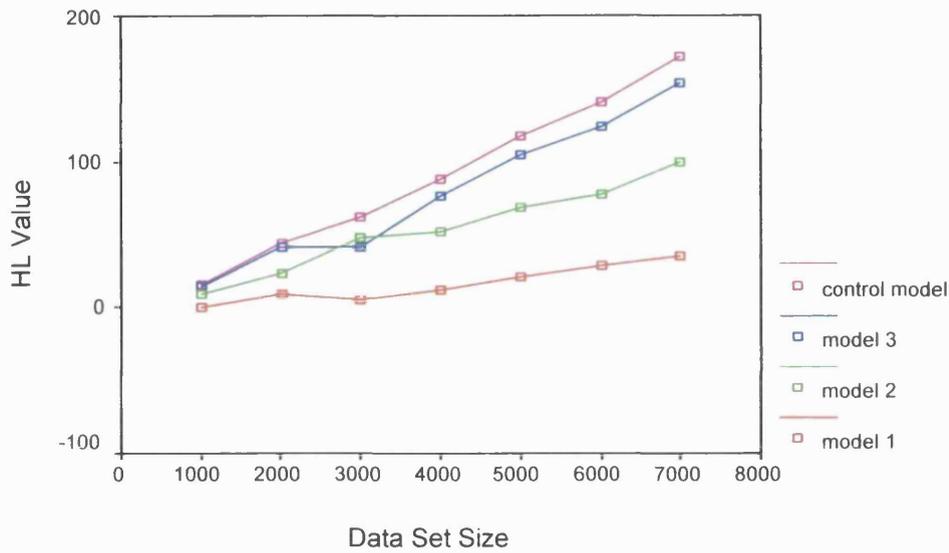
HL Values (Algorithm Method ) Plotted  
Against Data Set Size



**Graph 1**

Graph 1 shows that model 1 had the greatest amount of model over-fit (lowest HL value). The result was sustained for all data set values. Surprisingly, model 3 showed a greater degree of model over-fit when compared to model 2.

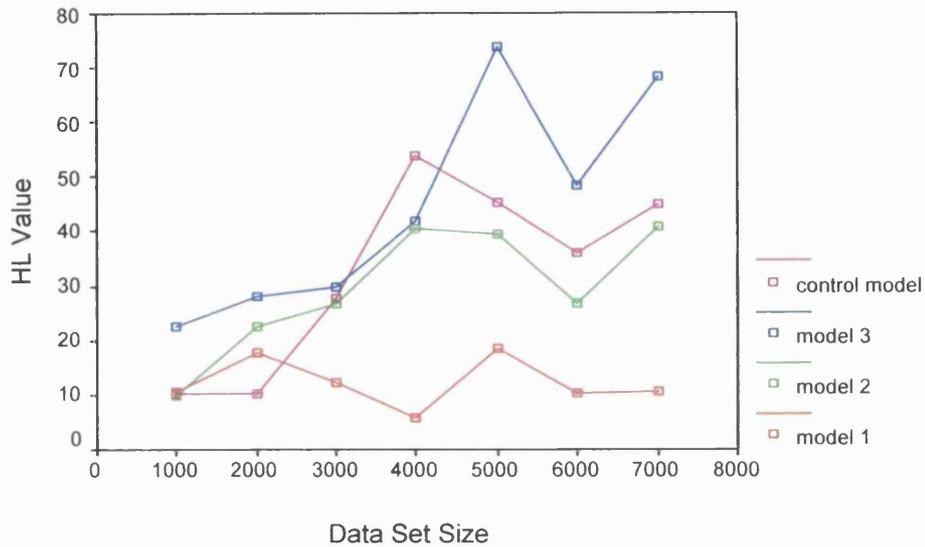
## HL Values (Fixed Percentile Method ) Plotted Against Data Set Size



**Graph 2**

Graph 2 demonstrates that the greatest degree of model over-fit was produced by the model with 6 groups using clinical cut-off points (model 1). The least degree of model over-fit was seen with the model with 12 random groups (model 3).

## HL Values (Fixed Group Method ) Plotted Against Data Set Size

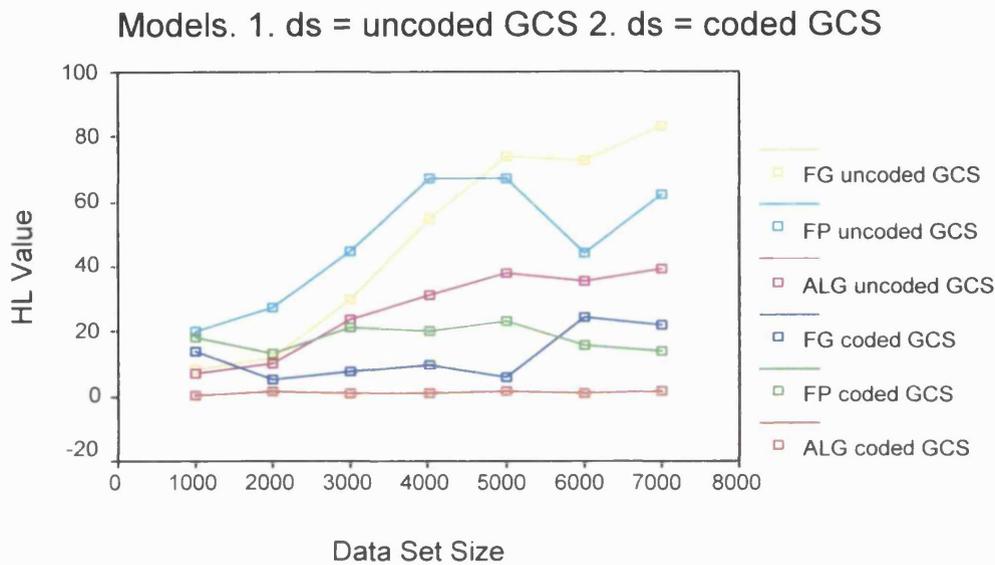


**Graph 3**

Graph 3 shows that model 1 again has the greatest degree of over-fitting. Model 2 also show predominantly over-fitting but less so than model 1. Random recoding the HCISS model into 12 groups (model 3) produced predominantly under-fitting rather than over-fitting as was originally hypothesised.

## RESULTS: Study 2

### HL Values (All Three Methods) Plotted Against Data Set Size

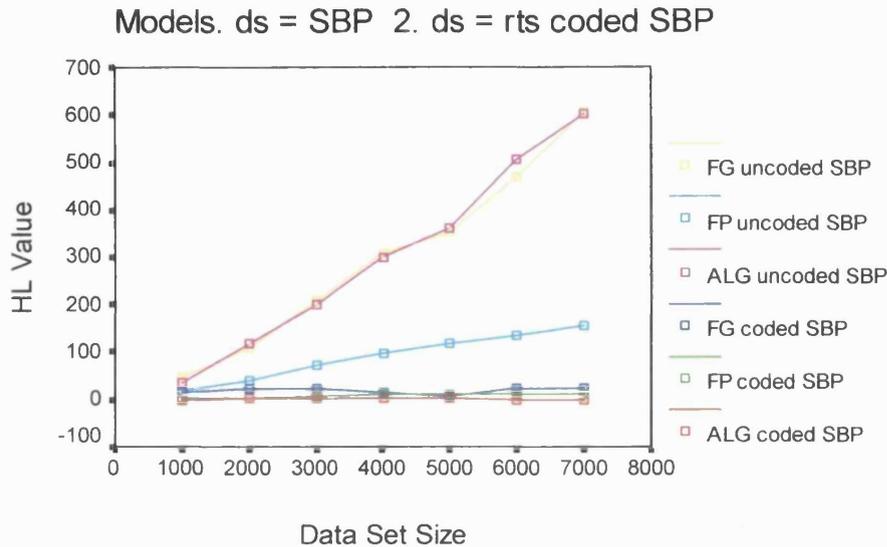


**Graph 4**

Graph 4 shows that for nearly all data set values reducing the covariate pattern for the GCS variable results in an apparent improvement in the goodness of fit for the model. This effect occurs for all three methods of calculating the Hosmer Lemeshow statistic. The most dramatic change is seen for the algorithm method. All three tests suggests that the reduced model (coded GCS model) has a much better fit than the uncoded GCS model as a simple consequence of reducing the covariate groupings.

### Results : Study 3

#### HL Values (All Three Methods) Plotted Against Data Set Size

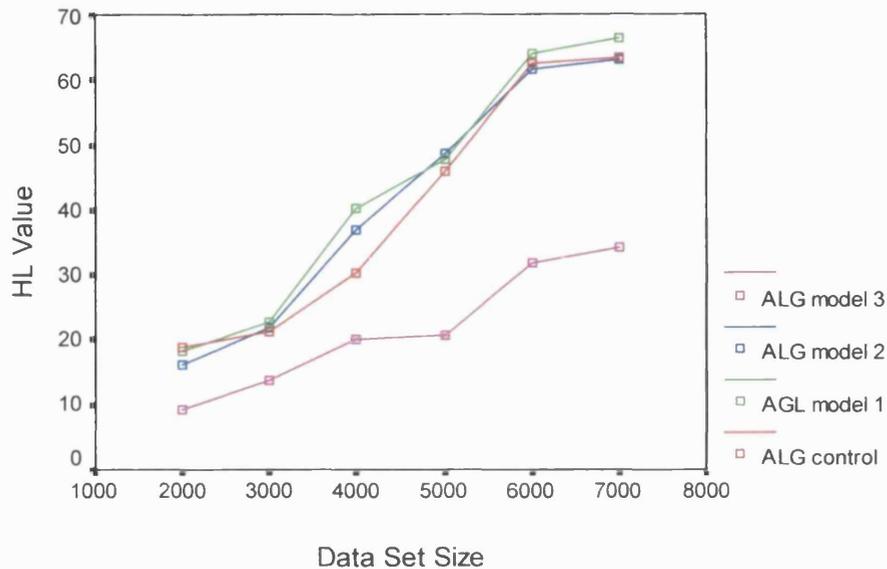


**Graph 5**

Graph 5 shows that for nearly all data set values reducing the covariate pattern for the SBP variable results in an apparent improvement in the goodness of fit for the model. This effect occurs for all three methods of calculating the Hosmer Lemeshow statistic. All three methods produce similar values for the reduced (coded SBP) model. All three tests suggests that the reduced model (coded SBP model) has a much better fit than the uncoded SBP model as a consequence of reducing the covariate groupings.

## RESULTS: Study 4

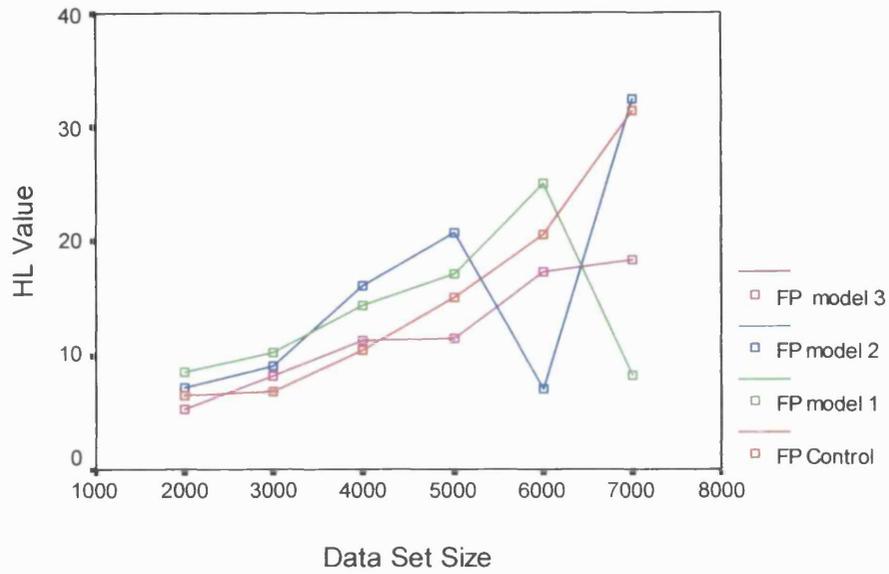
HL Value (Algorithm Method) Plotted  
Against Data Set Size



**Graph 6**

Graph 6 shows that only model 3 (smallest age covariate group) shows an appreciable difference from the control model. The difference is present over all data set values. Reducing the number of age groupings had little impact on the HL value except for the smallest grouping pattern.

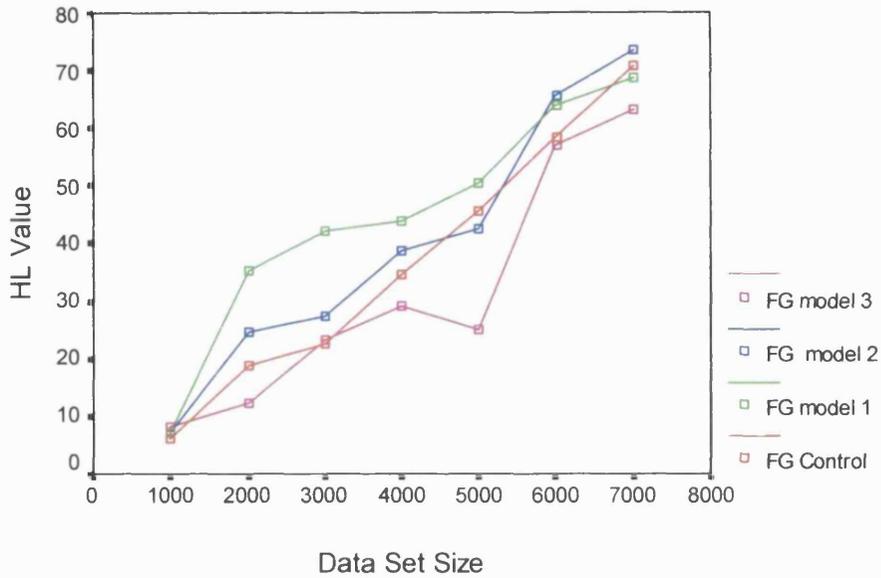
## HL Value (Fixed Percentile Method) Plotted Against Data Set Size



**Graph 7**

Graph 7 shows that changing the age covariate pattern has a variable effect on the HL value compared to the control model when using the Fixed Percentile method.

## HL Value (Fixed Group Method) Plotted Against Data Set Size

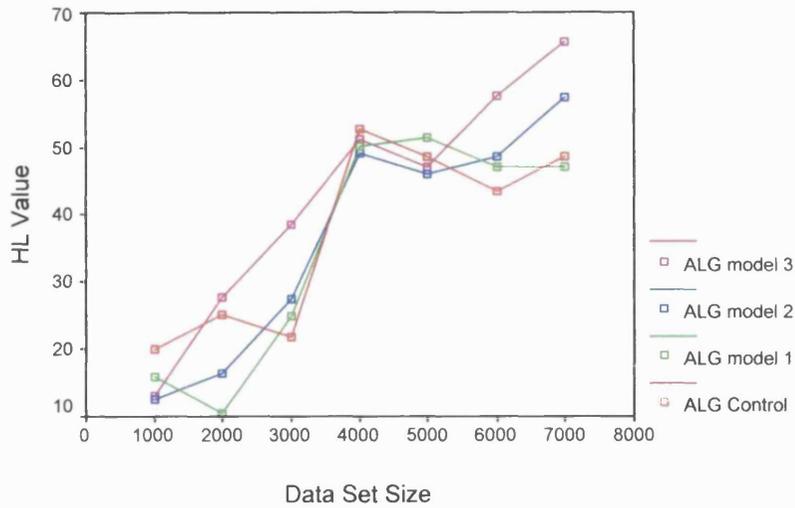


**Graph 8**

Graph 8 shows an increase in the HL value for model 1 and 2 compared to the control for the majority of the data set values. For model 3 the majority of the HL values are lower than the control.

## RESULTS: Study 5

HL Values (Algorithm Method) Plotted  
Against Data Set Size

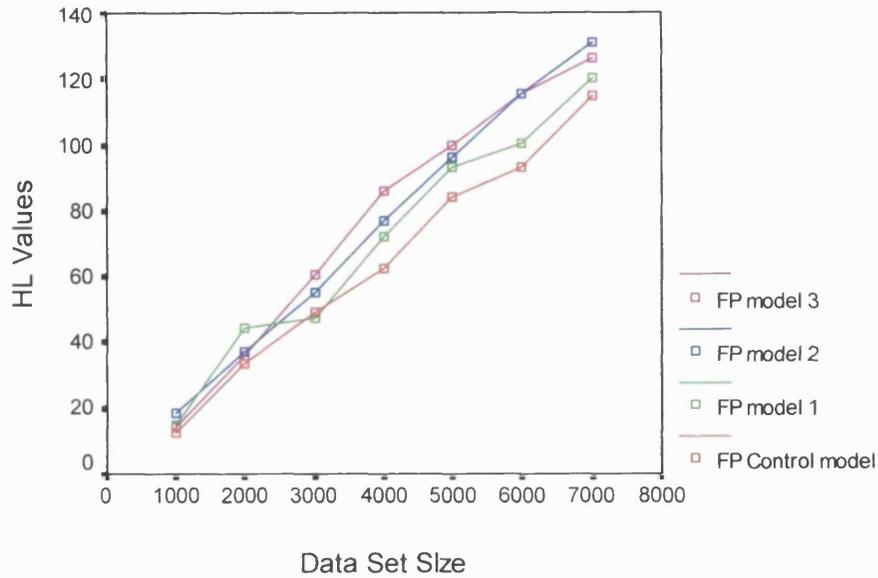


**Graph 9**

Graph 9 shows no clear effect after changing the age covariate pattern on the HL value for the HCISS + Age model using the Algorithm method.

## HL Values (Fixed Percentiles Method) Plotted

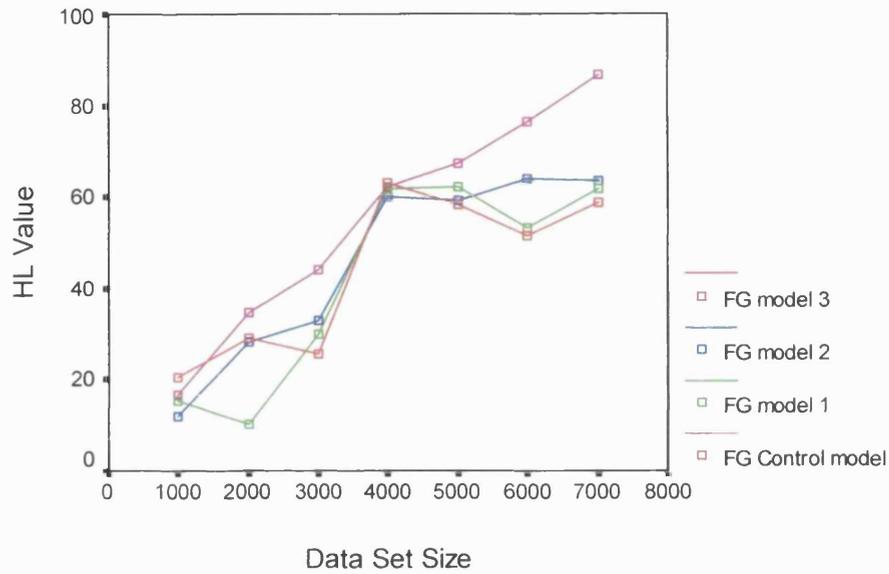
### Against Data Set Size



**Graph 10**

Graph 10 shows a marginal increase in the HL values for the three models when compared to the control model (HCISS + Age). Changing the age covariate pattern therefore had little effect on the HCISS + Age model using the fixed percentile HL method.

## HL Values (Fixed Group Method) Plotted Against Data Set Size

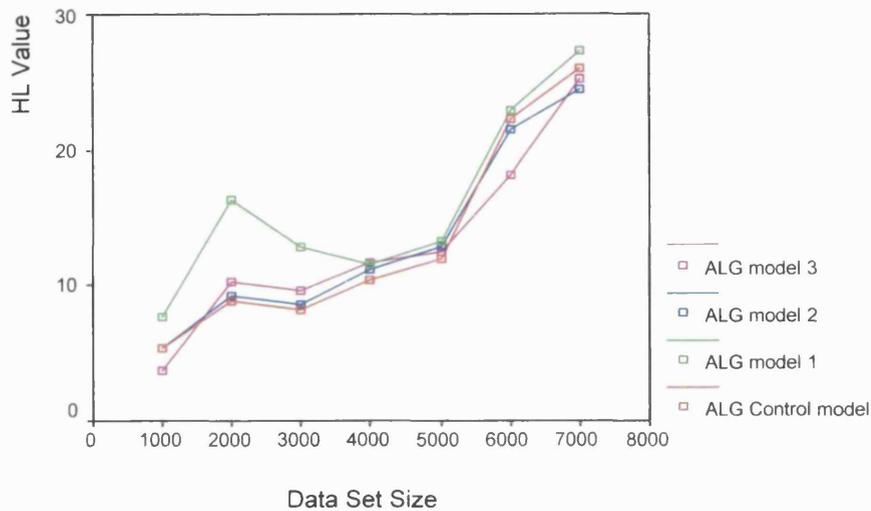


**Graph 11**

Graph 11 shows no clear effect on the HL values when compared to the control model (HCISS +Age). Changing the age covariate pattern therefore had little effect on the HCISS + Age model using the fixed group HL method.

## RESULTS: Study 6

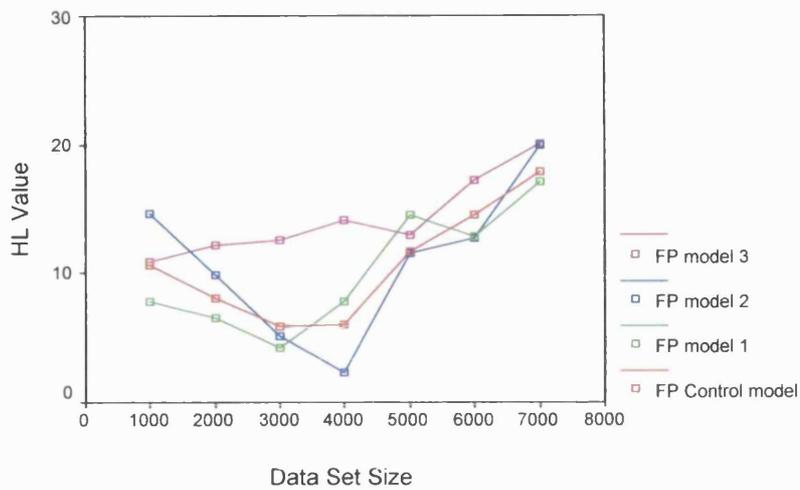
HL Value (Algorithm Method) Plotted  
Against Data Set Size



**Graph 12**

Graph 12 shows no real no appreciable effect on the HL values when compared to the control model (HCISS + RTS + Age) except for model 2, values 2000 and 3000. Changing the age covariate pattern therefore had little effect on the HCISS + RTS + Age model using the algorithm HL method.

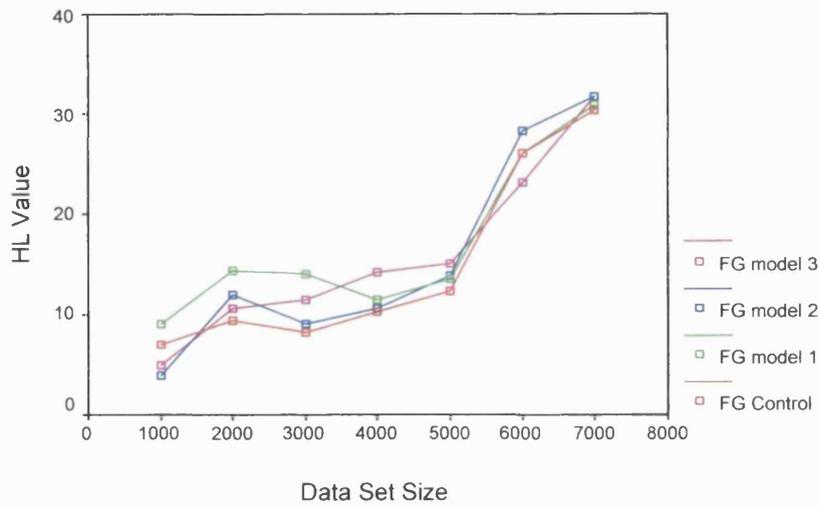
HL Value (Fixed Percentile Method) Plotted  
Against Data Set Size



**Graph 13**

Graph 13 shows that model 3 has a higher HL value over the full range of data set values when compared to the control model. The remaining two models (1 and 2) have values which fluctuate around the control model (HCISS + RTS + Age). Changing the age covariate pattern therefore had an unpredictable effect on the HCISS + RTS + Age model using the fixed percentile HL method.

### HL Value (Fixed Group Method) Plotted Against Data Set Size



**Graph 14**

Graph 13 shows little effect of changing the age covariate pattern on the HL value when compared to the control model. Changing the age covariate pattern therefore had little effect on the HCSS + RTS + Age model using the fixed group HL method.

## **Section 5: Discussion**

The results from these studies shows the variable effect on the HL statistic by reducing the covariate pattern. Study 1 demonstrated that using the fixed percentile method reducing the HCISS model into 6 or 12 groups resulted in model over-fit. This effect was greatest for the model with 6 groups using clinical cut-off points and least for the model with 12 random groups. The effect was more erratic with the other two methods of calculating the HL statistic. Using increase in data set size could be a confounding factor because of the corresponding potential increase in HCISS groups. An analysis of this showed that for a data set size of 1000 cases 38 HCISS groups were represented. For a data set size of 7000 only an additional 3 HCISS groups were added.

The problem of model over-fit by reducing the number of covariate groups was most marked for the model with SBP as the sole predictor variable (study 3). The HL results for this model imply significant over-fit as the HL values are close to zero for all three methods of calculating the HL test. A similar effect was seen in the model where GCS was the sole predictor variable (study 2). Coding the GCS using the triage RTS values again resulted in over prediction of the model fit. The effect was most marked for the algorithm method. The model with age as the sole predictor variable demonstrated a variable response (study 4). Only the smallest covariate pattern resulted in over prediction of the model, the effect was seen for all three methods of calculating the HL statistic.

In study 5 the effect of reducing the age covariate pattern on the model with two predictor variables (HCISS + Age) had little effect in terms of over prediction. The fixed percentile method resulted in slight under prediction for all three models compared to the control. The results for the fixed percentile method were more consistent compared to the algorithm and fixed group methods. The algorithm and fixed group methods produced some under and some over prediction for the three models. The model with the smallest covariate pattern (model 3) resulted in the greatest degree of over prediction compared to the control model. In study 6 the model with three predictors also produced variable results. The algorithm and fixed group methods showed no real trend with some over prediction and some under prediction.

Hosmer et al (1988) found that small cell sizes (less than 5) can result in large values of the HL test. No previous studies have looked specifically at the effect of changing the covariate pattern on the HL statistic.

## **Conclusions**

In summary the results of this study have shown that reducing the covariate pattern of a variable by recoding can have an appreciable impact on the HL value and may result in over prediction of the model goodness of fit. The effect is variable dependent. The effect was less with models with more than one predictor variable. All three methods of calculating the HL value were sensitive to changes in covariate pattern.

# **CHAPTER 7**

## **A COMPARISON OF SIX GOODNESS OF FIT TESTS BY SEQUENTIAL INCREASE IN DATA SET SIZE**

### **CONTENTS**

	<b>Page Number</b>
<b>Section 1: Introduction</b>	<b>148</b>
<b>Section 2: Aims</b>	<b>149</b>
<b>Section 3: Methodology</b>	<b>150</b>
<b>Section 4: Results</b>	<b>156</b>
<b>Section 5: Discussion</b>	<b>176</b>
<b>Appendix 1:</b>	<b>185</b>
<b>Appendix 2:</b>	<b>186</b>
<b>Appendix 3:</b>	<b>187</b>
<b>Appendix 4:</b>	<b>188</b>

## **Section 1: Introduction**

A variety of goodness of fit tests have been developed for the logistic regression model although only a few have been incorporated into statistical software packages. The Hosmer Lemeshow goodness of fit test combined with the area under the ROC has become the most common method for evaluating trauma scoring models. The Hosmer Lemeshow chi-squared test discussed previously in chapter 1 has several weaknesses. Firstly, its value increases often erratically as the data set size increases. Secondly, by simply reducing the size of the data set the goodness of fit of a logistic model can be changed from being a poor fit to a good fit. A number of other goodness of fit tests have been developed which may be potential alternatives to the Hosmer Lemeshow test. The aim of this chapter was to evaluate several of these tests using the revised USC data set.

## **Section 2: Aims**

The aim of this study was to determine the ability of five goodness of fit tests to distinguish between a model using the Injury Severity Score as the only predictor variable and a model using both the Injury Severity Score and the unweighted RTS as predictor variables. The Hosmer Lemeshow statistic was used as a comparison.

### **Hypothesis to be tested.**

That all five goodness of fit tests would be able to distinguish between the HCISS model and the HCISS + unweighted RTS model over a range of data set values.

## **Section 2: Methodology**

### **Section 2.1 Statistical software.**

SAS version 8.0 was used to perform the logistic regression modelling and the simulation studies. SPSS was used to plot the graphs.

**Section 2.2 Data Set:** The revised USC data set was used.

### **Section 2.3 Method for Model Selection**

Backward LR test (likelihood ratio).

(The calculated probabilities were for survival rather than death).

### **Section 2.4 Method of Validation**

Model validation was achieved by internal validation i.e. the goodness of fit tests were applied to the data set on which the models were derived.

### **2.5 Goodness of fit tests**

The five goodness of fit tests which were chosen for evaluation were :-

- (1) The Copas Test (Copas, 1989).
- (2) A modified standardized residuals test (le Cessie, 1991).
- (3) The Pearson chi-squared test.
- (4) The Deviance statistic.
- (5) The Brier test (Brier, 1950).

The Hosmer Lemeshow test (SAS algorithm method) was also evaluated as a comparison. Programs to calculate the Copas test, a modified standardized residuals test and the Brier test were written by this author using the SAS language (appendix 4). All of the above goodness of fit tests were incorporated into a simulation study of sequential increase in data set size.

The computation for the five tests are given below.

1. **Copas Test** - unweighted residual sum of squares (Copas, 1989).

$$S = \sum (y_i - \pi_i)^2$$

Where S is the test statistic

$y_i$  is the dependent variable value (one or zero) for the *ith* case.

$\pi_i$  is the calculated probability from the logistic model for the *ith* case.

2. **A modified ('Unweighted') Standardized Residuals Test.**

(for brevity it will be referred to as the USR test)

$$r = \sum [(y_i - \pi_i)] / \sqrt{\{\pi_i (1 - \pi_i)\}}$$

The value  $[(y_i - \pi_i)]$  is the absolute value not the arithmetic value.

r is the test statistic.

$y_i$  is the dependent variable value (one or zero) for the *ith* case.

$\pi_i$  is the calculated probability from the logistic model for the *ith* case.

This test statistic is a simplified version of the smoothed standardized residuals test proposed by le Cessie et al (1991). The smoothed standardized test multiplies the standardized residual by the weighted average of the residuals in a specified region:- the bandwidth, which is determined by a kernel function. The kernel function estimates the function of the logit model through a limited region of the model.

### 3.a Pearson chi-squared statistic (as calculated in *proc logistic*).

$$\chi^2 = \sum_{i=1}^I (y_i - \pi_i)^2 / \pi_i$$

$y_i$  = number of events (i.e. survivors) in covariate group i.

$\pi_i$  = expected number of events in covariate group i. The expected number of events equals the sum of the predicted probabilities for group i.

### 3.b Pearson chi-squared statistic (as calculated in *proc genmod*).

$$\chi^2 = \sum (y_i - \pi_i)^2 / \pi_i(1 - \pi_i)$$

SAS has two methods for calculating the Pearson chi-square statistic. In the *proc logistic* program the Pearson chi-square statistic is calculated by grouping the data into covariate groups. Each covariate group containing predictor variables with the same set of values. The degrees of freedom being equal to the number of covariate groups less than the number of estimated parameters in the model (which includes the intercept).

In the *proc genmod* program of SAS the Pearson chi-square is calculated for each case. The final Pearson chi-square statistic being the summation of all the individual chi-square values. This statistic is sometimes referred to as the global chi-square test (e.g. in SPSS) The degrees of freedom being the number of cases in the data set minus the number of estimated parameters in the model (which includes the intercept).

## 5. Deviance Statistic

SAS also has two methods of calculating the Deviance statistic. In *proc genmod* the Deviance is equal to twice the positive difference between the log-likelihood for the fitted model and the log-likelihood for the saturated model. The log-likelihood for the saturated model is 0, therefore the Deviance = -2 x log-likelihood (abbreviated form; -2LL) for the fitted model. A saturated model has one parameter for every probability and therefore has a perfect model fit.

In *proc logistic* the Deviance:-

$$D = 2 \sum_{i=1}^I O_i \log \frac{O_i}{E_i}$$

$O_i$  = number of events (i.e. survivors) in covariate group  $i$ .

$E_i$  = expected number of events in covariate group  $i$ . The expected number of events equals the sum of the predicted probabilities for group  $i$ .

The Deviance statistic in *proc logistic* is calculated by dividing the data into covariate groupings. The groupings being the same as that used to calculate the Pearson chi-square statistic in *proc logistic*.

For both the *proc logistic* and *proc genmod* methods evidence for lack of fit occurs when the value of these statistics (i.e. Pearson and Deviance) are large. P values are calculated using the  $n - p - 1$  distribution. P values are not however considered to be valid for either statistic when using a binary logistic model because the number of covariate patterns are not fixed (Hosmer, 1997). It should be noted that the log-likelihood can also be used as a method of model development (see chapter 4). The statistic  $-2 \times$  log-likelihood is computed as part of the output for model development in SPSS.

#### 5. Brier test.

$$Br = 2/n \sum(\pi_i - y_i)^2$$

*Br* is the test statistic.

$y_i$  is the dependent variable value (one or zero) for the *ith* case.

$\pi_i$  is the calculated probability from the logistic model for the *ith* case.  $n$  is the total number of cases.

Two models were used to evaluate the goodness of fit tests:-

#### **Model 1:**

Dependent variable:- death/survival

Independent variable:- HCISS

#### **Model 2:**

Dependent variable:- death/survival

Independent variables:-

HCISS + Revised Trauma Score (unweighted)

## **Section 2.6 Simulation Study**

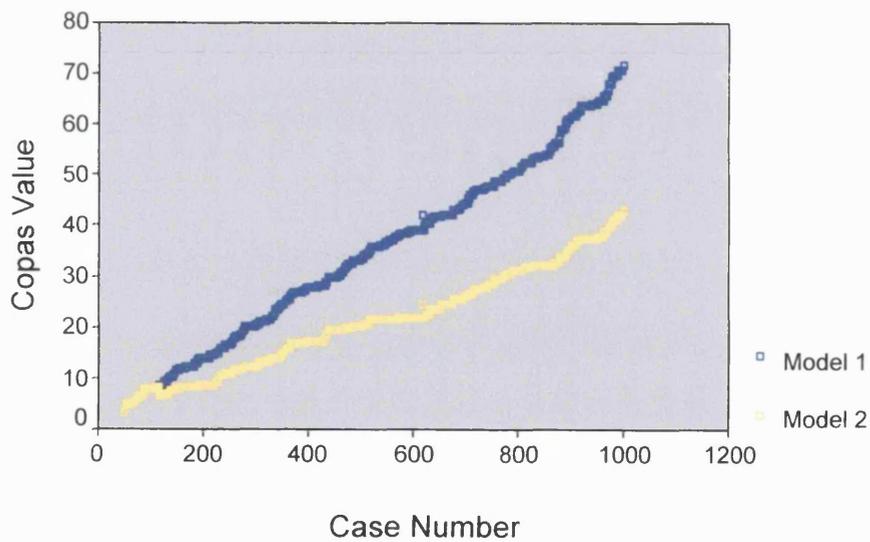
The two models were developed on the first 50 cases of the revised USC data set. Logistic regression was performed using the *proc genmod* and *proc logistic* programs in SAS. The Pearson, Deviance and Hosmer Lemeshow goodness of fit tests were outputted using the *ods* option in SAS. The values for the other goodness of fit tests were determined by writing a SAS program as previously mentioned in section 2.5. Using a *macro repeat* program the sequential goodness of fit test results were generated each time the data set was incrementally increased from case number 51 to 1000 (the ‘small’ data set). The same procedure was repeated by developing the two models on the first 6050 cases of the revised USC data set. Further goodness of fit test results were generated by sequentially increasing the size of the data set up to case number 7000 (the ‘large’ data set). The 951 goodness of fit test results for both the small and the large data sets were plotted for the two models against sequential increase in data set size. The programs for this simulation study are given in appendix 1 to 4.

## Section 3: Results

### Section 3.1: Copas Goodness of Fit Test.

Copas Value Plotted Against Sequential Increase In Data Set Size

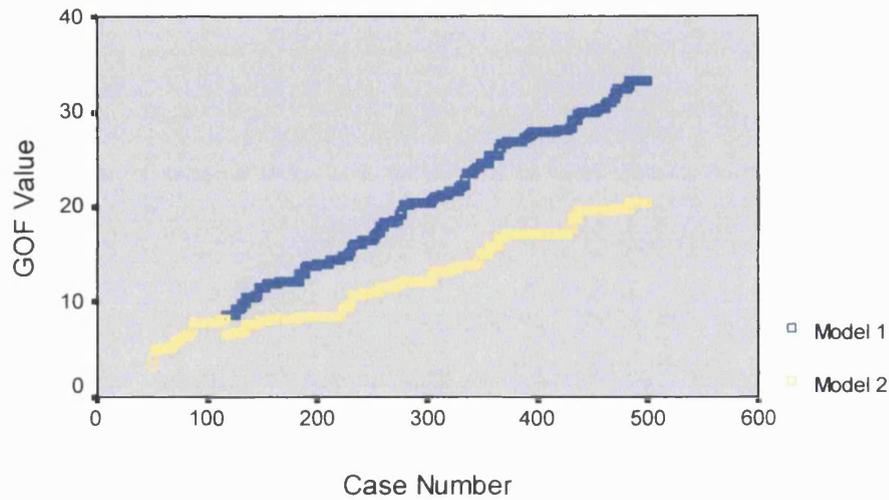
Models:- 1. DS=HCISS. 2. DS=HCISS+RTS



**Graph 1a**

Copas Value Plotted Against Sequential Increase in Data Set Size. N=50-500.

Models:- 1. DS=HCCISS. 2. DS=HCCISS+RTS

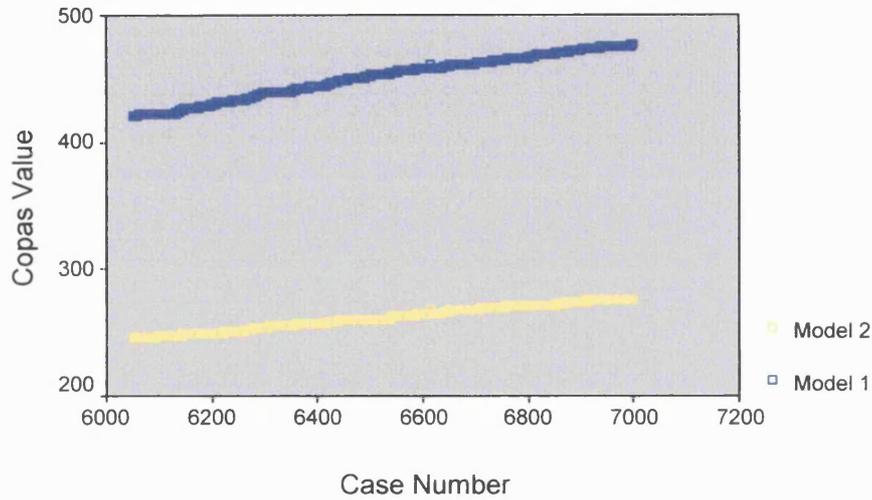


**Graph 1b**

Graph 1a and 1b show that with a data set of just over 100 cases the Copas test is able to correctly distinguish between a poor model (model 1) and a superior model (model 2). Note the relatively smooth increase in the Copas values with sequential increase in data set size.

## Copas Value Plotted Against Sequential Increase In Data Set Size (6050-7000)

Models:- 1. DS=HCCISS 2. DS=HCCISS+RTS



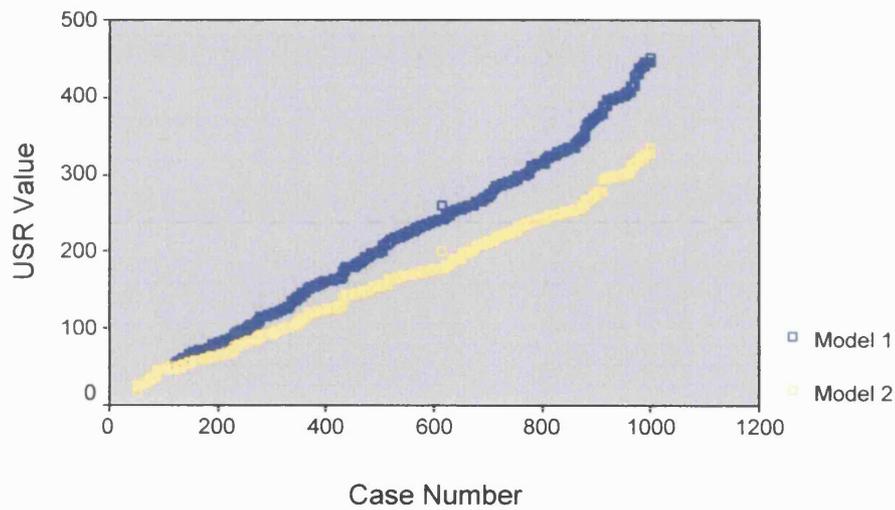
**Graph 1c**

Graph 1c shows that for the larger data set the Copas test is able to clearly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. Note the smoothness of the two lines as the Copas values increase with sequential increase in data set size.

## Section 3.2

### Unweighted Standardized Residual goodness of fit test (USR).

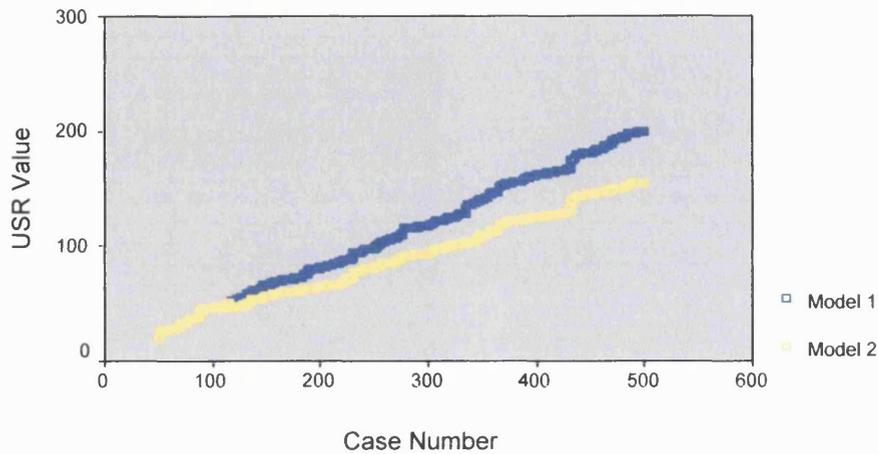
USR Value Plotted Against  
Sequential Increase In Data Set Size  
Models:- 1. DS=HCCISS. 2. DS=HCCISS+RTS



**Graph 2a**

USR Value Plotted Against Sequential Increase in Data Set Size. N=50-500.

Models:- 1. DS=HCCISS. 2. DS=HCCISS+RTS

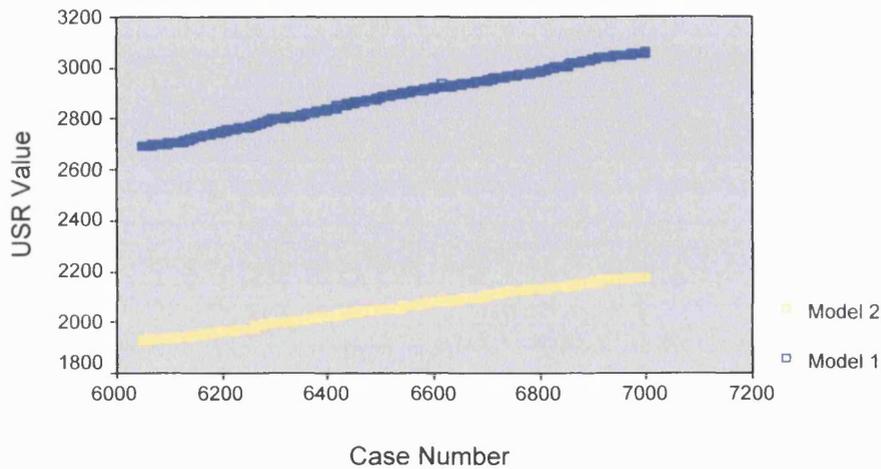


**Graph 2b**

Graph 2a and 2b show that with a data set of just over 100 cases the USR test is able to correctly distinguish between a poor model (model 1) and a superior model (model 2). Note the relatively smooth increase in the USR values with sequential increase in data set size.

## USR Values Plotted Against Sequential Increase In Data Set Size (6050-7000)

Models:- 1. DS=HCCISS 2. DS=HCCISS+RTS



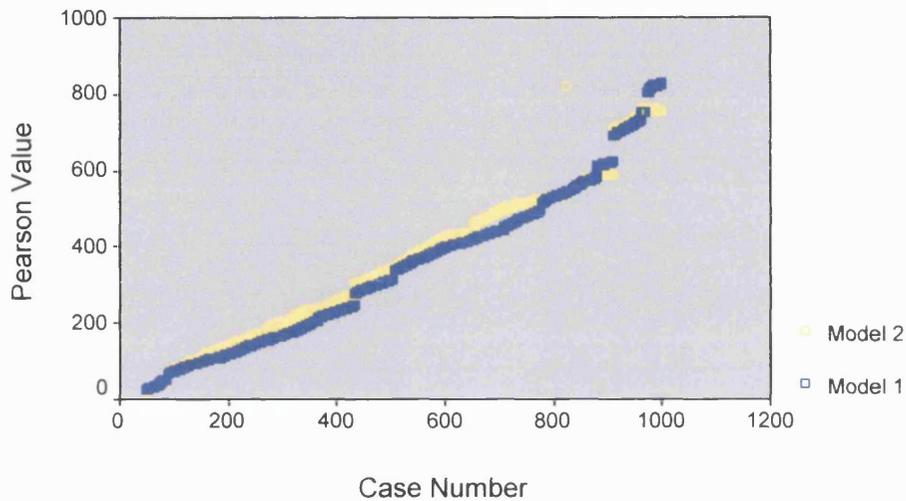
### Graph 2c

Graph 2c shows that for the larger data set the USR test again can clearly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. Note the smoothness of the two lines as the USR values increase with sequential increase in data set size.

### Section 3.3: Pearson Goodness of Fit Test.

Pearson Value Plotted Against Sequential Increase In Data Set Size (50-1000)

Models:- 1. DS = HCISS 2. DS = HCISS+RTS

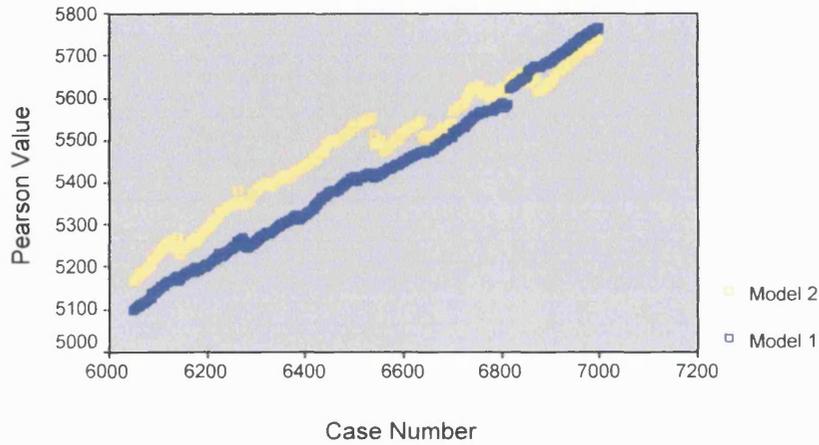


**Graph 3a<sub>1</sub>** (*proc genmod*)

Graphs 3a<sub>1</sub> shows that the Pearson chi-square test is unable to distinguish between the poor model (model 1) and the superior model (model 2) for virtually all data set values. The Pearson test predicted only 13.4% of cases correctly (n=127; there were no ties).

Pearson Value Plotted Against Sequential  
Increase In Data Set Size (6050-7000)

Models:- 1.DS=HCCISS 2. DS=HCCISS+RTS

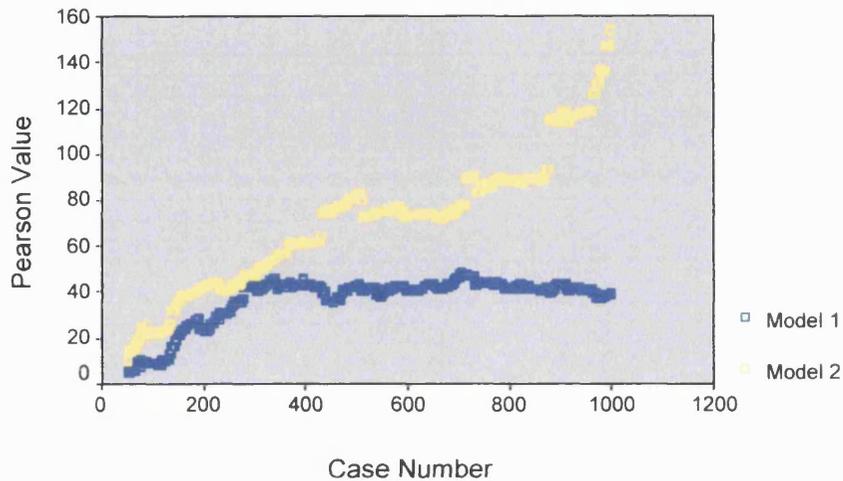


**Graph 3b<sub>1</sub>** (*proc genmod*)

Graph 3b<sub>1</sub> shows that for the larger data set the Pearson chi-square test again had poor ability to distinguish between the poor model (model 1) and the superior model (model 2) for the majority of data set values. The Pearson test predicted only 15.6% of cases correctly (n=148; there were no ties). All of the correct prediction cases were at the upper end of the data set range.

Pearson Value Plotted Against Sequential  
Increase In Data Set Size (proc logistic)

Models:- 1.DS=HCCISS 2. DS=HCCISS+RTS

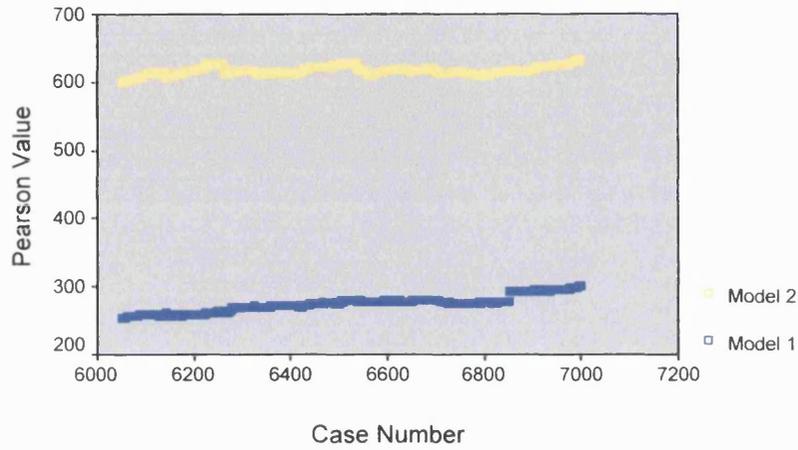


**Graph 3a<sub>2</sub>** (*proc logistic*)

Graphs 3a<sub>2</sub> shows that the Pearson chi-square test is unable to correctly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. The difference is exemplified as the data set size increases.

Pearson Value Plotted Against Sequential Increase In Data Set Size. (proc logistic)

Model:- 1. DS=HCISS 2. DS=HCISS+RTS

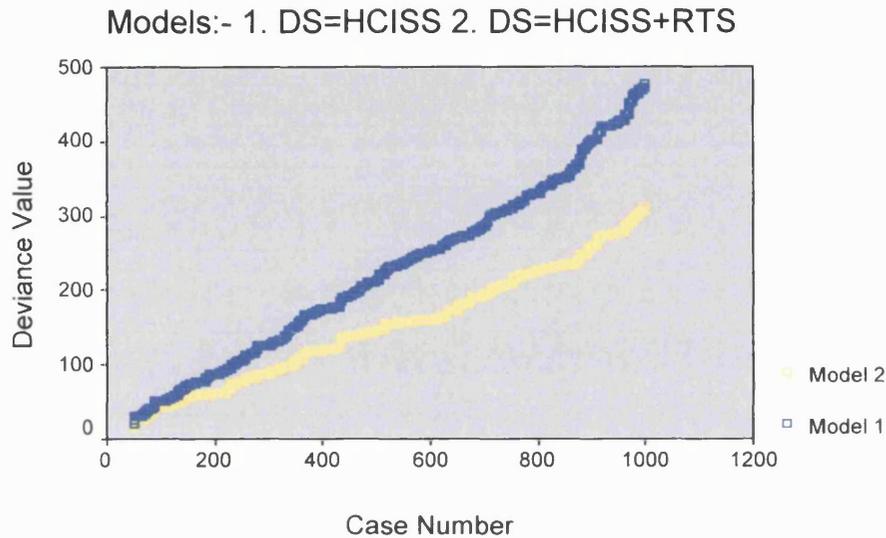


**Graph 3b<sub>2</sub>** (*proc logistic*)

Graphs 3b<sub>2</sub> shows that the Pearson chi-square test is unable to correctly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values.

### Section 3.4 Deviance Statistic

#### Deviance Value Plotted Against Sequential Increase In Data Set Size

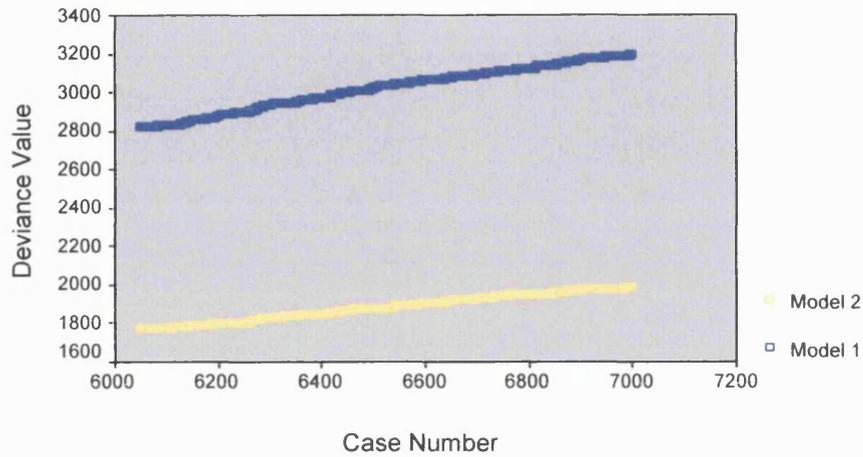


**Graph 4a<sub>1</sub>** (*proc genmod*)

Graph 4a shows that the Deviance statistic was able to correctly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. Note the relatively smooth increase in the deviance values with sequential increase in data set size.

## Deviance Value Plotted Against Sequential Increase In Data Set Size

Models:- 1.DS=HCCI 2. DS=HCCI+RTS

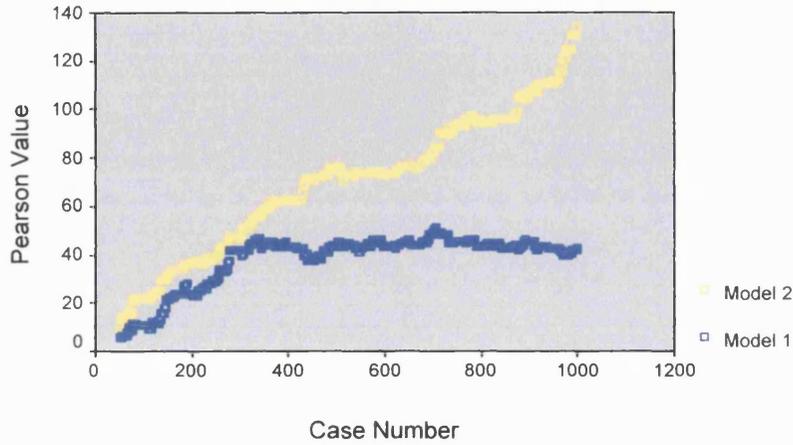


**Graph 4b<sub>1</sub>** (*proc genmod*)

Graph 4b shows that for the larger data set the Deviance statistic is again able to clearly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. Note the smoothness of the two lines as the deviance values increase with sequential increase in data set size.

Deviance Value Plotted Against Sequential  
Increase In Data Set Size (proc logistic)

Models 1.HCISS 2.HCISS+RTS

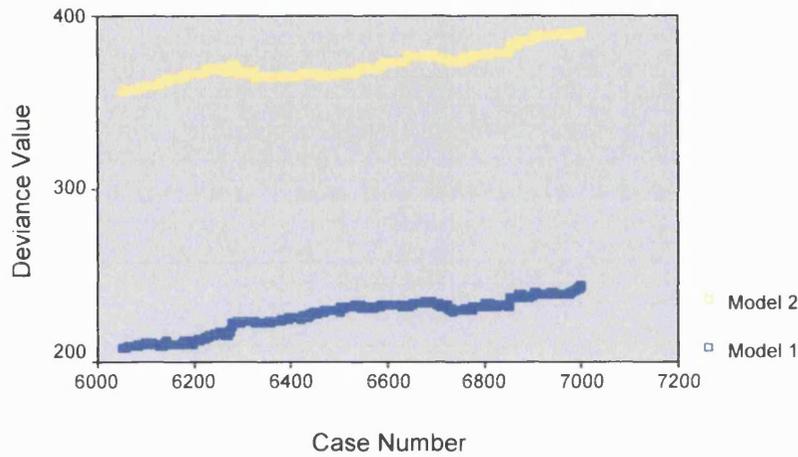


**Graph 4a<sub>2</sub>** (*proc logistic*)

Graphs 4a<sub>2</sub> shows that the Deviance chi-square statistic is unable to correctly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. The difference is exemplified as the data set size increases.

Deviance Value Plotted Against Sequential Increase In Data Set Size. (proc logistic)

Models:- 1. DS=HCCISS 2. DS=HCCISS+RTS



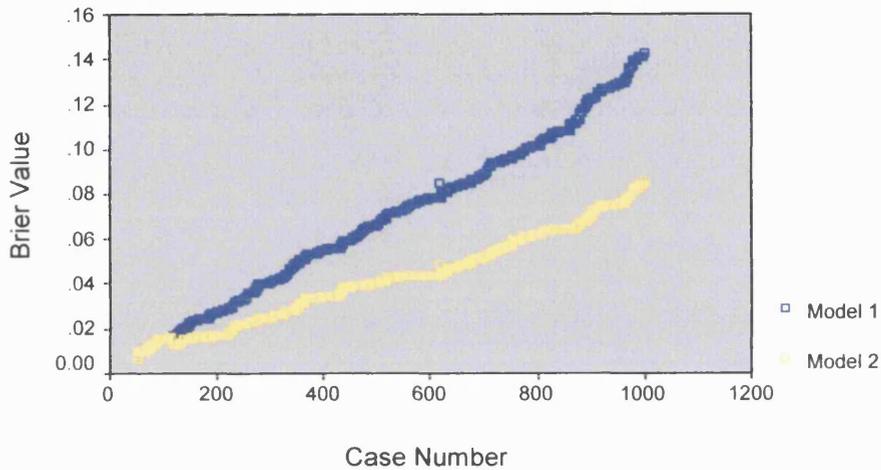
**Graph 4b<sub>2</sub>** (*proc logistic*)

Graphs 4b<sub>2</sub> shows that the Deviance chi-square statistic is unable to correctly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values.

### Section 3.5 Brier Goodness of Fit Test.

Brier Value Plotted Against Sequential Increase In Data Set Size

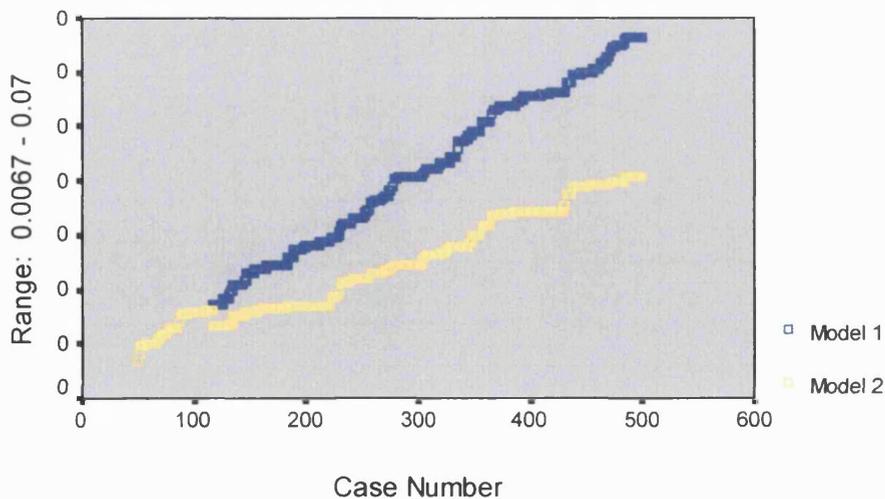
Models:- 1. DS=HCISS. 2. DS=HCISS+RTS



**Graph 5a**

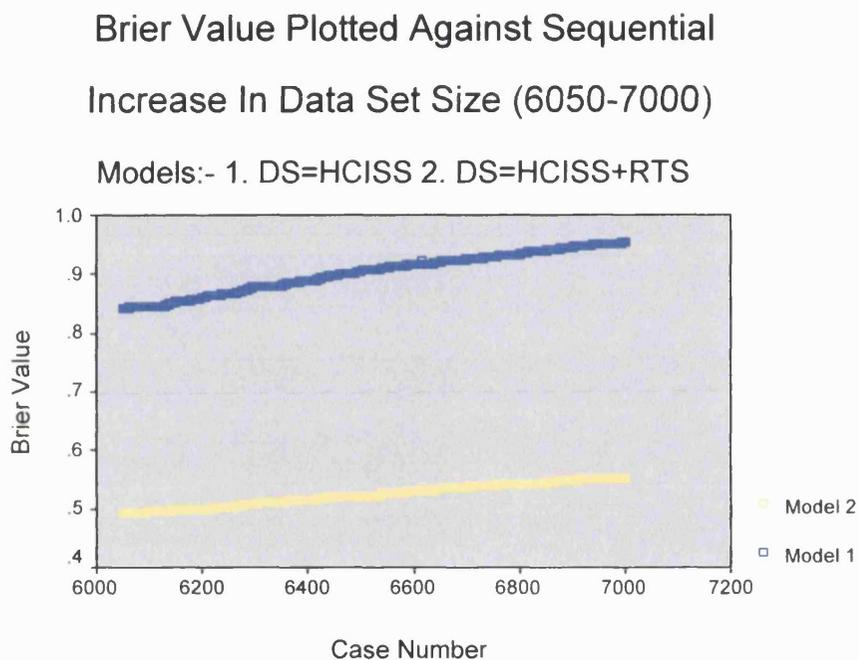
Brier Value Plotted Against Sequential Increase in Data Set Size. N=50-500.

Models:- 1. DS=HCISS. 2. DS=HCISS+RTS



**Graph 5b**

Graph 5a and 5b show that with a data set of just over 100 cases the Brier test is able to correctly distinguish between the poor model (model 1) and the superior model (model 2). Note the relatively smooth increase in the Brier values with sequential increase in data set size.



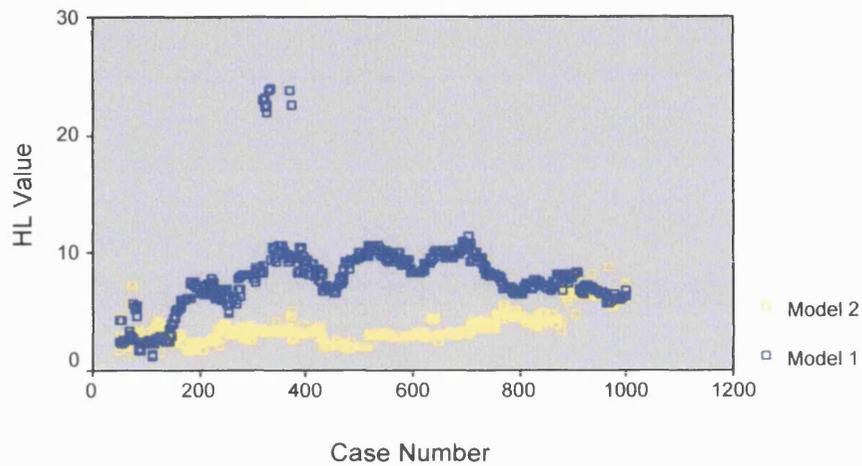
### Graph 5c

Graph 5c shows that for the larger data set the Brier test is again clearly able to correctly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. Note the smoothness of the two lines as the Brier values increase with sequential increase in data set size.

### Section 3.6 Hosmer Lemeshow goodness of fit test

HL Value (algorithm method) Plotted Against  
Sequential Increase In Data Set Size

Model:- 1. DS=HCISS 2. DS=HCISS+RTS

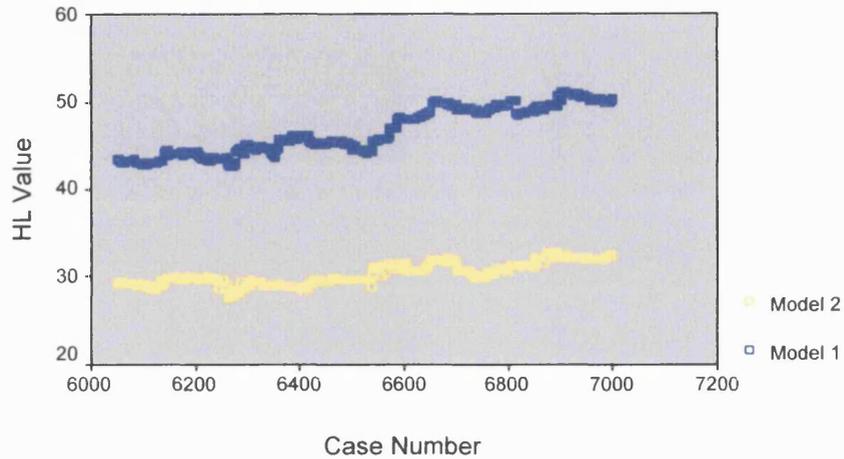


**Graph 6a**

Graph 6a shows that in the majority of cases (91.2%; n= 867, there were no ties) the HL value for the HCISS model was higher than the HL value for the HCISS plus RTS model. Both models show fluctuating values with sequential increase in data set size, more so for the HCISS model. At the upper and lower data set values the HL test fails to distinguish between the two models.

HL Value (algorithm) Plotted Against  
Sequential Increase In Data Set Size

Models:- 1.DS=HCISS 2. DS=HCISS+RTS



**Graph 6b**

Graph 6b shows that for the larger data set the HL test is able to clearly distinguish between the poor model (model 1) and the superior model (model 2) for all data set values. Both models show fluctuating values with sequential increase in the data.

### **Section 3.7 Further Analysis of Results**

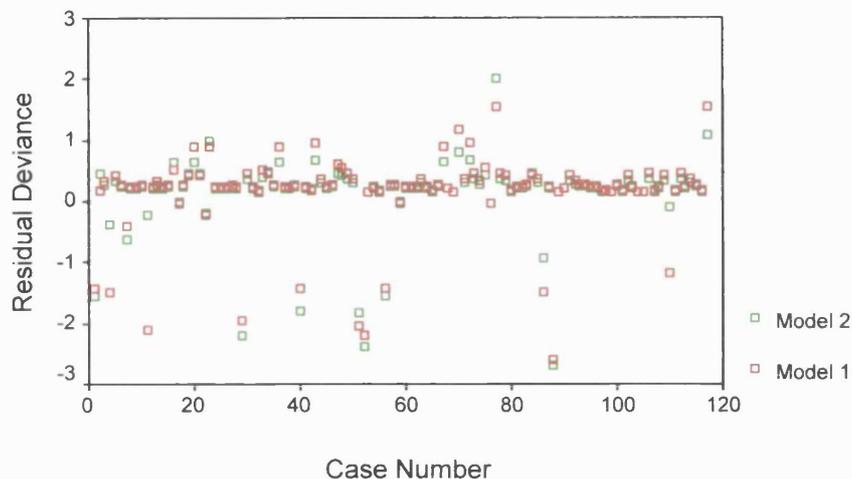
Three of the goodness of fit tests (Copas, USR and Brier) were unable to distinguish between the two models (HCISS and HCISS + RTS) up to case number 116. This was inferred by the fact that the two lines were superimposed. The latter finding was validated by checking the actual results. At case 117 the model 2 line abruptly diverges away from the model 1 line. The values of the variables for case 117 being HCISS = 34, RTS = 12, Outcome = 1 (survival). The predicted probability for survival for this case changes from 0.30 for model 1 to 0.56 for model 2 when the data set size was 117 (i.e. cases 1-117 inclusive). This unexpected survivor for model 1 becomes an expected survivor in model 2 and results in separation in the two curves. This is reflected in the large residual deviance\* values for this case (Graph 8: last two squares on the graph). The presence of other cases with large residual deviance values (graph 8) prior to case number 117 did not result in separation of the two lines. From case 117 three of the goodness of fit tests studied (i.e. Copas, USR and Brier) correctly discriminated between the poor model (model 1) and the superior model (model 2) for all case values up to 1000.

\*The residual deviance =  $-2 \times \log$  (predicted probability for that case)

## Residual Deviance Value Plotted Against

Case Number. Data Set Size = 117

Model 1. DS=HCCI 2. DS=HCCI+RTS



### Graph 8: Residual Deviance Values

The residual deviance values for the two models were produced using a data set composed of the first 117 cases. The residual deviance values were generated using the logistic program in SPSS (backward likelihood ratio method). The latter method produced the same coefficient values for both models when compared to the *proc logistic* program in SAS. The residual deviance is not routinely outputted in SAS which is why the logistic program in SPSS was used.

## **Section 4: Discussion**

The basis of many simulation studies is to calculate the number of times a statistic correctly accepts the right model (or vice versa) using random number generation to produce the data set. Many simulation studies use specific distributions (e.g. binomial, Gauchy). These distributions do not accurately reflect the distribution of injury pattern seen in trauma data sets. For this reason the actual revised USC data set was used rather than a specific distribution generated by random numbers (Monte Carlo Simulation). Multiple values were generated by sequentially increasing the size of the data set thus maintaining the basic composition of the data set. P values were not used because all of the statistics used in the simulation study (except for the Brier test where p values have not been developed) have the inherent problem in that their test result increases as the data set size increases. A model which appears to fit well on a small data set can be made to fit poorly simply by increasing the size of the data set. Trauma scoring modelling requires large data sets so that a reasonable number of the less common injury patterns particularly in the middle range are represented. Reliance on p values in data sets containing several thousand cases may lead to rejection of potentially good fitting models if the p value is used.

The combination of a physiological variable (e.g. RTS) with an anatomical injury variable (e.g. ISS) is well recognised as being superior to a model containing only ISS (Champion, 1983) using death/survival as the dependent variable. The ability of a goodness of fit test to distinguish between these two models over a large range of data set values provides a reasonable assessment of its

accuracy in the context of trauma scoring. Internal validation was used rather than external validation methods such as data splitting or cross-validation firstly because opinion varies as to which is the superior method. Secondly, the use of data splitting or cross-validation methods may be a confounding factor in the analysis.

The Deviance statistic using *proc genmod* was able to distinguish between the two models for all data set values. This was in contrast to the deviance statistic using *proc logistic* which was unable to correctly distinguish between the two models for all data set values. The Pearson statistic was able to correctly distinguish between the two models in only a small number of cases using *proc genmod* and in none of the cases using *proc logistic*.

The Copas test, the Brier test and the USR test were all able to distinguish between the above two models for all case values when the data set reached case number 117 (i.e. in 92.9% of cases). The HL test was able to correctly distinguish between the two models in 91.2% of cases for the smaller data set and in 100% of cases in the larger data set.

The results of this study suggest that the deviance statistic (-2LL) had the largest power to detect a difference between the Injury Severity Score model and the model with the Injury Severity Score variable plus the Revised Trauma Score variable. The Copas, Brier and USR all showed the same power to detect a difference between the two models. The latter three tests all showed poor performance with very small data set sizes. A simplified unweighted standardized residuals test (USR) was chosen for this study

because of uncertainty about the best way of calculating the bandwidth. This problem was highlighted by le Cessie et al in their original monograph (1991) and also by Hosmer et al (1997). In reality the smoothed standardized residuals test is probably too complex for routine use by non-statisticians and as a result has not been widely adopted. The simplified USR test however had a similar performance to the Copas and Brier tests.

The Brier test was developed in 1950 (Brier, 1950) as a means of evaluating weather forecasts when expressed as a probability. The test which was initially used in meteorology has recently become popular as a goodness of fit test in the logistic regression model.

The Brier Score in its full mathematical text is :-

$$Br = 2/n \sum_{i=1}^n \sum_{j=1}^r (f_{ij} - E_{ij})^2$$

where  $E_{ij}$  takes the value of 1 or zero depending on whether or not the  $j$  event (rain or no rain) occurred on the  $i$ th occasion. The score is the average squared probability and has a minimum value of zero (perfect forecasting) and 2 (worst possible forecasting). An illustrative example is given below as to how the Brier score is calculated for weather forecasting.

Ten forecasts are given (i.e.  $n=10$ ) for rain and no rain (i.e.  $r = 2$ ). The forecasts are 0.9, 0.7, 0.8, 0, 0.4, 0, 0, 0.1, 0.2, 0.1. It actually rained on the third, fourth and ninth occasions. The Brier score is therefore:-

$$1/10 \times 2 \times (0.9^2 + 0.7^2 + 0.2^2 + 1^2 + 0.4^2 + 0^2 + 0^2 + 0.1^2 + 0.8^2 + 0.1^2) = 0.632$$

The Brier test was found to have a similar performance to the Copas Test. The main disadvantage of this statistic is its small range of values which makes it difficult to judge the magnitude of a difference between two models.

Many goodness of fit tests have been proposed for assessing the logistic model. Kuss (2002) identified 28 goodness of fit tests in his review. Many of these goodness of fit tests have either not been fully evaluated, are not available in commercial software packages or are too complex to write as a program for the non-statistician. The Hosmer Lemeshow test has been adopted by many statistical packages and is generally considered to be the standard means of assessing the calibration aspect of goodness of fit. The Hosmer Lemeshow statistic does however have some limitations as outlined in chapters 5 and 6.

The Pearson chi-squared statistic and the Deviance statistic were two of the earliest tests used for assessing goodness of fit in the logistic model. The deficiencies of these two statistics are well recognised, most notably both tend to give erratic results when the cells contain sparse data (Pulkstenis, 2002). Neither statistic has a true chi-square distribution and there is a tendency for the statistics to over predict model fit i.e. underdispersion. The formulae for the Pearson chi-squared statistic and Deviance statistic are given in the methodology section. A third method of calculating the Pearson chi-square statistic is used by some statisticians\*\* (Hosmer, 1997).

$$** \chi^2 = \sum_{i=1}^I (y_i - \pi_i)^2 / \pi_i (1 - \pi_i)$$

$y_i$  = number of events (i.e. survivors) in covariate group  $i$ .

$\pi_i$  = expected number of events in covariate group  $i$ . The expected number of events equals the sum of the predicted probabilities for group  $i$ .

As mentioned previously the Deviance statistic as computed in *proc genmod* compares the fitted model with the saturated model. The Deviance statistic and the Pearson chi-square in *proc logistic* compare the fitted model with the null model (intercept only). Several other variations of the two methods given for calculating the Deviance statistic are available (Hosmer, 1997; Kuss, 2002). Although the Deviance statistic (-2 x log-likelihood method) had the greatest power to correctly distinguish between a poor model and a superior model the different ways of generating the statistic affords it little advantage over the Hosmer Lemeshow test. The other drawback with using the deviance statistic as a goodness of fit statistic is that the log-likelihood statistic is also used for model development in the likelihood ratio methods.

Farrington (1996) suggested a modified version of the Pearson chi-squared statistic\*\* (third method) by the addition of a constant. Unfortunately as  $m$  (the size of the covariate groups) approaches 1 (extreme sparseness) the  $\chi^2$  value approximates to  $N$  (the sample size) and hence provides little information about model fit.

The Unweighted Sum of Squares Test was advocated by Copas (1989) as an alternative to the standard Pearson chi-squared test. The test statistic gives greater weight to covariate groups with large values of  $m$  and proportionately smaller weight to covariate groups with smaller values of  $m$ .

The Unweighted Sum of Squares Test:-  $S = \sum(x-n\mu)^2$

$\mu$  is the weighted probability  $\mu = \sum x / \sum n^2$

$x$  represents the number of successes out of  $n$  trials.

The probability of a covariate group with  $x$  successes in a group of size  $n = \sum x/n$

The statistic advocated by Copas was modified by Hosmer et al (1997) into its more familiar form:-

$$S = \sum(y_i - \pi_i)^2$$

$y_i$ :- is the outcome variable in the logistic model.

$\pi_i$ :- is the probability for survival (or death) derived by the logistic model.

Hosmer et al (1997) provide details of how to calculate a p value for the test. The Copas statistic is a measure of *predictive accuracy* of the model in contrast to the Hosmer Lemeshow statistic which measures model calibration.

Le Cessie et al (1991) developed a smooth standardized residual test. As mentioned in the methodology section problems with defining the bandwidth has meant that the statistic has not been introduced into routine statistical practice. In this simulation study a simplified version appears to perform as well as the Copas and Brier tests.

Hosmer et al (1997) performed the first systematic simulation study of goodness of fit tests used in logistic regression. These authors studied the performance of the HL test in its different formats i.e. equal sized groups, fixed percentile and algorithm methods. The authors noted that different computer software packages (SYSTAT, STATISTIX, STATA, SAS, LOGXACT and BMDPLR) produced differing results due to the selection of different cut-off points for the deciles. The SPSS and SAS software packages were not included in this study. The other goodness of fit statistics included were the Pearson chi-square statistic\*\* (third method), the unweighted sum of squares test (Copas, 1989), a scores test for the logistic regression model (Stukel, 1988), the smooth standardized residuals (le Cessie, 1991 and Royston, 1992). The overall results from this study showed no superiority for any particular test statistic. The Pearson chi-square test and the Copas test had the highest power to detect omission of a quadratic term from the model. All tests had low power to detect a continuous dichotomous variable interaction. The authors made no strong recommendations about the use of any one particular statistic but highlighted the lack of power of all the goodness of fit tests to detect small deviations in goodness of fit when using small data sets.

A more recent simulation study of goodness of fit tests in logistic regression has been undertaken by Kuss (2002). In total 28 goodness of fit tests were studied although the results of only eight tests were given in his monograph. The Pearson (third method\*\*) and Deviance statistic (-2LL type method) were used as reference tests. The six remaining tests were three modified Pearson (third

method\*\*) chi-square tests ( $\chi^2_{\text{McC}}$  {McCullagh, 1985},  $\chi^2_{\text{F}}$  {Farrington, 1996} and  $\chi^2_{\text{O}}$  {Osious, 1992}), the Residual Sum of Squares Test (Copas, 1989), the Information Matrix (White, 1982) and the Hosmer Lemeshow statistic. The basis of the simulation studies involved an analysis of p values for several models with various sized data sets and covariate patterns. The author found that the Pearson and the Deviance chi-squared tests under predicted lack of fit in the majority of the simulation studies. The family of modified Pearson tests all behaved similarly although the Farrington test ( $\chi^2_{\text{F}}$ ) outperformed its counterparts. The Copas test performed better with a wrong functional form of the covariate compared to the HL test and the Farrington test ( $\chi^2_{\text{F}}$ ). The family of modified Pearson tests outperformed the Copas test in models with missing covariates. The HL test was found to perform in between these two test statistics (i.e. the modified Pearson tests and the Copas test).

As mentioned in chapter 5, Pigeon (1999b) proposed a modified version of the Pearson chi-squared statistic (*proc logistic* method) which he named the  $\mathcal{J}^2$  statistic. His simulation studies showed that the new statistic, which contained a correction factor had less tendency to underdispersion (model over-fitting) than the Pearson chi-squared test (*proc logistic* method). Further simulation studies are required to determine whether the correction factor ( $\Phi$ ) will prevent underdispersion in models derived from large data sets containing a large number of covariate patterns. In his simulation studies the maximum data set size was only 200.

Pulkstenis et al (2002) also proposed a modified Pearson and Deviance statistic (*proc logistic* methods). The basis of their new statistics was a subgrouping of the covariate patterns based upon a categorical predictor variable. Their simulation studies showed that both of their modified statistics had an increased power over the Pearson and Deviance statistics (*proc logistic* method) to detect a model which required an interaction term in order to achieve a statistically good model fit. The authors also advocated that a similar approach could be used to modify the HL statistic in order to identify the need for possible interaction terms. The modified HL test however seems to convey little advantage over the conventional HL methods. In fact one could predict that the results would be more erratic due to the further increase in the number of cells, thereby increasing the risk of producing cells with sparse data.

## Conclusions

The Deviance statistic had the best performance of all the tests when calculated as  $-2 \times \log\text{-likelihood}$ . Neither the Deviance statistic nor the Pearson chi-square statistic generated by *proc logistic* were able to correctly distinguish between the two models. The Copas test produced results which were inferior to the Deviance (-2LL) statistic but were less erratic than the Hosmer Lemeshow test. Several methods are available to calculate the Deviance statistic which affords it little benefit over the HL test which can also be calculated several different ways. The Copas test may therefore be a potential viable alternative or addition to the HL test in the context of trauma scoring modelling.

## Appendix 1

### Program for outputting the Pearson chi-square test using the *proc genmod* program.

Only the first 12 cases of the data set have been included. The program shown will output the Pearson chi-squared test, scaled Pearson chi square test, the deviance test and the scaled deviance test for all data sets from the first 50 cases up to the first 1000 cases inclusive (%do i=50 %to 1000). To output all data cases from 6050 to 7000 the %do statement should be changed to %do i=6050 %to 7000.

```
data usc;
  input ds hciss rts;
datalines;
.00 26.00 11.00
1.00 5.00 8.00
1.00 13.00 12.00
.00 25.00 4.00
1.00 16.00 12.00
1.00 10.00 12.00
.00 43.00 11.00
1.00 9.00 12.00
1.00 9.00 12.00
1.00 10.00 12.00
.00 17.00 .00
1.00 9.00 12.00
;
run;

%macro repeat;

  %do i=50 %to 1000;

ods trace on ;
ods listing close ;
ods output Modelfit = pearson ;
proc genmod data = usc (FIRSTOBS=1 OBS=&i) DESC;
model ds = hciss rts / D=B;
run;
ods listing ;
proc print data=pearson;
run ;

  %end;
%mend;
%repeat
```

The free text from the SAS output had to be removed manually before an active data set was created. The active data set was then saved as an *excel* file and subsequently imported into SPSS. Scatter plots were then produced using the SPSS graphs option.

## Appendix 2

### Program for outputting the Pearson chi-square test using the *proc logistic* program.

Only the first 12 cases of the data set have been included. The program shown will output the Pearson chi-squared test and the deviance test for all data sets from the first 50 cases up to the first 1000 cases inclusive (`%do i=50 %to 1000`). To output all data cases from 6050 to 7000 the `%do` statement should be changed to `%do i=6050 %to 7000`.

```
data usc;
  input ds hciss rts;
datalines;
.00 26.00 11.00
1.00 5.00 8.00
1.00 13.00 12.00
.00 25.00 4.00
1.00 16.00 12.00
1.00 10.00 12.00
.00 43.00 11.00
1.00 9.00 12.00
1.00 9.00 12.00
1.00 10.00 12.00
.00 17.00 .00
1.00 9.00 12.00
;
run;

%macro repeat;

  %do i=50 %to 1000;
ods trace on ;
ods listing close ;
ods output GoodnessOfFit = pearson ;
proc logistic desc data = usc (FIRSTOBS=1 OBS=&i) ;
model ds = hciss rts / aggregate scale=none;
```

```
run;
ods listing ;
proc print data=pearson;
run ;
```

```
    %end;
```

```
  %mend;
```

```
%repeat
```

The free text from the SAS output had to be removed manually before an active data set was created. The active data set was then saved as an *excel* file and subsequently imported into SPSS. Scatter plots were then produced using the SPSS graphs option.

### Appendix 3

#### Program for outputting the Hosmer Lemeshow statistic using the *proc logistic* program.

The *lackfit* option is used after the model selection. The results is outputted using the *LackFitchiSq=lackfit* option in *ods*.

```
%macro repeat;
```

```
  %do i=50 %to 51;
```

```
ods trace on ;
ods listing close ;
ods output LackFitchiSq = lackfit ;
proc logistic desc data = usc (FIRSTOBS=1 OBS=&i) ;
model ds = hciss rts / lackfit;
run;
ods listing ;
proc print data=lackfit;
run ;
```

```
  %end;
```

```
%mend;
```

```
%repeat
```

The free text from the SAS output had to be removed manually before an active data set was created. The active data set was then

saved as an *excel* file and subsequently imported into SPSS. Scatter plots were then produced using the SPSS graphs option.

## APPENDIX 4

**Programs A and B for outputting the Copas test, the unweighted Standardized Residuals test and the Brier goodness of fit test.**

Only the first 12 cases of the data set have been included. The program shown will output the three goodness of fit statistics for all data sets from the first 50 cases up to the first 1000 cases inclusive (`%do i=50 %to 1000`). Only the first 50 output data sets *sum* 50 to *sum* 100 have been included in the program. To output all 951 results additional data sets from *sum* 101 to 1000 have to be included in the program. The output calculated is the mean goodness of fit value (sum /data set size). The actual goodness of fit test (i.e. mean value x data set size) is outputted by using program B. To output values for data set values from 6050 to 7000 the `%do` statement is again changed to:- `%do i=6050 %to 7000`. The *proc logistic* rather than *proc genmod* was used in this section.

### Program A

```
data usc;
  input ds hciss rts num;
datalines;
.00 26.00 11.00 1.00
1.00 5.00 8.00 2.00
1.00 13.00 12.00 3.00
.00 25.00 4.00 4.00
1.00 16.00 12.00 5.00
1.00 10.00 12.00 6.00
.00 43.00 11.00 7.00
1.00 9.00 12.00 8.00
1.00 9.00 12.00 9.00
1.00 10.00 12.00 10.00
```

```

.00 17.00 .00 11.00
1.00 9.00 12.00 12.00
;
run;

%macro repeat;
  %do i=50 %to 1000;

*close listing destination. i.e. output window;
ods listing close;
proc logistic data=usc.(FIRSTOBS=1 OBS=&i) des;
  model ds = iss rts / selection = backward;
  output out=lout&i p=pred;
run;

data work.lout&i;
  set lout&i;
  copas = (ds-pred)**2;
  ssr1 = ds-pred;
  ssr2 = ABS(ssr1);
  ssr3 = pred*(1-pred);
  ssr4 = SQRT(ssr3);
  ssr = ssr2/ssr4;
  brier1 = (pred-ds)**2;
  brier = (2/1000)*brier1;
run;

*re-open listing destination;
ods listing;
proc means data=work.lout&i;
var royston copas ssr brier;
output out=sum&i;
  run;
  %end;
%mend;

options symbolgen mprint mlogic;

%repeat

data final;
set sum50 sum51 sum52 sum53 sum54 sum55 sum56 sum57 sum58
sum59 sum60 sum61 sum62 sum63 sum64 sum65 sum66 sum67 sum68
sum69 sum70 sum71 sum72 sum73 sum74 sum75 sum76 sum77 sum78
sum79 sum80 sum81 sum82 sum83 sum84 sum85 sum86 sum87 sum88
sum89 sum90 sum91 sum92 sum93 sum94 sum95 sum96 sum97 sum98
sum99 sum100
;
run;

```

The data set *final* was exported and saved as an *excel file*. The data was then copied and pasted below the *datalines* statement (line of x's) into program B (page 190). This program strips out the

unnecessary statistics and calculates the actual goodness of fit value by multiplying its mean value by the size of its corresponding data set.

## Program B

```
data final;
  input freq stat $ royl copasl ssrl pearl brierl;
  if stat = 'N' then gof=1;
  if stat = 'MAX' then gof=2;
  if stat = 'MIN' then gof=3;
  if stat = 'MEAN' then gof=4;
  if stat = 'STD' then gof=5;
  if gof ne 4 then delete;
  datalines;
XXXXXXXXXX
;
run;
  data work.final;
  set final;
  royston = royl*freq;
  copas = copasl*freq;
  ssr = ssrl*freq;
  pearson = pearl*freq;
  brier = brierl*freq;
run;
```

**CHAPTER 8**  
**A STUDY TO EVALUATE**  
**MODEL VALIDATION USING DATA**  
**SPLITTING.**

**CONTENTS**

<b>Part 1.</b>	<b>Page Number</b>
<b>Section 1: Introduction</b>	<b>61</b>
<b>Section 2: Aims</b>	<b>62</b>
<b>Section 3: Methodology</b>	<b>63</b>
<b>Section 4: Results</b>	<b>66</b>
<b>Section 5: Discussion</b>	<b>84</b>
 <b>Part 2.</b>	
<b>Section 1: Introduction</b>	<b>61</b>
<b>Section 2: Aims</b>	<b>62</b>
<b>Section 3: Methodology</b>	<b>63</b>
<b>Section 4: Results</b>	<b>66</b>
<b>Section 5: Discussion</b>	<b>84</b>
 <b>Appendix (1-5):</b>	 <b>205</b>

## **Section 1: Introduction**

Data splitting is a well established means of validating a prognostic model. The simplest method is to perform a single split. The data set is divided into the training data set used for model development and the test data set used for model validation. The method by which the data set is split is an important issue. Several methods have been advocated. The simplest method is to split the data set into two equal halves by case number. Champion et al (1990b) in the development of ASCOT used an alternating case number split. Alternate survivors and non-survivors were placed into the training and test data set. An alternative method of data splitting is to perform a chronological split of the data accepting that the training and test data sets may not be of equal size. This can cause problems with interpretation when the Hosmer Lemeshow statistic is used. Data splitting is simple and easy to perform but does have several disadvantages which have been previously highlighted by Harrell et al (1996), Altman et al (2000) and Harrell (2002). Firstly, data splitting results in a significant reduction in the size of the data set used to develop the model. Secondly, data splitting does not validate the final model if the full data set is used. Thirdly, if the data set is homogenous then data splitting may not be a significantly stringent test to validate the model. Equally if the data set is not homogenous then different results may be obtained depending upon the method or point at which the data is split. Despite these drawbacks data splitting has become an integral part of model validation in the development of trauma scoring models. Multiple data splitting is in essence cross-validation and will be discussed in detail in the next chapter.

## **Part 1.**

### **Section 2: Aims of the study**

The aim of the study was to determine the variability in the prediction error using data splitting as a method of model validation.

### **Section 3: Methodology**

#### **Section 3.1 Statistical software.**

SAS version 8.0 was used to perform the logistic regression modelling and the simulation studies. SPSS was used to plot the graphs.

#### **Section 3.2 Method for Model Selection**

Backward LR test (likelihood ratio).

(The calculated probabilities were for survival rather than death).

#### **Section 3.3 Models Used**

Two models were used in this study: -

1. Dependent variable:- death/survival.

Predictor variable:- HCISS

2. Dependent variable:- death/survival.

Predictor variables:- HCISS + RTS (unweighted)

These two models have been evaluated in chapter 4 with regard to calibration, discrimination and log odd plots.

### **Section 3.4 Method of Data Splitting**

The training data set was selected to be the first 1000 cases from the revised USC data set. Six test data sets were used in the study. The first test data set was chosen to be the second 1000 cases. The second test data set was chosen as the third 1000 cases. The third test data set was chosen as the fourth 1000 cases. The fourth test data set was chosen as the fifth 1000 cases. The fifth test data set was chosen as the sixth 1000 cases. The sixth test data set was chosen as the seventh 1000 cases. The remaining cases in the revised USC data set were not used in the study. Multiple 50% data splits were therefore generated using this method.

### **Section 3.5**

The coefficients for both models were obtained from the training data set (first 1000 cases) using the SAS *proc logistic* program (appendix 1). The coefficients from the training data set were then used to calculate the predicted probabilities on each of the six test data sets using a second program (appendix 2).

### **Section 3.6**

The test statistic used in this study was the mean square error.

$$\text{Mean Square Error} = \sum(DS - \text{pred prob})^2 / n$$

Where DS is the dependent variable:- death/survival.

Pred pob: - is the predicted probability of survival.

n = the size of the data set.

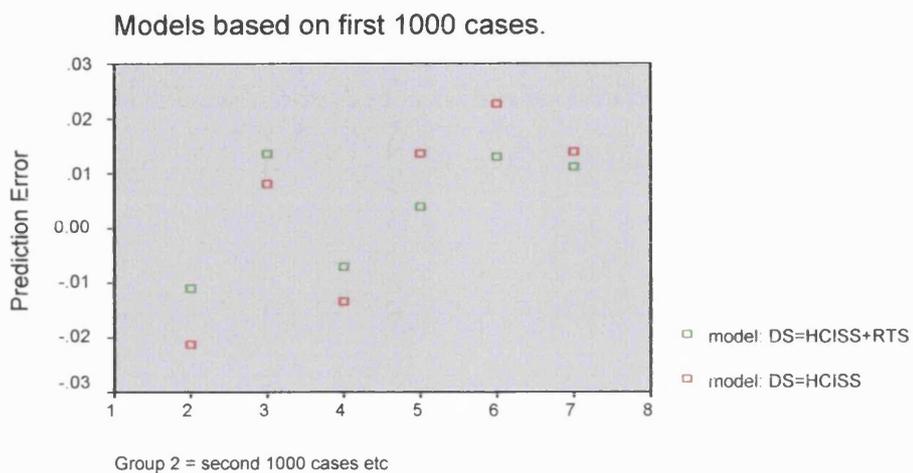
This statistic is also known as the residual squared error.

The mean square error for both models was calculated on the training data set (first 1000 cases). The mean square error was then calculated on the six test data sets. The difference between the mean square error for the training data set and each test data set (the prediction error\*) was calculated for both models. A negative result indicated that the training data set was over-fitted compared to the test data set.

\* Prediction error = MSE (training data set) – MSE (test data set)

## Section 4: Results

### Prediction Error Plotted Against Sequential Data Splitting



Graph 1

Table 1 shows the variability in the prediction error with repeated sequential data splitting. The graph also shows that for groups 2 and 4 the prediction error has a negative value indicating that the model fit for the training data set was better than that for the test data set.

## **Part 2**

### **Section 2: Aim**

The aim of the study was to look at the variability of the data splitting method using randomly selected data sets and three different size data splits.

### **Section 3: Methodology**

#### **Section 3.1 Statistical software.**

SAS version 8.0 was used to perform the logistic regression modelling and the simulation studies. SPSS was used to plot the graphs.

#### **Section 3.2 Method for Model Selection**

Backward LR test (likelihood ratio).

(The calculated probabilities were for survival rather than death).

#### **Section 3.3 Model Used**

Dependent variable:- death/survival

Predictor variable:- HCISS

**Section 3.4** Method for determining goodness of fit (using a measure for predictive accuracy) for the test and training data sets was the Mean Square Error (MSE).

$$MSE = \sum (dl - \text{pred prob})^2 / n$$

Where dl is the dependent variable:- death/survival

Pred prob: -is the predicted probability of survival

n = the size of the data set.

### **Section 3.5 Data Splitting**

Three different sized data splits were used: -

1. 50% data split.
2. 70% data split i.e. training data set = 70% test data set = 30%.
3. 90% data split i.e. training data set = 90% test data set = 10%.

### **Section 3.6 Data Sets.**

100 data sets containing 1000 cases were randomly selected from the revised USC data set. The program used to generate 100 data sets is given below (appendix 3). The SAS program in appendix 3 is the same program used in the cross validation chapter (chapter 9). The program randomly selects 1000 cases without replacement from the full data set. An individual case can only be selected once. For a more detailed account of this program see chapter 9. The 100 data sets were stored in the temporary SAS library as samples 1-100.

The next step was to generate the coefficients (constant and predictor variable) for the training data using the *proc logistic* program. This program was run 100 times in order to generate coefficients for the 100 data sets. With each run the appropriate data set name was written into the program (i.e. *sample1-sample100*). The appropriate sized training data set was generated from the *sample* data set using the (*firstobs=1 obs=700*) command within the *proc logistic* program (see appendix 4).

Prior to calculating the MSE for the test and training data sets an additional program (appendix 5) was run which added a number variable to each data set (sample 1-100). The number variable

produced a sequential series of numbers from 1–1000 corresponding to the 1000 cases in each data set. This additional variable enabled a program to be written which would generate the appropriate split in the data set and thus generate the test data set.

The final step was to calculate the MSE for the training and test data sets. This was performed using the program given in appendix 6. The MSE for the training data set was calculated firstly by outputting the predicted probabilities from *proc logistic* using the *output out=lout* statement. Following this a program was written to calculate the squared error for each data set.

i.e. Squared Error =  $\sum(\text{dl} - \text{pred prob})^2$

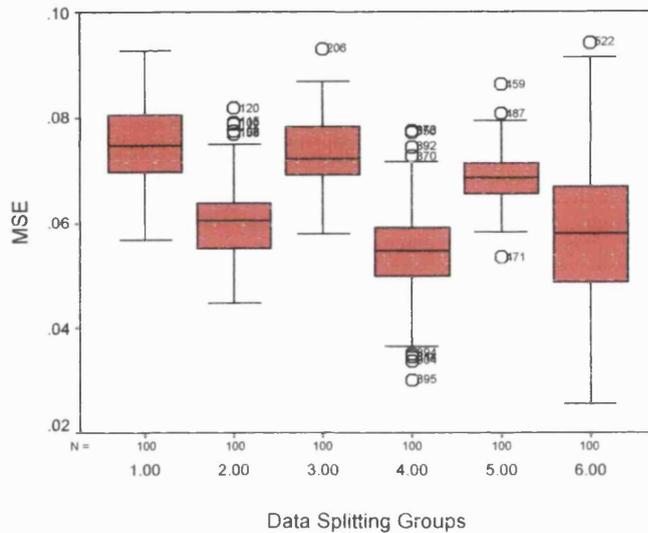
The MSE for the training data set was then calculated by using the *proc means* program which calculates the mean for the output variable SE.

#### **MSE for the test data set.**

The coefficients for the constant and the predictor variable for each sample data set had to be inserted manually for each run. To calculate the MSE for the test data set a program was written which selects appropriate cases from the sample data set and deletes the unwanted cases. (using an *if ... then delete* statement). The coefficients were used to calculate the predicted probabilities using the standard logistic regression equation. The squared error and mean squared error were calculated using the method described above for the training data set. The program was run 100 times and the MSE values stored in data sets called *sumtest* and *sumtrain*.

## Section 4: Results

### Box Plots For MSE For Three Data Splits



**Graph 2**

Group 1 = 100 Training data sets; 50% split. (first 500 cases).

Group 2 = 100 Test data sets; 50% split. (last 500 cases).

Group 3 = 100 Training data sets; 70% split. (first 700 cases).

Group 4 = 100 Test data sets; 30% split. (last 300 cases).

Group 5 = 100 Training data sets; 90% split. (last 900 cases).

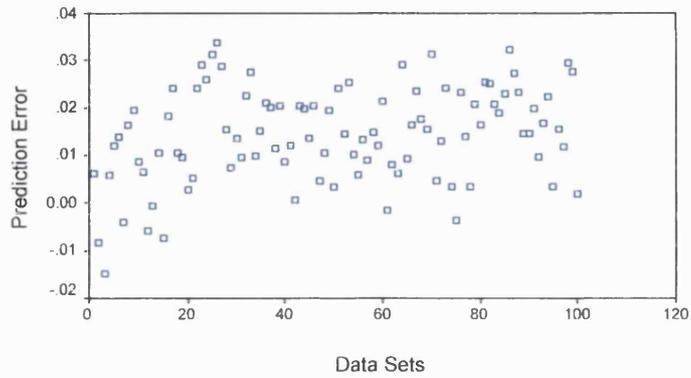
Group 6 = 100 Test data sets; 10% split. (last 100 cases).

The MSE for the full revised USC data set = 0.069.

All six groups had a MSE value less than the MSE for the full revised USC data set. The mean MSE was less than the mean MSE for the corresponding training data sets for all three data splits. The smallest variability for the training data sets was seen for the 90% split group. The largest variability for the test data sets was seen for the 10% split group.

Scatter Plot of Prediction Error Plotted  
Against 100 Data Sets

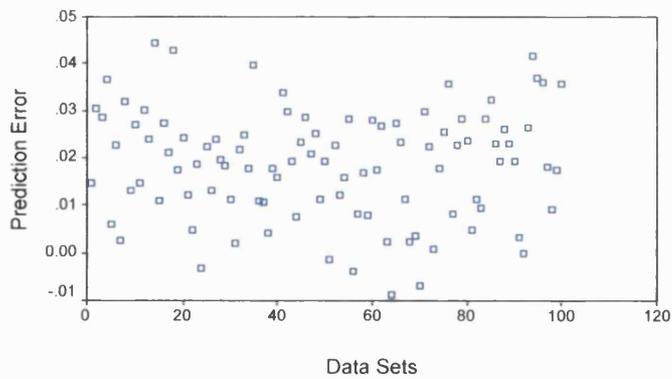
50% Data Split



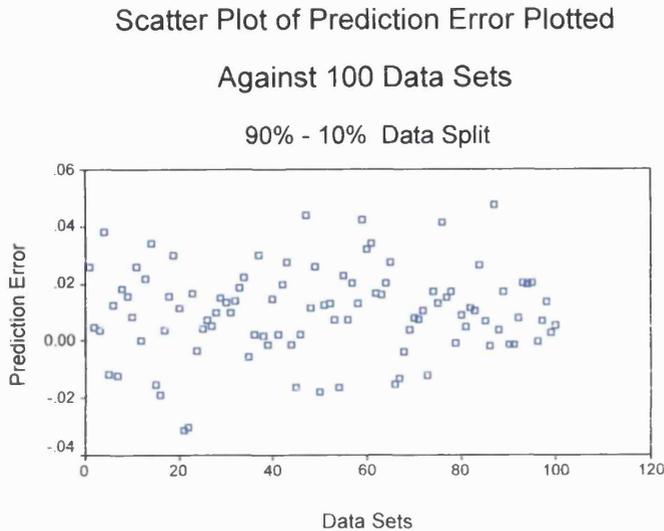
**Graph 3**

Scatter Plot of Prediction Error Plotted  
Against 100 Data Sets

70% - 30% Data Split



**Graph 4**



### Graph 5

Graphs 3-5 show that not all of the 100 data sets had a positive prediction error.

Graph 3 (50% split): The Prediction Error had a negative value in 92% of cases. There were no ties.

Graph 4 (70% split): The Prediction Error had a negative value in 95% of cases. There were no ties.

Graph 5 (90% split): The Prediction Error had a negative value in 78% of cases. There were no ties.

When the Prediction Error has a negative value the MSE for the training data set is smaller (better fit) than its corresponding value for the test data set. Thus indicating that the data splitting method has not corrected for over prediction of the internally validated model.

## Section 5: Discussion

The two models used in this study have been previously evaluated in Chapter 4. The results from chapter 4 showed that the calibration and discrimination values for the HCISS + RTS model was superior to the HCISS model. These two models are therefore representative of a poor model (HCISS) and a superior model (HCISS + RTS). The results from part 1 of this chapter showed that the variability in the prediction error was similar for both models. In two groups the training group MSE was smaller than the test MSE value. The results from part 2 of the study also demonstrated that not all cases had training values greater than their corresponding test values. This effect was seen for all three data split sizes. Both studies demonstrate therefore that data splitting can be an ineffective method to correct for over-fitting in an internally validated model. The methodology used in part 2 of the study was based upon the method used by Steyerberg et al (2001). In their simulation study they used the GUSTO-I study data which recorded mortality in patients with acute myocardial infarction. These authors found that data splitting using 50%-50% and 70%-30% splits resulted in overly pessimistic estimates of the test statistic. They also compared cross-validation and bootstrapping to data splitting. They found that bootstrapping gave the most accurate measure of model performance. Despite its limitations data splitting remains an important method of validating a logistic model. Pickard et al (1990) in a monograph on the theory of data splitting highlighted two problems associated with making an inappropriate data split. Firstly, if the training data set is too small the ability to predict future observations is impaired. Secondly, the ability to assess the fitted model is impaired if the

test data set is too small. The best way of splitting the data still remains an unresolved problem. The optimal data split may have to be tailored around the data set, a less than ideal situation. A 50% split of the data is probably the most balanced and is the method used by many investigators.

## Appendix 1

SAS program for model development using the training data set.

```
proc logistic data=sasuser.Hcfinal2 (firstobs=1 obs=1000) DES;  
model ds = hciiss / selection = backward;  
run;
```

## Appendix 2

SAS program for calculating predicted probabilities (for survival)  
and also the mean squared error.

Model: dependent variable = D/S (death/survival).

Predictor variable = HCISS.

```
data Pred;  
input hciiss ds;  
res1 = (hciiss*0.830)-7.326;      LINE A  
res2 = -(res1);  
res3 = EXP(res2);  
result = 1/(1+res3);  
PE = (ds-result)**2;  
cards;  
1.00 .00  
9.00 1.00  
12.00 1.00  
16.00 .00  
1.00 1.00  
1.00 1.00  
25.00 .00  
9.00 1.00  
12.00 1.00  
16.00 1.00  
12.00 1.00  
10.00 1.00  
1.00 1.00  
2.00 1.00  
10.00 1.00  
37.00 .00  
1.00 1.00  
1.00 1.00  
12.00 1.00  
11.00 1.00  
75.00 .00  
;  
proc univariate data=pred;  
var PE;  
run;
```

The program line named Line A (page 205) shows the derived value for the HCISS coefficient (+0.830) and also the derived value for the constant (-7.326).

*Proc univariate* produces the mean value for  $(ds\text{-result})^{**2}$  for all the cases in the test data set  $(\sum(ds\text{-result})^{**2}/n)$ , which is the Mean Squared Error.

### Appendix 3

Program for random sampling without replacement. This program generates random numbers using the *ranuni* random numbers generator. The program needs a starting value or seed, a negative value for the *ranuni* generator is used in this program and this tells the computer to use the time value from its internal clock. A different sequence of numbers and hence data sets are outputted (named:- sample) each time the program is run that day. A SAS *macro repeat* program was also added so that multiple data sets could be generated. The set statement in the program defines the name of the original data set, in this case named *test*. The final 10 cases of the revised USC data set are given for brevity. For further information regarding this program see chapter 9, page 214.

```
data test;
  input ds iss rts num;
datalines;
1.00 34.00 12.00 7060
1.00 13.00 12.00 7061
.00 17.00 11.00 7062
1.00 5.00 12.00 7063
.00 26.00 8.00 7064
1.00 14.00 11.00 7065
1.00 5.00 12.00 7066
.00 26.00 4.00 7067
1.00 14.00 11.00 7068
1.00 4.00 12.00 7069
;
run;
```

```

%macro repeat (n=100);
%do i=1 %to &n;
data sample&i (drop=sampsize obsleft) ;
    sampsize = 1000 ;
    obsleft = totobs ;
    do while (sampsize > 0) ;
        readit + 1 ;
        if ranuni(0) < sampsize/obsleft then do ;
            set test point=readit nobst=totobs ;
            output ;
            sampsize = sampsize - 1 ;
            end ;
            obsleft = obsleft - 1 ;
        end ;
    stop ;
run ;

    %end;
%mend;
%repeat

```

## Appendix 4

Program to generate the coefficients for the 100 data sets (named sample 1- 100). Only the first three *proc logistic* programs are given for data sets 1-3 (work.sample1, work.sample2, work.sample3).

```

proc logistic data=work.sample1 (firstobs=1 obs=700) des;
    model ds = iss / selection = backward;
    output out=lout501 p=pred;
run;
proc logistic data=work.sample2 (firstobs=1 obs=700)des;
    model ds = iss / selection = backward;
    output out=lout502 p=pred;
run;
proc logistic data=work.sample3 (firstobs=1 obs=700)des;
    model ds = iss / selection = backward;
    output out=lout503 p=pred;
run;

```

## Appendix 5

Program which adds a numbers variable to each data set. Each case in the data set is now renamed sample1r through to sample100r and has a corresponding number. The numbers are sequential 1-1000 and therefore correspond to the order in which the cases appear

within the data sets. The program has to be run 100 times. Only the first 3 programs are given for brevity.

```
data sample1r;
merge work.sample1 sasuser.number;
run;
data sample2r;
merge work.sample2 sasuser.number;
run;
data sample3r;
merge work.sample3 sasuser.number;
run;
```

## Appendix 6

Program for calculating MSE for the training and test data sets.

Programs for data sets 99 and 100 are given for brevity.

```
ods listing close;
proc logistic data=work.sample99r (firstobs=1 obs=700)des;
  model ds = iss / selection = backward;
  output out=lout2599 p=pred;
run;

data work.lout2599;
  set lout2599;
  mse = (ds-pred)**2;
run;

*re-open listing destination;
ods listing;
proc means data=work.lout2599;
var mse;
output out=sum599;
run;

data sample99r ;
  set sample99r;
  if digits < 701 then delete;
  res1 = (iss*-0.1654)+4.9526;
  res2 = -(res1);
  res3 = EXP(res2);
  result = 1/(1+res3);
  msetest = (ds-result)**2;

proc means data=sample99r;
var msetest;
output out=sum1599;
run;

options symbolgen mprint mlogic;
```

```

ods listing close;
proc logistic data=work.sample100r (firstobs=1 obs=700)des;
  model ds = iss / selection = backward;
  output out=lout2600 p=pred;
run;

data work.lout2600;
  set lout2600;
  mse = (ds-pred)**2;
run;

*re-open listing destination;
ods listing;
proc means data=work.lout2600;
  var mse;
  output out=sum600;
run;

data sample100r;
  set sample100r;
  if digits < 701 then delete;
  res1 = (iss*-0.1575)+4.3785;
  res2 = -(res1);
  res3 = EXP(res2);
  result = 1/(1+res3);
  msetest = (ds-result)**2;

proc means data=sample100r;
  var msetest;
  output out=sum1600;
run;
  options symbolgen mprint mlogic;

data sumtrain;
  set sum599 sum600;
run;

data sumtest;
  set sum1599 sum1600;
run;

```

# **CHAPTER 9**

## **A STUDY TO EVALUATE CROSS VALIDATION ON TRAUMA SCORING MODELLING**

### **CONTENTS**

	<b>Page Number</b>
<b>Section 1: Introduction</b>	<b>211</b>
<b>Section 2: Aims</b>	<b>212</b>
<b>Section 3: Methodology</b>	<b>213</b>
<b>Section 4: Results</b>	<b>218</b>
<b>Section 5: Discussion</b>	<b>235</b>
<b>Appendix 1:</b>	<b>240</b>
<b>Appendix 2:</b>	<b>242</b>

## **Section 1: Introduction**

Several methods are available which attempt to overcome the problem of over-fitting which often occurs when the model is validated against the data set from which it was developed. Data splitting is arguably the simplest method, although the optimal way of splitting the data set is unresolved (Harrell, 1996, 2002). Splitting the data into two halves at its median point may be appropriate in some data sets but not in others. Data splitting can result in loss of potentially important data, especially when the original data set is small or certain covariate patterns are sparse. Cross-validation is in essence repeated data splitting. A number of reduced data sets are developed which are randomly sampled from the full data set. The sampling is performed such that individual cases can only be selected once i.e. random sampling without replacement. This is in contrast to the bootstrap method which is random sampling with replacement. The bootstrap data set being the same size as the original data set. One of the problems with cross-validation is that there are no clear evidenced based guidelines regarding the appropriate size of the validation data set. For small data sets of 100 or less then the Jackknife (Tukey, 1958) method of cross-validation may be appropriate. This involves leaving out one case and validating the model on  $n-1$  cases. The procedure is repeated for all  $n-1$  cases. For large data sets however leaving only one case out results in a validation set which is too similar to the training set. The method is therefore not stringent enough to effectively validate the model. The size of the validation sample in large data sets is therefore of importance. The effect of varying the sampling size on the validation result will be evaluated in this chapter in the context of trauma scoring modelling.

## **Section 2: Aims**

The aim of this study is to determine whether cross-validation is a reliable method of validating a logistic model using trauma scoring models.

Hypotheses to be tested.

1. That the sampling size will effect the validation test result.
2. That the size of the original data set will affect the validation result.
3. That the validation results will be affected by the model selected.
4. That the sampling method will affect the test result i.e. random sampling without replacement compared to random sampling with replacement.

## **Section 3: Methodology**

### **Section 3.1 Statistical software.**

SAS version 8.0 was used to perform the logistic regression modelling and the simulation studies. SPSS was used to plot the graphs.

**Section 3.2 Data Set.** The revised USC data set was used.

### **Section 3.3 Method for Model Selection**

Backward LR test (likelihood ratio).

(The calculated probabilities were for survival rather than death).

### **3.4 Goodness of fit tests**

The goodness of fit tests which were chosen to validate the model were:- (1) The Copas test, (2) The Brier score.

### **3.5 Models**

The two models chosen for the simulation studies were: -

Model 1. Dependent variable = Death /survival

Independent variable = HCISS

Model 2. Dependent variable = Death /survival

Independent variables = HCISS and RTS

### **Section 3.6: Study 1**

To determine the effect of increasing the size of the sampling data set on the cross validated Brier score.

### **Model**

Dependent variable = Death /survival

Predictor variable = HCISS

### **Random Sampling Without Replacement**

A program for *random sampling without replacement* was downloaded from the SAS website. The program performs random sampling without replacement using the *ranuni* random numbers generator. The program needs a starting value or seed, a negative value for the *ranuni* generator will tell the computer to use the time value from its internal clock. A different sequence of numbers and hence parameter estimates are produced each time the program is run that day. A SAS *macro repeat* program was also added so that 1000 cross validated test data sets could be generated. The program for random sampling without replacement works by selecting the first case using the random numbers seed. This first case is then excluded from the original data set. The program then randomly selects the second case from the remaining cases in the data set. These two selected cases are again excluded from the original data set and the program randomly selects the third case etc. The full complement of selected cases comprises the cross validated test data set. The full SAS program is given in appendix 1 at the end of this chapter.

The cross validation was first performed selecting 1000 cases from the first 1100 cases in the USC data set. 1000 cross validated data sets were generated and the Brier score was calculated for each data set. The cross validation method was then repeated by selecting 1000 cases from the first 1200 cases, 1000 cases from the first 1300 cases, 1000 cases from the first 1400 cases and lastly 1000 cases from 1500 cases. The Brier results were plotted as

histograms and the mean cross validated Brier score was calculated for each of the selected data set sizes. The Brier score for the model derived from the first 1000 cases (internally validated Brier score) was also calculated.

### **Section 3.6: Study 2.**

To determine the effect of increasing the size of the sampling data set on the cross validation Copas result. Exactly the same method was used as that in study 1 except that the Copas goodness of fit test was used instead of the Brier test.

### **Section 3.6: Study 3.**

The aim of this study was to determine whether any variability of the cross validated results was due to the sampling method or the increase in data set size. Exactly the same method was used as that in study 2 (Copas goodness of fit test) except that 1000 cases were randomly selected from additional data sets sizes. These were: the first 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 3000, 4000, 5000, 6000, and 7000 cases. A second set of results was generated for the same model and same data set sizes by running the program for a second time. This generated a second set of 1000 cross validated data sets. The purpose of this was to determine whether any variability present was due to the sampling method or the increase in data set size.

#### **Section 3.6: Study 4.**

The aim of this study was to determine the effect of changing the sampling method on the cross validated results. The same methodology and data set points were used as in study 3. The sampling method chosen was *random selection with replacement*. A program to perform this function was down loaded from the SAS website and a SAS *macro repeat* program was also added so that 1000 samples could be generated. The program was written so that it randomly selected 1000 cases from the specified data set size. This program also generates random numbers using the *ranuni* random numbers generator. The *ranuni* value in the program is positive so that the same sequence of numbers are generated every time the program is run. The *random sampling with replacement* method returns the first selected case back into the data set. The second random number selected is therefore selected from the entire data set. All the cases in the data set have the same probability of being selected. The full complement of selected cases comprises the cross validated test data set. The complete program is shown in appendix 2 at the end of this chapter.

#### **Section 3.6: Study 5.**

The aim of this study was to evaluate the effect of changing the sampling data set on the cross validated results. Exactly the same method was used as that in study 3 for the Copas goodness of fit test except that the second 1000 cases from the data set were used. A series of cross validation results were then obtained by sequentially increasing the size of the data set. The first Copas cross validated value being derived by random selection without replacement from case 1001 to case 2100. The second cross

validated Copas value being derived by random selection without replacement from case 1001 to case 2200 etc.

### **Section 3.6: Study 6.**

The aim of this study was determine the effect of increasing the data set size on the cross validated results using the model DS = HCISS + RTS. The same method was used as that in studies 2 and 3. The first 1000 cases were used to derive the Copas value by internal validation. Cross validated results were derived using data set sizes of: 1100, 1200, 1300, 1400, 1500, 1600, 1700 1800, 1900, 2000 and 2100.

### **Section 3.6: Study 7.**

The aim of this study was determine the effect of increasing the data set size on the cross validated results using the model; DS = HCISS, but using a much smaller sized data set. The same method was used as that in studies 2 and 3 except that the first 100 cases were used to derive the Copas value by internal validation. Additional values were derived using data set sizes of:- 110, 120, 130, 140, 150, 160, 170, 180, 190 and 200.

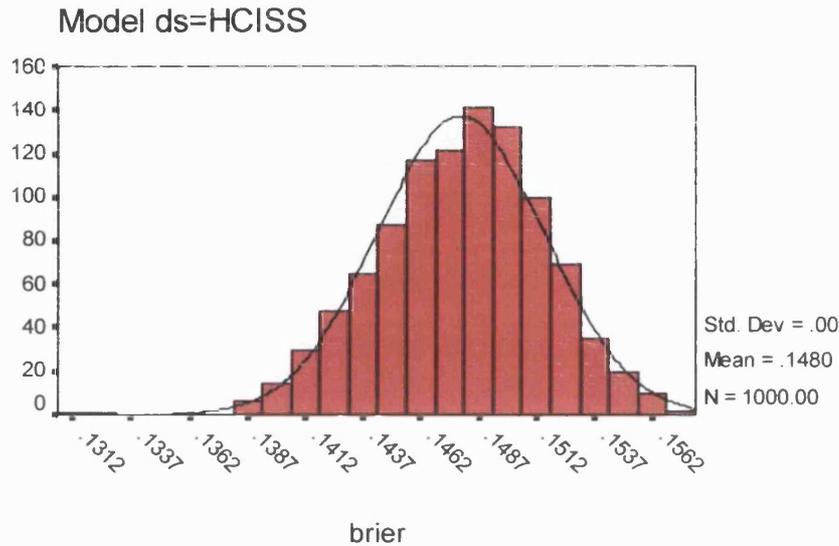
### **Section 3.6: Study 8.**

The aim of this study was determine the effect of increasing the data set size on cross validation using the model; DS = HCISS + RTS. The same method was used as that in study 7 i.e. the first 100 cases were used to derive the Copas value by internal validation. Additional values were derived using data set sizes of: 110, 120, 130, 140, 150, 160, 170, 180, 190 and 200.

## Section 4: Results

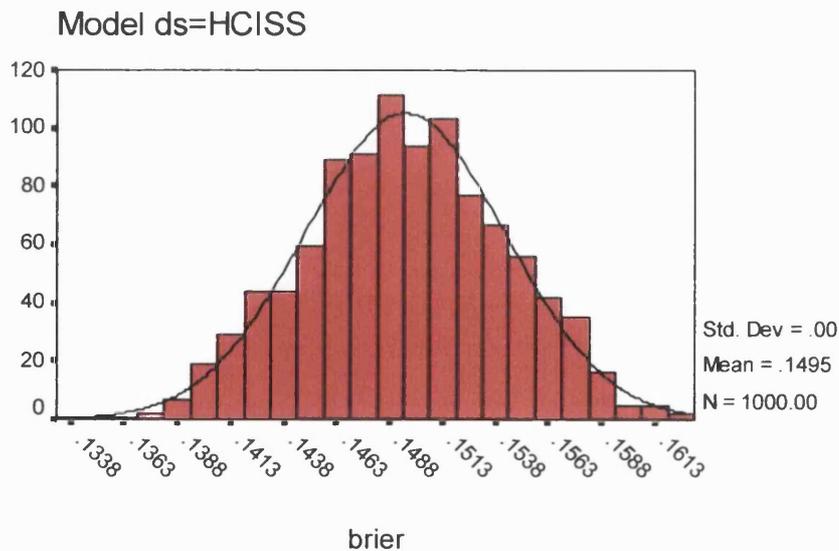
### Section 4.1 Study 1. Goodness of Fit Test: Brier

Frequency Distribution of 1000 Cases Randomly  
Selected From 1100 Cases Without Replacement



**Graph 4.1a**

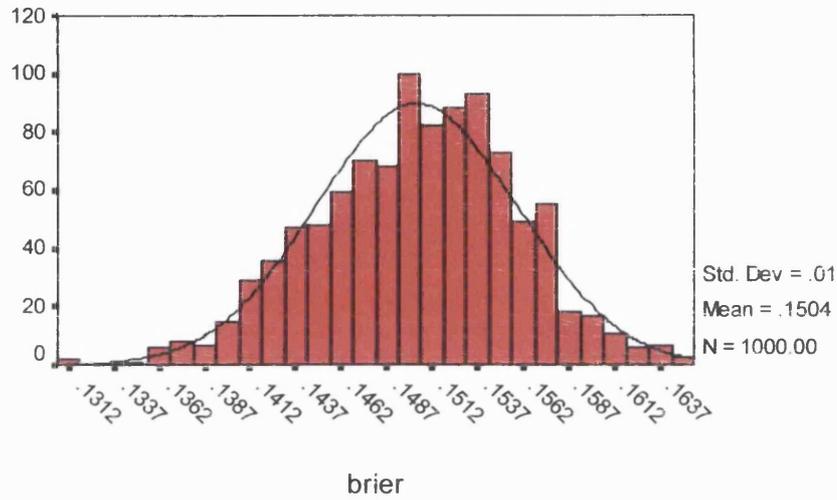
Frequency Distribution of 1000 Cases Randomly  
Selected Without Replacement From 1200 Cases



**Graph 4.1b**

Distribution of 1000 Randomly Selected Cases  
From 1300 Cases Without Replacement

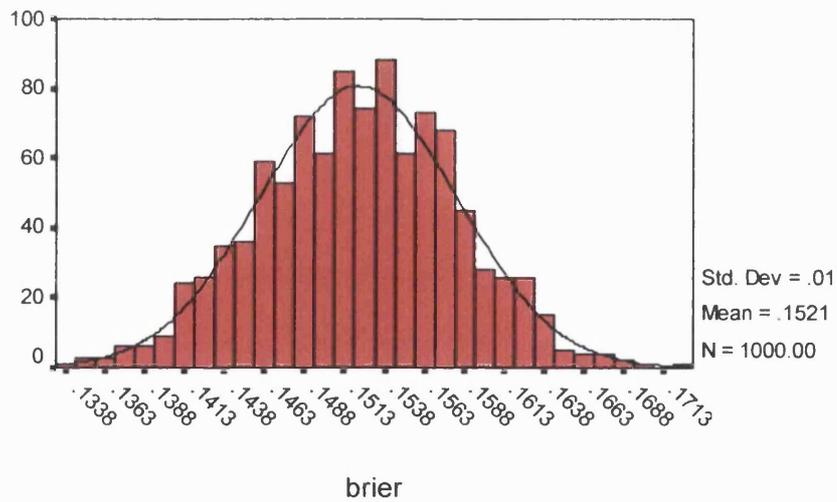
Model ds=HCISS



**Graph 4.1c**

Frequency Distribution of 1000 Cases Randomly  
Selected from 1400 Cases Without Replacement

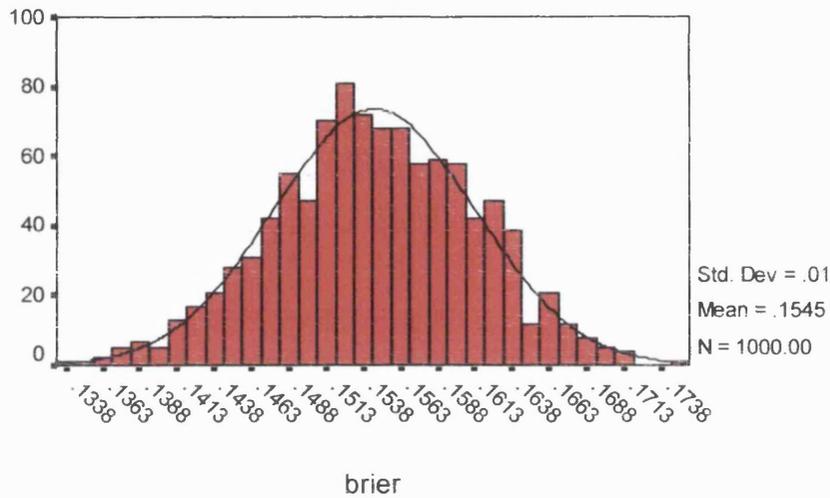
Model ds=HCISS



**Graph 4.1d**

Frequency Distribution of 1000 Cases Randomly  
Selected From 1500 Cases Without Replacement

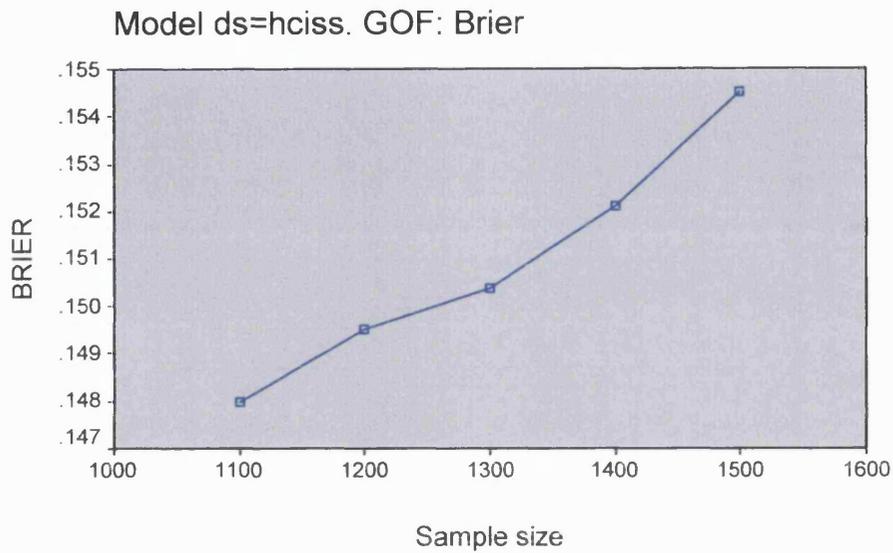
Model ds=HCISS



**Graph 4.1e**

Graphs 4.1a to 4.1e show that the cross validated Brier scores have a good approximation to the normal distribution. This indicates that the random sampling method is balanced.

Mean Cross Validated Result Plotted  
Against Sample Size.



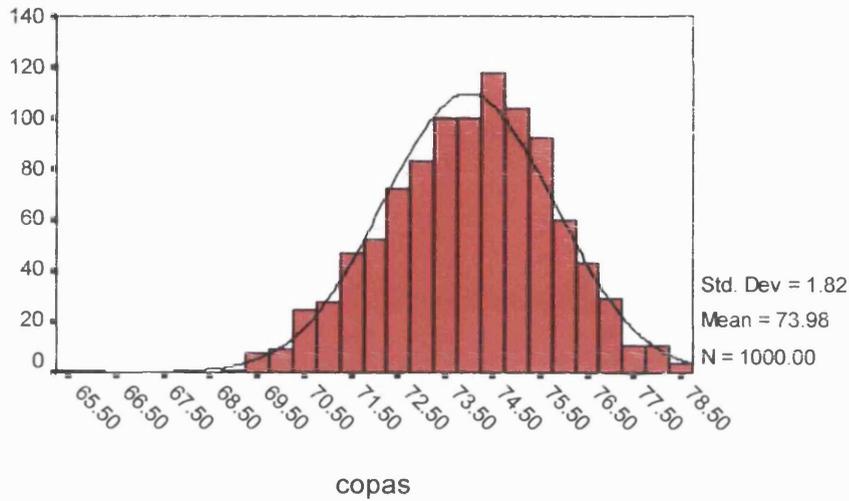
**Graph 4.1f**

Graph 4.1f shows that the cross validated Brier score increases in an approximate linear fashion as the size of the sampling data set increases. The Brier score for the model DS = HCISS derived from the first 1000 cases = 0.143. The results from graph 4.1f therefore shows that increasing the size of the sampling data set increases the cross validated Brier score in an approximately linear fashion.

## Section 4.2. Study 2. Goodness of fit test: Copas

Frequency Distribution of 1000 Cases Randomly Selected From 1100 Cases Without Replacement

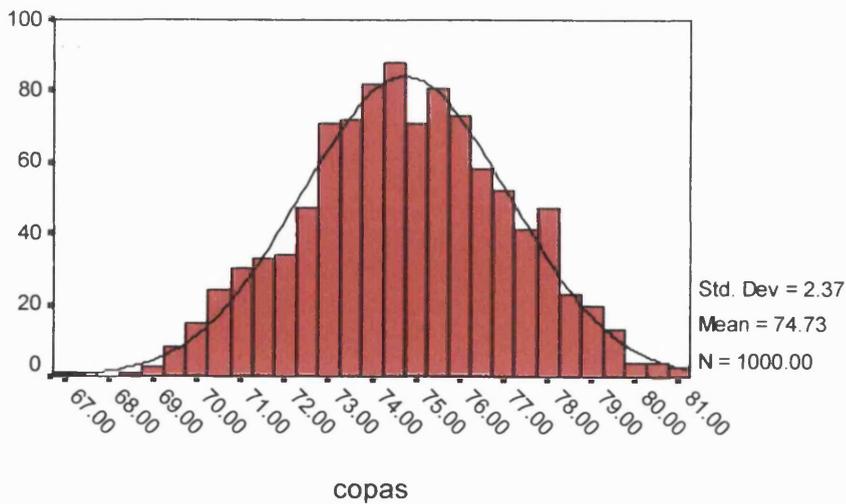
Model ds=HCISS



**Graph 4.2a**

Frequency Distribution of 1000 Cases Randomly Selected Without Replacement From 1200 Cases

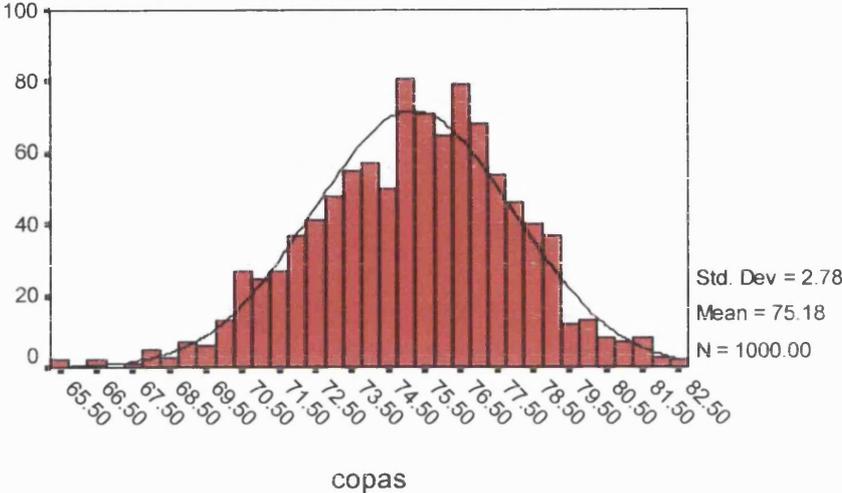
Model ds=HCISS



**Graph 4.2b**

Distribution of 1000 Randomly Selected Cases  
From 1300 Cases Without Replacement

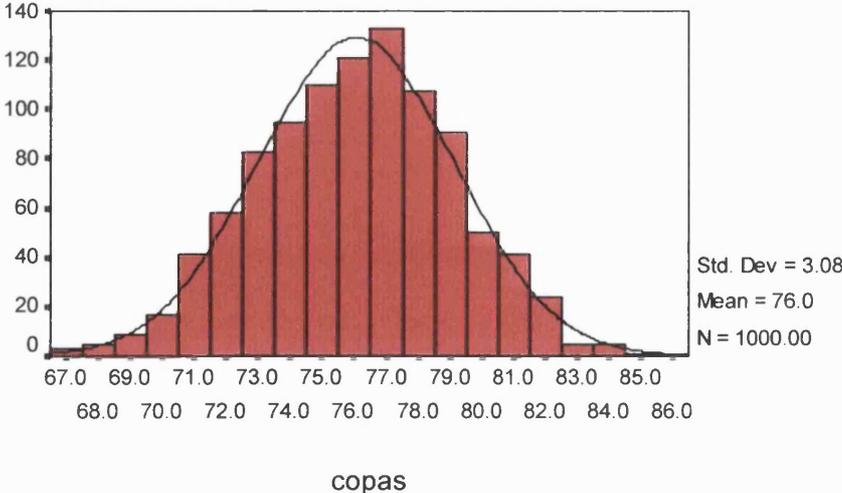
Model ds=HClSS



**Graph 4.2c**

Frequency Distribution of 1000 Cases Randomly  
Selected from 1400 Cases Without Replacement

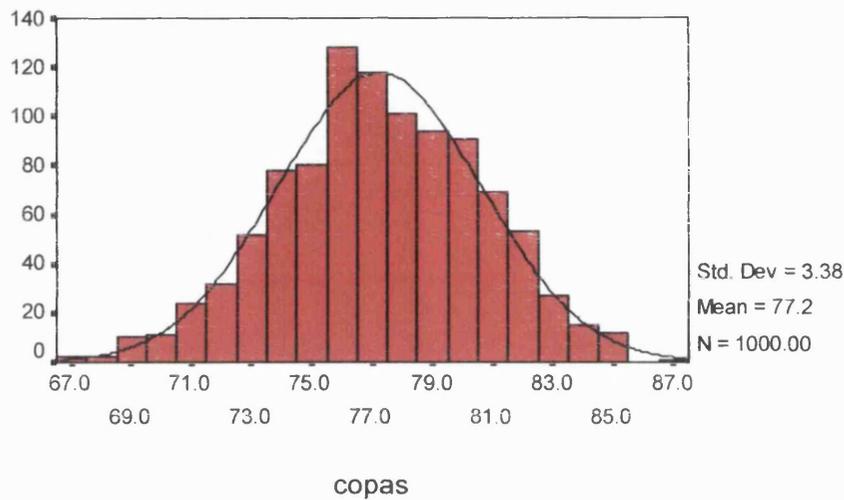
Model ds=HClSS



**Graph 4.2d**

Frequency Distribution of 1000 Cases Randomly  
Selected From 1500 Cases Without Replacement

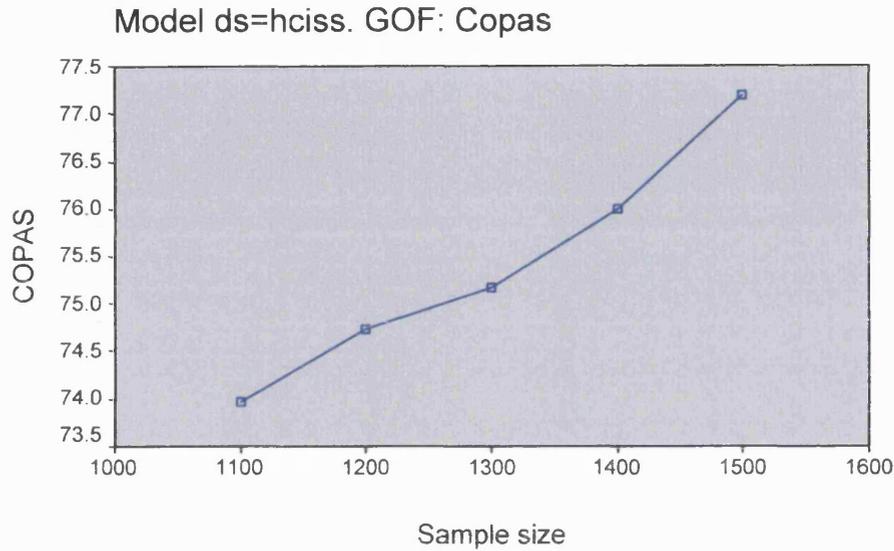
Model ds=HCISS



**Graph 4.2e**

Graphs 4.1a to 4.1e show that the cross validated Copas test values have a reasonable approximation to the normal distribution. This indicates that the random sampling method is balanced.

Mean Cross Validated Result Plotted  
Against Sample Size.



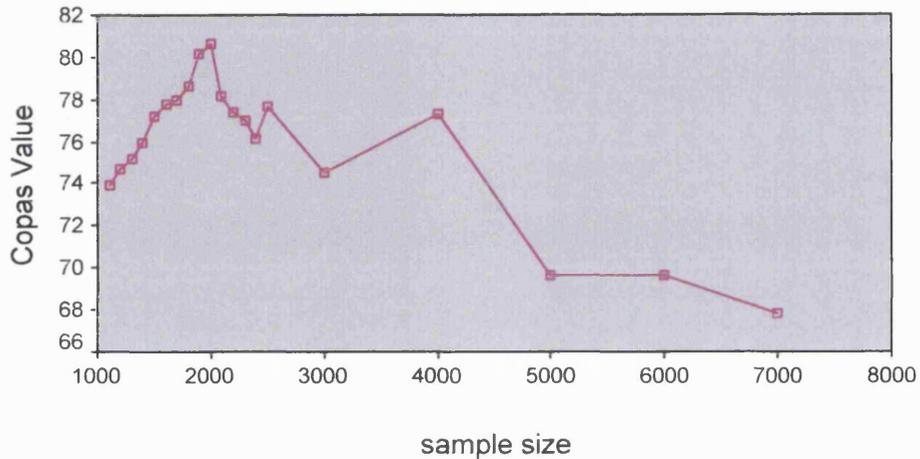
**Graph 4.2f**

Graph 4.2f shows that the cross validated Copas test increases in an approximate linear fashion as the size of the sampling data set increases. The Copas test for the model DS = HCISS derived from the first 1000 cases = 71.54. The results from graph 4.1f therefore shows that increasing the size of the sampling data set increases the cross validated Copas test in an approximately linear fashion.

### Results 4.3. Study 3. Goodness of fit test: Copas.

Cross Validated Copas Value Plotted Against Sequential Increase In Sample Size.

Model ds=hciss. Data set size = first 1000 cases



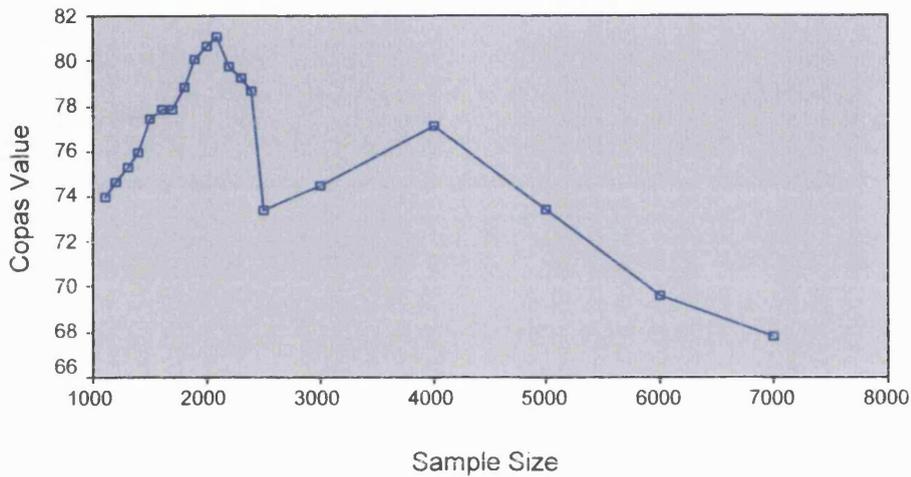
Copas Value for the model = 71.54

**Graph 4.3a**

Graph 4.3a shows that further enlargement of the sampling data set results in a gradual reduction in the cross validated Copas value. The final three data set points produce cross validated Copas values which are less than the internally validated Copas Value of 71.54.

## Cross Validated Copas Value Plotted Against Sequential Increase In Sample Size

Model ds=hciss Data Set Size = first 1000 cases



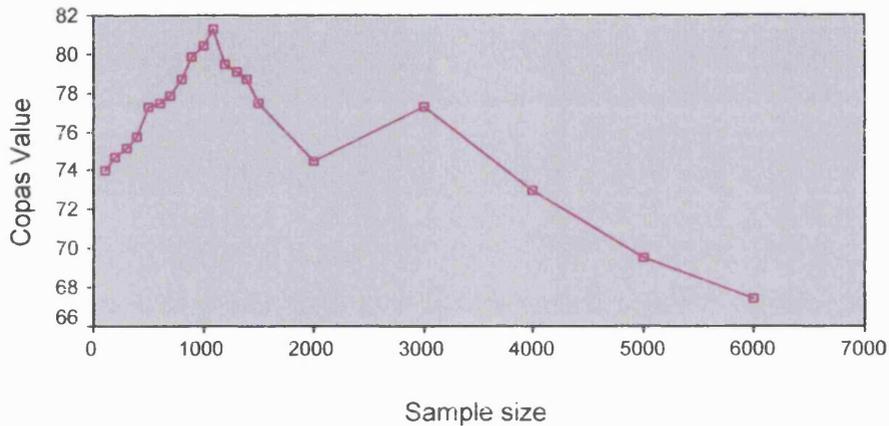
**Graph 4.3b**

Graph 4.3b shows a similar pattern of variation for the Copas value for the second run of bootstrap samples. This indicates that the variability in the Copas bootstrap values is predominantly due to the increase in the data set size rather than an intrinsic problem with the sampling (*ranuni*) program.

## Results 4.4. Study 4. Goodness of fit test: Copas.

Cross Validated Copas Value Plotted Against Sequential Increase In Sample Size.

Model ds=hciss. First 1000 Cases.



Sampling with replacement.

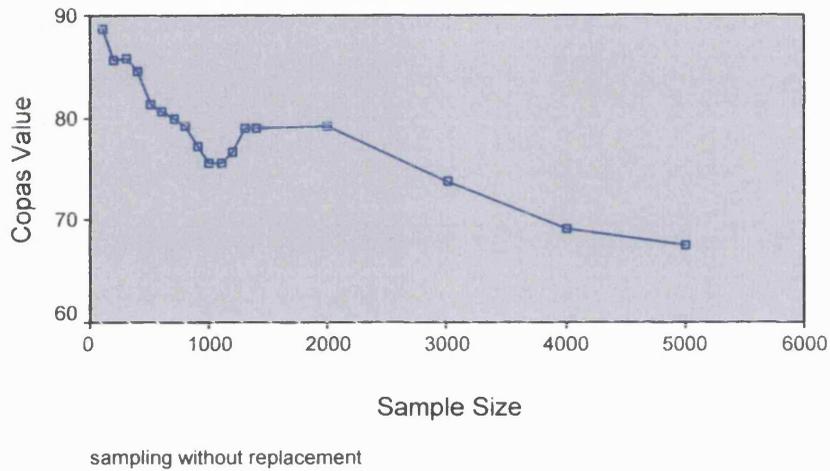
### Graph 4.4a

Graph 4.4a shows the cross validated Copas values derived by sampling with replacement follows a similar pattern to the Copas values derived by sampling without replacement. The final two data points produce cross validated points which are less than the internally validated Copas value of 71.54.

## Results 4.5. Study 5. Goodness of fit test: Copas.

Cross Validated Copas Value Plotted Against  
Sequential Increase In Sample Size

Model ds=hciss. Second 1000 cases.



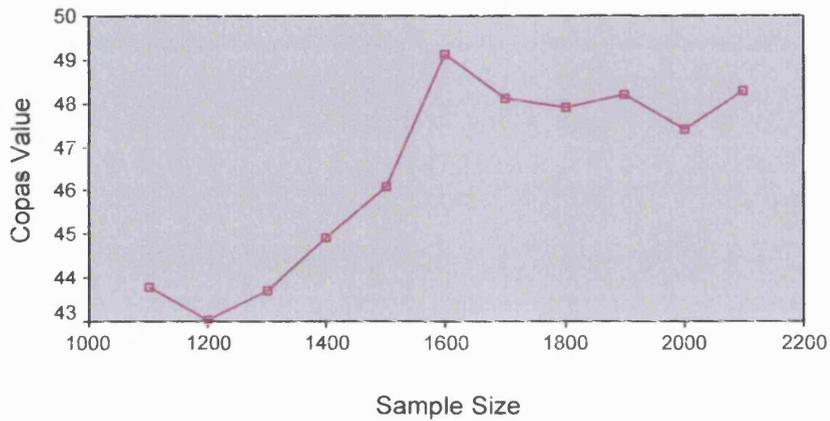
**Graph 4.5a**

Graph 4.5a shows that the first data point exceeds the internally validated Copas value of 88.75. The remaining cross validated data points show a gradual reduction in the Copas value with increasing sampling data set size.

## Results 4.6. Study 6. Goodness of fit test: Copas.

Cross Validated Copas Value Plotted Against  
Sequential Increase In Sample Size.

Model:  $ds=hciss+rts$ . First 1000 cases.



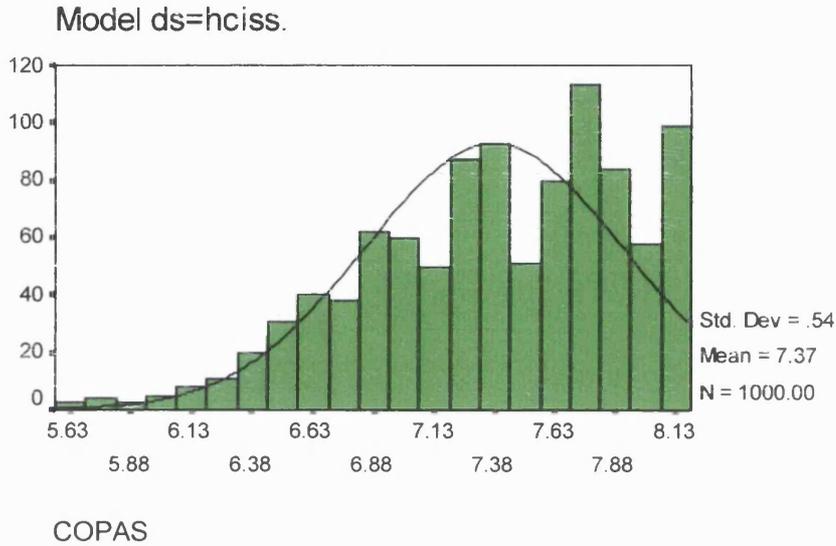
Sampling without replacement.

**Graph 4.6a**

Graph 4.6a shows an erratic increase in the cross validated Copas value with increase in sampling data set size. The internally validated Copas value derived from the first 1000 cases = 43.02.

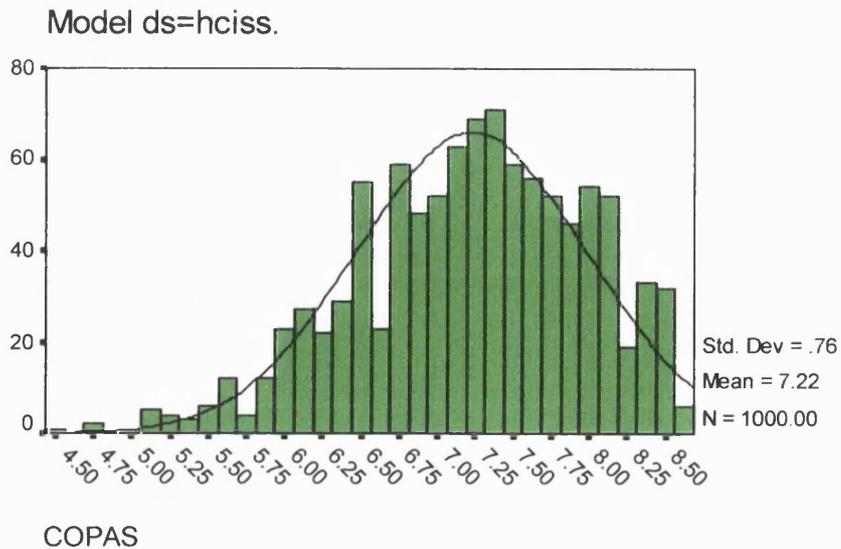
## Results 4.7. Study 7. Goodness of fit test: Copas.

Distribution of 1000 Cases Randomly  
Selected From 110 Cases



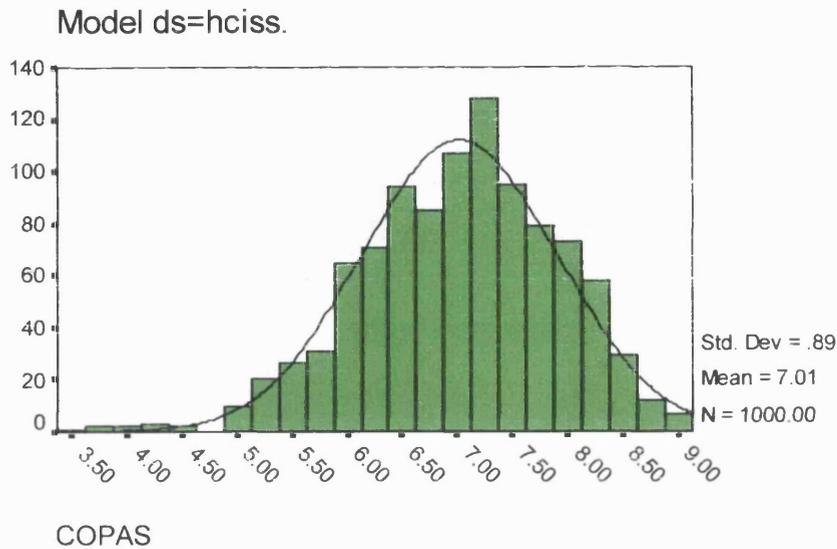
**Graph 4.7a**

Distribution of 1000 Cases Randomly  
Selected From 120 Cases.



**Graph 4.7b**

Distribution of 1000 Cases Randomly  
Selected from 130 Cases.

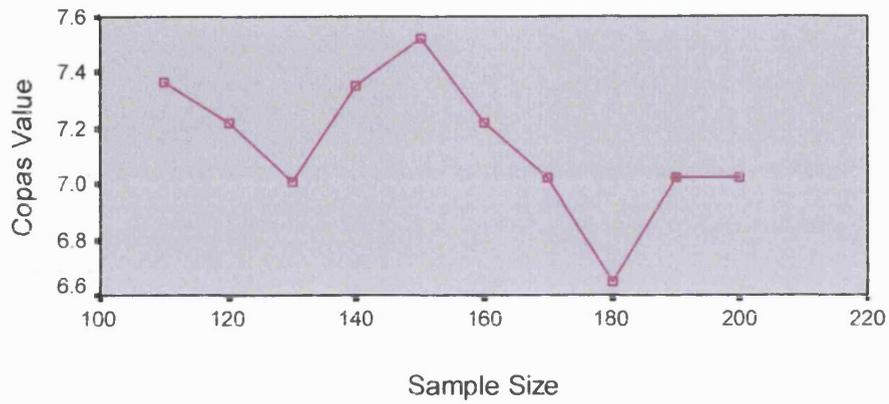


**Graph 4.7c**

Graph 4.7a shows that the distribution for the Copas cross validated samples has a poor approximation to a normal distribution, thus indicating that the sampling without replacement method is not well balanced using this data set size. Graphs 4.7b and 4.7c show a more balanced sampling distribution. For brevity only the first three sampling distributions have been included in the results section for study 7. The remaining data set sizes produced sampling distributions which were similar to those seen in graphs 4.7b and 4.7c.

## Cross Validated Copas Value Plotted Against Sample Size.

Model: ds=hciss. First 100 cases.



Sampling without replacement.

Non validated value = 7.87

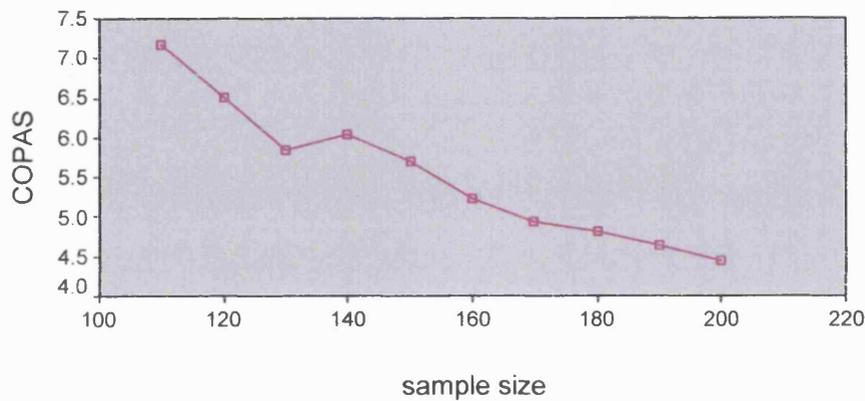
### Graph 4.7d

Graph 4.7d shows the erratic response of the cross validated Copas values when plotted against increase in sampling data set size. All data set values are less than the internally validated Copas value of 7.87.

## Results 4.8. Study 8. Goodness of fit test: Copas.

Cross Validated Copas Value Plotted  
Against Sample Size.

Model:  $ds=hciss+rts$ . First 100 cases.



Sampling without replacement.

Nonvalidated value = 6.38

### Graph 4.8a

Graph 4.8a shows a gradual reduction in the cross validated Copas value when plotted against sampling data set size. All but the first two data set values are less than the internally validated Copas value of 6.38.

## Section 5: Discussion

The results from this series of studies highlights several of the limitations of the cross validation method. Firstly, the cross validated Copas result varied erratically with sequential increase in sampling data set size. Study 3 showed an initial linear increase in the cross validated result then a downward trend (graph 4.3a). The same model (HCISS) using the second 1000 cases as the internal validation data set showed a gradual reduction in the cross validated Copas results with increasing sampling data set size (graph 4.5a). These results demonstrate that the cross validation result is affected by the size of the sampling data set. The effect however is erratic and varies with the variability within the composition of the data set (graph 4.3a c.f. 4.5a). The cross validation result also varies with the model, in this study HCISS c.f. HCISS + RTS (graph 4.3a versus 4.6a and 4.7d versus 4.8a). There was little difference in this study when comparing random sampling without replacement to random sampling with replacement (4.3a c.f. 4.4a). A second run of 1000 samples for the HCISS model showed a similar pattern to the first run (graph 4.3a c.f. graph 4.3b) thus indicating that the variability was largely due to the increase in the sampling pool rather than variability of the sampling method. Using a small data set of 100 cases produced erratic results for the HCISS model. The HCISS + RTS model showed a gradual reduction in cross validation values with increasing sampling data set size. In study 7 the distribution for the 1000 samples randomly selected from 110 cases (HCISS model) showed an unbalanced sampling distribution. This was the only data set value in all the studies which produced an unbalanced sampling distribution. The later effect can result in bias in the

sample mean. The main finding which has been demonstrated by this series of studies is that the optimal size of the sampling data set is unclear. The results from this series of simulation studies suggest that cross-validation is not a reliable method of validating a logistic trauma model.

The Copas model was selected for 7 out of the 8 studies because earlier work in this thesis (chapter 7) has shown that it produces less erratic results than the Hosmer Lemeshow statistic when used for trauma scoring modelling. The Brier score was used in one study as a simple comparison. Both the Copas test and Brier score produced balanced sampling except in one study (study 7). It was anticipated that both statistics being simple summation statistics would produce balanced sampling. Unbalanced sampling is a potential cause of error using this method. The two models chosen in this series of simulation studies were used to contrast a poor model (HCISS) with a superior one (HCISS + RTS).

Cross-validation has its origins in the Jackknife method. This method was first proposed by Quenouille (1949). Tukey (1958) recognised its potential value and named it the Jackknife. The Jackknife method involves leaving out one data point and calculating the model statistic on the n-1 data set. The process is repeated leaving out a different data point each time until all data points have been excluded and the model has been validated n-1 times. The mean result of the samples is the cross validated Jackknife result. This method can work well with small data sets but for larger data sets produces too narrow a sampling distribution. Stone (1974) and Geisser (1975) were two of the early authors to

popularise the concept of cross validation by leaving out more than one case at a time. Although the latter method was considered superior to the simple Jackknife at the time, it was soon overshadowed by the bootstrap method following the work of Efron and colleagues (Efron 1979, 1982, 1983; Efron and Tibshirani 1986; Efron and Gong 1983). An alternative to the standard cross validation method is  $k$ -fold cross-validation (Efron and Tibshirani 1998, page 240). This involves splitting the data set into  $k$  approximately equal sized parts (commonly 10). Then performing leave-one-out cross-validation on the subsets. The final result is the average of the  $k$  subsets. A variant to this is to develop the model on all subsets and to test the models on all the remaining subsets. The final result is again the average of the subsets. This method was recently used by Meredith et al (2002). Whichever method of  $k$ -fold cross-validation is used the data set can be either split sequentially or randomly. The optimal number of subsets remains unresolved. McCarthy (1976) emphasised that whatever cross validation method is used the sampling procedure should be balanced.

The cross validation method has not been widely used as a method for validating the logistic model by investigators working in the field of trauma scoring. This may in part be due to the fact that many statistical software packages do not have cross validation programs. Efron and colleagues (Efron 1979, 1982, 1983; Efron and Tibshirani 1986; Efron and Gong 1983) during the development of the bootstrap also evaluated the cross validation method. These authors found that with cross validation there was a high variation in the accuracy of the parameter estimates. Despite

these limitations Efron and Tibshirani (1998) writing in their classic monograph *An Introduction to the Bootstrap* still consider the cross validation method as being a potentially useful way of validating a model. They used the *prediction error* in many of their simulation studies to evaluate the cross validation method. The prediction error being the difference between the residual squared error (RSE) of the internally validated model and the residual square error derived by cross validation. The RSE is in fact the Copas statistic divided by  $n$  (the number of cases in the data set). Efron (1983) recommends that at least 200 cross validation samples may need to be generated in order to determine the RSE. Efron and co-workers despite their extensive work in the area of cross validation make no recommendations about the size of the sampling data set. Harrell et al (1996) in a review of multivariate regression models somewhat arbitrarily suggests that in a data set of 1000 cases cross validation might be performed on 950 cases; 50 cases being randomly left out for each cross validation sample generated. The simulation studies performed in this thesis were partly designed to address this key question and therefore complement the work of Efron and colleagues (Efron 1979, 1982, 1983; Efron and Tibshirani 1986; Efron and Gong 1983).

## **Conclusions**

The variability of the cross validation Copas results which occurred by increasing the sampling pool suggests that cross validation is not a reliable method of reducing model over-fitting in the context of trauma scoring modelling using logistic regression.

The large number of samples generated in these simulation studies makes the variability unlikely to be due to the sampling size. A control model used for study 4 showed that most of the variability could be explained by the increase in the data set size rather than due to variability of the sampling method.

## Appendix 1

### Program A: Random Selection Without Replacement.

The program below contains a truncated data set of 10 cases. The program as currently written will randomly select 8 cases from the data set (`sampsize=8;`). The `%macro repeat (n=20);` line will generate twenty Copas values. The program actually generates the mean Copas value for each run as a consequence of using the *proc means* statement. The output for each data set point was saved as a SAS file and then copied and pasted into program B, below the line statement `datalines;` (replacing the line of x's). This second program strips out the unwanted output and also multiplies the mean Copas value by the data set size (in this case = 8) to give the actual Copas value. This output was exported and saved as an *excel* file. The *excel* file was then imported into SPSS and the distribution of the cross validated samples and the cross validated mean value was obtained using the *graphs* and then *histogram* options in SPSS.

```
data usc;
  input ds iss;
datalines;
  .00 34.00
  1.00 10.00
  1.00 13.00
  .00 25.00
  1.00 25.00
  1.00 10.00
  .00 25.00
  1.00 9.00
  1.00 9.00
  1.00 9.00
;
run;

%macro repeat (n=20);
%do i=1 %to &n;
data sample&i (drop=sampsize obsleft) ;
  sampsize = 8 ;
  obsleft = totobs ;
  do while (sampsize > 0) ;
  readit + 1 ;
  if ranuni(0) < sampsize/obsleft then do ;
```

```

        set usc point=readit nobs=totobs ;
        output ;
        sampsiz = sampsiz - 1 ;
        end ;
        obsleft = obsleft - 1 ;
    end ;
stop ;
run ;

*close listing destination. i.e. output window;
ods listing close;
proc logistic data=work.sample&i des;
    model ds = iss / selection = backward;
    output out=lout&i p=pred;
run;

data work.lout&i;
    set lout&i;
    copas = (ds-pred)**2;
run;

*re-open listing destination;
ods listing;
proc means data=work.lout&i;
var copas;
output out=sum&i;
    run;
    %end;
%mend;

options symbolgen mprint mlogic;
%repeat

data final;
    set sum1 sum2 sum3 sum4 sum5 sum6 sum7 sum8 sum9 sum10
        sum11 sum12 sum13 sum14 sum15 sum16 sum17 sum18 sum19 sum20;
    run;

```

## Program B

```

data final;
    input stat $ roy1 copas1 ssr1 pear1 brier1;
    if stat = 'N' then gof=1;
    if stat = 'MAX' then gof=2;
    if stat = 'MIN' then gof=3;
    if stat = 'MEAN' then gof=4;
    if stat = 'STD' then gof=5;
    if gof ne 4 then delete;
    datalines;
xxxxxxx
;
run;

data work.final;
    set final;
    copas = copas1*8;
run;

```

## Appendix 2

### Program For Random Selection With Replacement.

The program below contains a truncated data set of 10 cases. The program as currently written will randomly select 8 cases from the data set (if  $j > 8$  then stop; and if  $i > 8$  then stop; statements). The `%macro repeat (n=20);` line will generate twenty Copas values. The program actually generates the mean Copas value for each run as a consequence of using the *proc means* statement. The output for each data set point was saved as a SAS file and then copied and pasted into program B, below the line statement `datalines;` (replacing the line of x's). The second program strips out the unwanted output and also multiplies the mean Copas value by the data set size (in this case = 8) to give the actual Copas value. This output was exported and saved as an *excel* file. The *excel* file was then imported into SPSS and the distribution of the cross validated samples and the cross validated mean value was obtained using the *graphs* and then *histogram* options in SPSS.

```
data test;
  input ds hciss rts num;
datalines;
  .00 26.00 11.00 1.00
  1.00  5.00  8.00 2.00
  1.00 13.00 12.00 3.00
  .00 25.00  4.00  4.00
  1.00 16.00 12.00 5.00
  1.00 10.00 12.00 6.00
  .00 43.00  4.00 7.00
  1.00  9.00 12.00 8.00
  1.00  9.00 12.00 9.00
  1.00 10.00 12.00 10.00
;
run;

%macro repeat (n=20);
%local i;
%do i=1 %to &n;
```

```

data test&i;
  set test point=selection nobs=n;
  k=&i*12345;
selection=int(ranuni(k)*n)+1;
j+1;
  if j > 8 then stop;
run ;

*close listing destination. i.e. output window;
ods listing close;
proc logistic data=work.test&i des;
  model ds = hciss regcs / selection = backward;
  output out=lout&i p=pred;
run;

data work.lout&i;
  set lout&i;
copas = (ds-pred)**2;
run;

*re-open listing destination;
ods listing;
proc means data=work.lout&i;
var copas;
output out=sum&i;
  run;
  %end;
%mend;

options symbolgen mprint mlogic;
%repeat;

data final;
  set sum1 sum2 sum3 sum4 sum5 sum6 sum7 sum8 sum9 sum10
  sum11 sum12 sum13 sum14 sum15 sum16 sum17 sum18 sum19 sum20;
run;

data test1;
selection=int(ranuni(12345)*n)+1;
set test point=selection nobs=n;
i+1;
  if i > 8 then stop;
  drop i;
run;

```

## Program B

```

data final;
  input stat $ royl copasl ssrl pearl brierl;
  if stat = 'N' then gof=1;
  if stat = 'MAX' then gof=2;
  if stat = 'MIN' then gof=3;
  if stat = 'MEAN' then gof=4;
  if stat = 'STD' then gof=5;
  if gof ne 4 then delete;
  datalines;
xxxxxxxxxxxx
;

```

```
run;  
  
data work.final;  
  set final;  
  copas = copas1*8;  
run;
```

# CHAPTER 10

## A STUDY TO DETERMINE THE ACCURACY OF BOOTSTRAP GENERATED CONFIDENCE INTERVALS FOR TWO GOODNESS OF FIT TESTS IN LOGISTIC REGRESSION

### CONTENTS

	<b>Page Number</b>
<b>Section 1: Introduction</b>	
<b>Overview of bootstrapping</b>	<b>246</b>
<b>Bootstrap confidence intervals</b>	<b>247</b>
<b>Bootstrap-<math>t</math> interval</b>	<b>248</b>
<b>Percentile method</b>	<b>250</b>
<b>The BC method</b>	<b>250</b>
<b>The BCa method</b>	<b>251</b>
<b>Section 2: Aims</b>	<b>252</b>
<b>Section 3: Methodology</b>	<b>252</b>
<b>Section 4: Results</b>	<b>257</b>
<b>Section 5: Discussion</b>	<b>295</b>
<b>Appendix 1:</b>	<b>297</b>
<b>Appendix 2:</b>	<b>299</b>

## **Section 1: Introduction**

### **Overview of Bootstrapping.**

Bootstrapping is a computer based method which provides measures of accuracy to statistical estimates. The bootstrap procedure is one of several methods that generates multiple new samples by resampling from the original data. The method was developed and popularised by Efron and Tibshirani in the 1970s. Resampling procedures have been around for several decades, for example the Jackknife method was proposed by Quenouille in 1949. A sentinel paper by Efron (1979) crystallised statistical thinking by unifying resampling procedures such as the Jackknife method with the bootstrap procedure. A series of papers by Efron and co-workers (Efron 1982, Efron and Gong 1983, Efron and Tibshirani, 1986) helped to establish the bootstrap procedure as an important tool in the statisticians armamentarium. Since the early work by Efron and co-workers numerous simulation studies have been performed and these have highlighted both the strengths and limitations of the bootstrap method.

The bootstrap procedure enables standard errors and confidence intervals to be calculated for a variety of statistical parameters. The simplest example being the mean. A bootstrap sample is produced by generating a sample of size  $n$  (where  $n$  equals the size as the original sample) by randomly selecting cases from the original sample. Once the first case has been randomly selected it is replaced back into the sample and the second case is again randomly selected from the whole of the original sample. This is in contrast to random selection without replacement where a case

which once selected is not returned to the original sample. In random selection with replacement the probability of selecting a case is equal to its frequency with which it occurs in the sample. A distribution where the frequency or probability of each case is defined within a sample is called the *empirical distribution*.

The bootstrap procedure is repeated k times (usually at least 100) and the parameter of interest is calculated from each bootstrap sample ( $\theta^*$ ). If  $\theta^*$  represents the mean then the standard deviation of the bootstrap samples is the standard error for  $\theta^*$ . Bootstrapping programs are unfortunately not routinely available in all statistical software packages. The SAS website has a random sampling with replacement program that can be used for bootstrapping.

## Bootstrapping Confidence Intervals

### Confidence Intervals: Theory

Given a situation where an estimator  $\bar{\theta}$  is normally distributed the statistic equalling  $(\bar{\theta} - \theta) / se$  has a standard normal distribution.

i.e. 
$$Z = \frac{\bar{\theta} - \theta}{se} \sim N(0,1) \quad \{se = \text{standard error}\}$$

Let  $z^{(\alpha)}$  indicate the  $\alpha$ th percentile point of a N (0,1) distribution as given in a standard normal distribution table.

The interval  $[\bar{\theta} - z^{(1-\alpha)} \cdot \bar{se}, \bar{\theta} - z^{(\alpha)} \cdot \bar{se}]$  is known as the standard confidence interval with coverage probability equal to  $1 - 2\alpha$ . Since  $z^{(\alpha)} = -z^{(1-\alpha)}$  the formula can be written in the more familiar form:-  $\bar{\theta} \pm z^{(1-\alpha)} \cdot \bar{se}$

A better approximation for the confidence interval has been shown to be :-

$$Z = \frac{\bar{\theta} - \theta}{se} \sim N(0,1) \quad t_{n-1}, \text{ where } t_{n-1} \text{ represents the Student's } t$$

distribution with  $n-1$  degrees of freedom. The confidence interval using the Student's  $t$  distribution with  $n-1$  degrees of freedom becomes:-  $\bar{\theta} - t^{1-\alpha} \cdot se, \bar{\theta} - t^{\alpha} \cdot se$

The Student- $t$  confidence interval becomes exact if the observations are normally distributed. The disadvantage of the Student- $t$  based confidence interval is that the  $t$  distribution doesn't adjust for skewness in the population.

### **The bootstrap- $t$ interval**

(also known as the 'percentile- $t$ ' in Hall, 1992 and 'studentized- $t$ ' in Hjorth, 1994)

The bootstrap- $t$  interval was developed by Efron (1979, 1982). It is the simplest of all the bootstrap methods particularly for location statistics such as the sample mean. A location statistic is a parameter which when each data value  $x$  is increased by a constant  $c$ , the parameter also increases by  $c$ . According to Chernick (1999) the bootstrap- $t$  method can produce confidence intervals of high accuracy. The bootstrap- $t$  confidence intervals can be asymmetric about 0. This according to Efron and Tibshirani (1998; page 161) can provide improved coverage over the standard Student- $t$  confidence interval. The bootstrap- $t$  confidence intervals can be very long according to SAS (1995). Efron and Tibshirani (1998; page 160) have shown from their own simulation studies that the

bootstrap- $t$  method can produce erratic results and can be overly sensitive to outliers.

The bootstrap- $t$  method calculates the  $t$  value directly from the data. A table of percentiles are created from  $B$  bootstrap samples from which the confidence intervals are constructed. The strength of the bootstrap- $t$  method is that it can produce accurate intervals for location statistics such as the sample mean.

The annotation for the bootstrap- $t$  method is:-

$$Z^*(b) = \frac{\bar{\theta}^*(b) - \bar{\theta}}{se^*(b)}$$

given  $B$  bootstrap samples  $x^{*1}, x^{*2}, x^{*3} \dots x^{*B}$

$Z^*(b)$  is the standardized value of the bootstrap sample  $x^{*B}$

$\bar{\theta}$  = the test statistic (such as the sample mean)

$\bar{\theta}^*(b)$  = is the value of the  $\bar{\theta}$  for the bootstrap sample  $x^{*b}$

$se$  is the estimated standard error of  $\bar{\theta}^*$  for the bootstrap sample  $x^{*B}$

### **Calculation of the $\alpha$ percentile of the $Z^*(b)$ :**

#### **bootstrap- $t$ method.**

If  $B = 1000$  bootstrap samples, the estimate for the 5% point is the 50<sup>th</sup> largest value of the  $Z^*(b)$ s. The 95% point is the 950 largest value of the  $Z^*(b)$ s. A difficulty with the bootstrap- $t$  method is that it requires an estimate of the standard error of each statistic being bootstrapped. If the standard error is not estimated accurately, for instance because the distribution of the bootstrap samples is not

close to a normal distribution then the bootstrap- $t$  method may give poor results. Once the 5% and 95% bootstrap- $t$  values have been determined the confidence interval for the parameter is calculated in the standard way ( $\bar{\theta} \pm \{t^{1-\alpha} \times \bar{se}\}$ ).

## **Percentile Method**

The percentile method uses the  $\alpha/2$  and  $1-\alpha/2$  percentiles of the bootstrap histogram to define the interval. Where the bootstrap distribution of a parameter  $\bar{\theta}^*$  is not normally distributed the percentile method also gives poor results. Occasionally using a transformation (e.g. logarithmic) the distribution of  $\bar{\theta}^*$  can be normalised. The percentile method according to SAS (1995) works well for parameters which have a symmetric sampling distribution. Efron and Tibshirani (1998) have shown that the percentile method works well in some cases i.e. median even though it is not exact. The percentile method tends to work well if exactly 50% of the bootstrap distribution for the estimated parameter is less than the bootstrap mean for the parameter.

## **The BC Method**

The BC method corrects the percentile interval for bias-median not bias mean. The correction is performed by adjusting the percentile points to values other than  $\alpha/2$  and  $1-\alpha/2$ . The bias is calculated by subtracting the value of the parameter at its 50 percentile from the parameter mean of the bootstrap distribution. If a large correction is required one of the percentiles will be very small. A large

number of samples will be required to approximate the intervals accurately.

### **The BCa Method**

The BCa (full name :- bias correction and acceleration) method corrects the percentile method for both bias and for skewness. This method requires an estimate of the bias and acceleration constants. The acceleration constant is related to the skewness of the bootstrap sampling distribution. If the acceleration constant is not estimated accurately the BCa interval will not work well. For large values of the acceleration constant, the BCa interval is excessively short (SAS, 1995). The BCa confidence interval is no better than the BC for non-smooth statistics such as the median (SAS 1995).

Several other methods for calculating bootstrap confidence intervals have been developed in the last two decades. A review of these methods is provided by Carpenter (2000). All of these methods have their strengths and weaknesses, only a few have been incorporated into statistical software packages. The SAS bootstrapping program which can be downloaded from the SAS website has several methods for calculating bootstrap programs but is very cumbersome to use. STATA has a procedure for calculating percentile and BC confidence intervals but requires a user specific program. S-plus has the most comprehensive range of bootstrapping programs but again requires knowledge of the program (s-plus) language. The percentile method despite its limitations is intuitively and computationally the easiest to use and is the method which was used in this study.

## **Section 2: Aims**

The aim of this study was to determine whether the bootstrap method could be applied to two goodness of fit tests used in logistic regression in order to provide confidence intervals for these statistics.

## **Section 3: Methodology**

### **Section 3.1**

The revised USC data set was used in this study.

### **Section 3.2**

Logistic regression was performed using SAS version 8. Method of model selection chosen was:- Backward Selection.

The descending option was used so that predicted probabilities were calculated for survival rather than death.

### **Section 3.3**

Two models were selected;

First model: Dependent variable:- death/survival

Predictor variable:- HCISS (hand calculated ISS)

Second model: Dependent variable:- death/ survival

Predictor variable:- HCISS and RTS (unweighted)

These two models were chosen because Champion et al (1983) have shown that ISS + RTS is a superior model to ISS alone. An accurate confidence interval should be able to distinguish between

these two models. A confidence interval which is not able to distinguish between these two models is probably too wide in its coverage.

### **Section 3.4**

Copas and HL values were calculated for both models over a range of data set sizes:

Data set size = 1000 (first 1000 cases: subgroup 1)

Data set size = 3000 (first 3000 cases: subgroup 2)

Data set size = 5000 (first 5000 cases: subgroup 3)

Data set size = 7000 (first 7000 cases: subgroup 4)

### **Section 3.5**

The method for calculating the Copas statistic has been previously described. The Copas test is a measure of predictive accuracy.

$$\text{Copas} = \sum (ds - \text{pred prob})^2$$

ds : outcome variable – death or survival

pred prob – predicted probabilities

The method used to calculate the Hosmer Lemeshow statistic was the algorithm method used by the *proc logistic* program in SAS. This method has been previously discussed in chapter 5.

### **Section 3.6**

A data set control group was also run for the Copas HCISS model. This control data set was designed so that the data set sizes of 3000, 5000 and 7000 were created by having multiples of the first 1000 cases i.e. The control data set of 3000 cases was composed of three sets of the first 1000 cases. The reason for the control group was to more clearly define the effect of increasing the data set size

on the confidence interval coverage. It was postulated that the confidence interval coverage values would increase as the data set was increased. This increase may be due to an increase in the variability of the data set rather than just a simple increase in data set size. The use of a control group should enable these two confounding effects to be distinguished.

### **Section 3.7**

A bootstrap program was obtained from the SAS website ([www.sas.com](http://www.sas.com)). The program performs random sampling with replacement using the *ranuni* random numbers generator. The *ranuni* value was set by the programmer so that for a given data set sample the same sequence of numbers and hence parameter estimates are produced each time the seed is activated (pseudo random number generator). The bootstrap program (*macro*) was merged with the standard program for performing logistic regression (*proc logistic*) in SAS. The Copas statistic was generated by writing an additional program due to the fact that the SAS *proc logistic* program does not output this statistic. A *macro repeat* program was also added to these two programs so that the final program (program A; see appendix 1) would automatically repeat itself multiple times and output each goodness of fit test result.

### **Section 3.8**

The *proc logistic* program as previously mentioned was used for modelling in program A. The statistic  $(DS - \text{Predicted probability})^2$  was calculated for each case. The mean value for all the cases was obtained and outputted using the *proc means* program. Each mean

result was initially outputted into a separate data set called *sum*. All of these data sets were merged (strictly speaking concatenated i.e. the data is merged vertically) using the *set* statement to produce a single output data set named *final*. The *final* data set was exported and saved as an excel file. The *stat* and *Copas* variable columns were copied and then pasted into program B (see appendix 2). The latter program converted the mean Copas result into the actual Copas value for the bootstrap data set by multiplying the mean Copas value by the data set size value. The final data set was again exported and saved as an excel file. This file was read into an SPSS spread sheet. A histogram was generated for the data which also gives the bootstrap mean of the parameter by using the *graphs* option on the menu bar. The 50<sup>th</sup> and 950<sup>th</sup> largest parameter estimates (Copas) values were obtained by using the *sort* option (by smallest value) within the data option, which is also on the main menu bar.

### **Section 3.9**

1000 bootstrap samples were produced for both models. The 10% confidence interval was calculated using the percentile method described by Efron et al (1998) i.e. the 5% confidence value (percentile) is that result corresponding to the 50<sup>th</sup> highest Copas value. The 95% confidence value (percentile) is that result corresponding to the 950<sup>th</sup> highest value. The 10% coverage value is defined as the difference between the upper (95%) and lower (5%) confidence values.

### **Section 3.10**

The Copas values for each model were calculated for the four

different data set subgroups (plug in estimates) using the same method as that used to generate the bootstrap samples. i.e. SAS version 8, model selection:- backward likelihood ratio. The bootstrap mean was calculated as the  $\sum$ bootstrap values/1000. The difference between the plug in estimate and the bootstrap mean is defined as the bias for that parameter.

Using the method used by Efron et al (1998, page 138):-

Bias corrected Copas (BCC) = Copas (plug in value) – Bias

Bias = Bootstrap mean Copas – Copas (plug in value)

### **Section 3.10**

Evidence that the bootstrap method was effective in producing confidence intervals was based upon:-

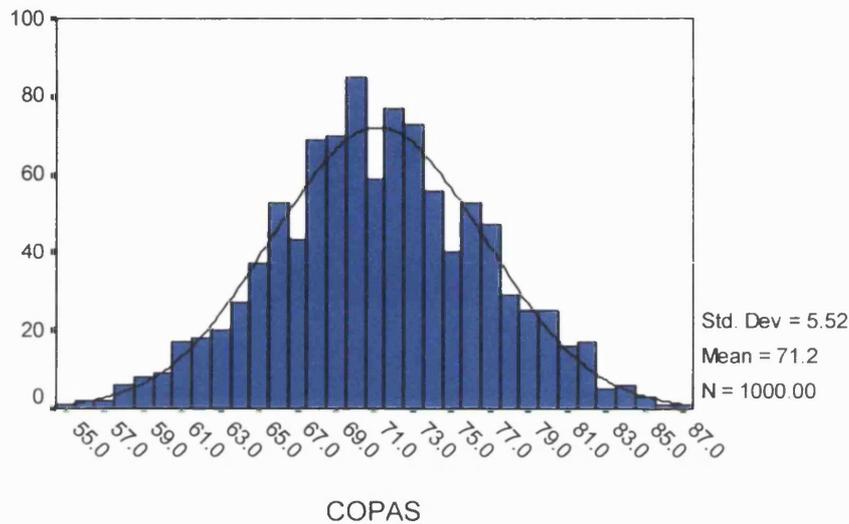
- (1) The near normal distribution of the bootstrap values which should be apparent when histograms are created for the bootstrap samples.
- (2) A near normal distribution following a logarithmic transformation when the distributions are clearly non-normal.
- (3) The ability of the two confidence intervals to distinguish between the two models.
- (4) A small difference between the bootstrap mean and the plug in estimate (i.e. in the order of one or less). Efron et al (1998, page 138) warn that if the bias is large compared to the standard error (in this case the confidence interval) then bootstrapping may not be the appropriate means of estimation.

## Section 4: Results

### Section: 4.1

#### Histogram of 1000 Bootstrap Copas Values

Model DL=HCISS.

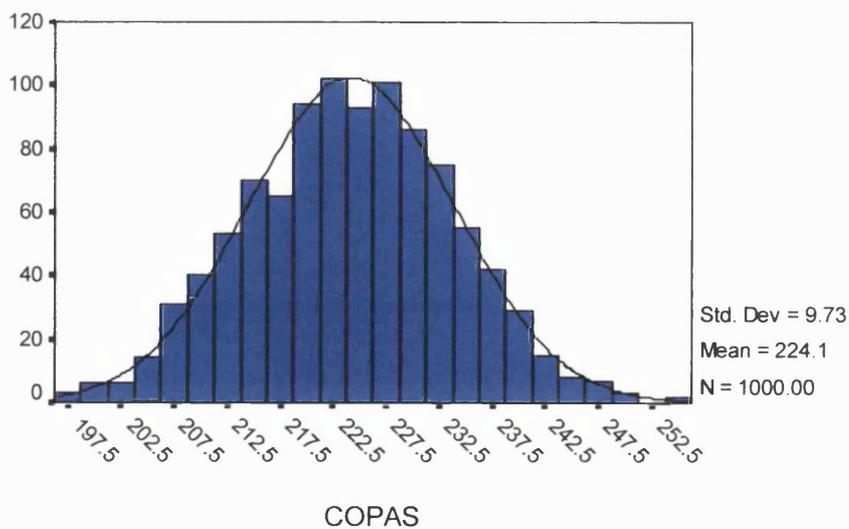


First 1000 Cases

**Graph 4.1a**

#### Histogram of 1000 Bootstrap Copas Values

Model: DL=HCISS

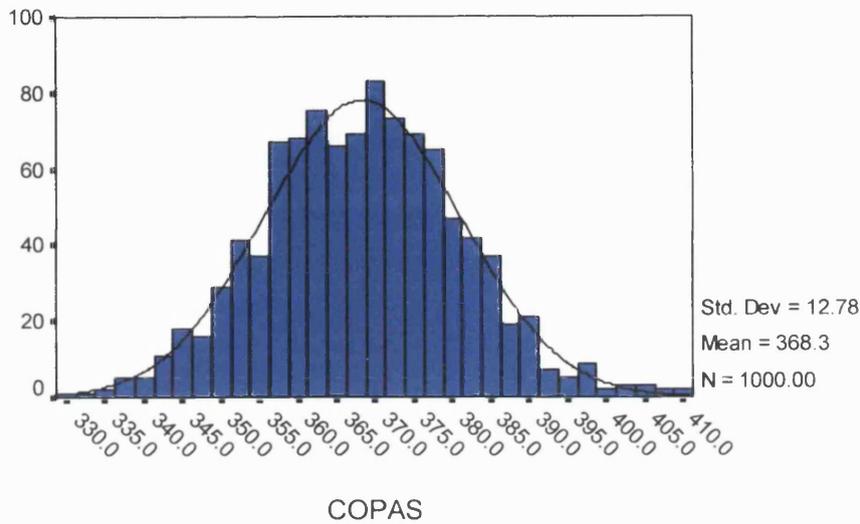


First 3000 Cases

**Graph 4.1b**

### Histogram of 1000 Bootstrap Copas Values

Model: DL=HCCISS

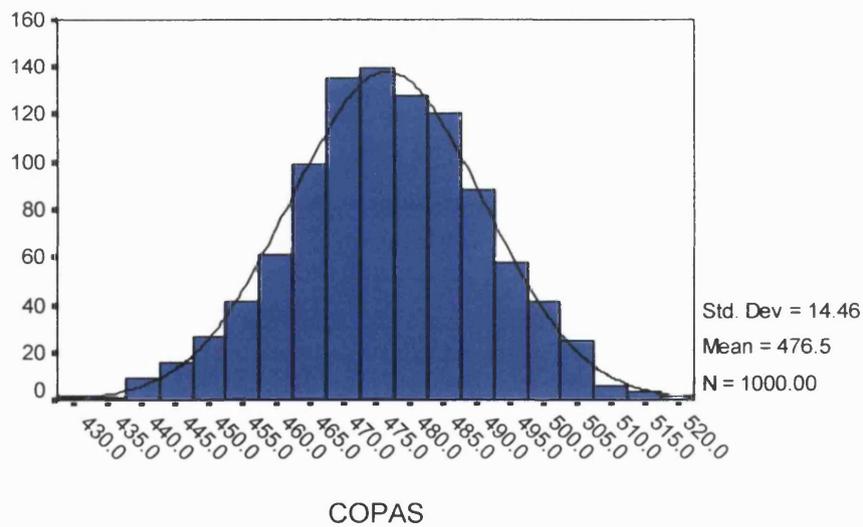


First 5000 Cases

**Graph 4.1c**

### Histogram of 1000 Bootstrap Copas Values

Model: DL=HCCISS



First 7000 Cases.

**Graph 4.1d**

Graphs 4.1a to 4.1d show the approximate normal distribution of the bootstrap samples for the Copas values using the HCISS model. A normal distribution curved is superimposed on the histograms for clarity.

**Table 4.1**

Model: DL=HCISS. Copas 90% Confidence Interval.  
(Percentile Method)

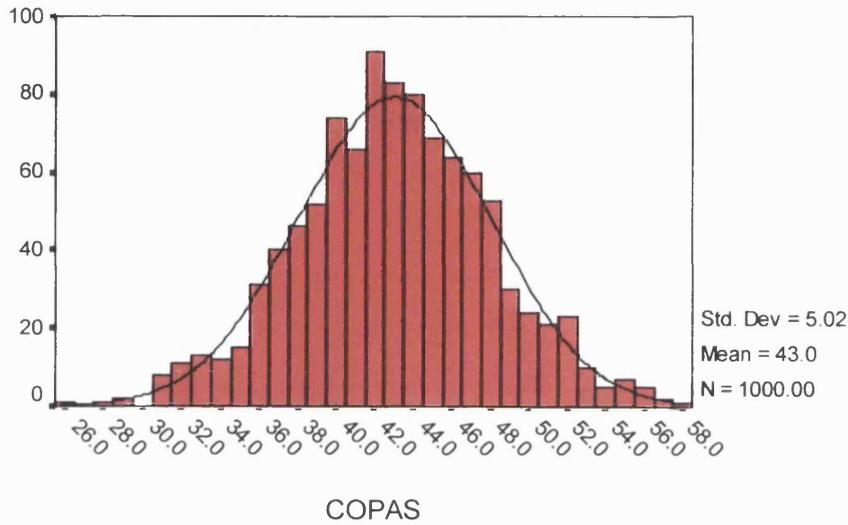
Data Set Size	Copas 95%:LCI	Copas 95%:UCI	Bootstrap Copas mean	Plug in Estimate Copas value	coverage value
1000	61.71	80.47	71.2	71.54	18.76
3000	207.96	240.05	224.1	224.47	32.09
5000	347.26	388.94	368.3	368.29	41.68
7000	451.79	500.96	476.5	476.81	49.17

Table 4.1 shows that as the data set size increases the Copas coverage values also increase. The table also shows that the bootstrap mean is close in value to the plug in estimate.

**Section: 4.2**

**Histogram of 1000 Bootstrap Copas Values**

Model: DL=HCCISS+RTS

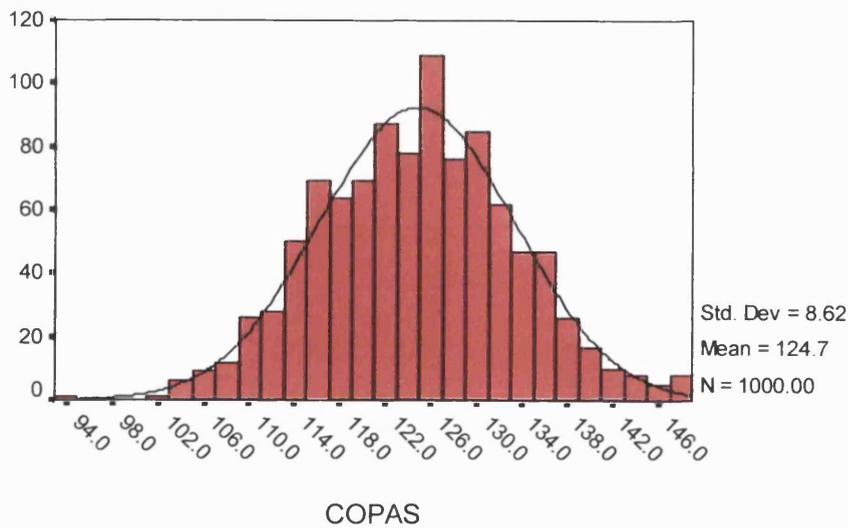


First 1000 Cases.

**Graph 4.2a**

**Histogram of 1000 Bootstrap Copas Values**

Model: DL=HCCISS+RTS

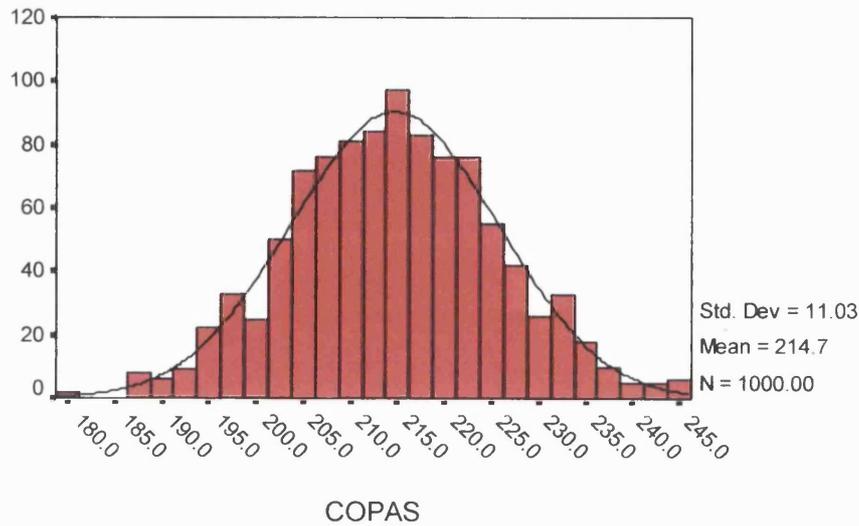


First 3000 Cases.

**Graph 4.2b**

## Histogram of 1000 Bootstrap Copas Values

Model: DL=HCISS+RTS

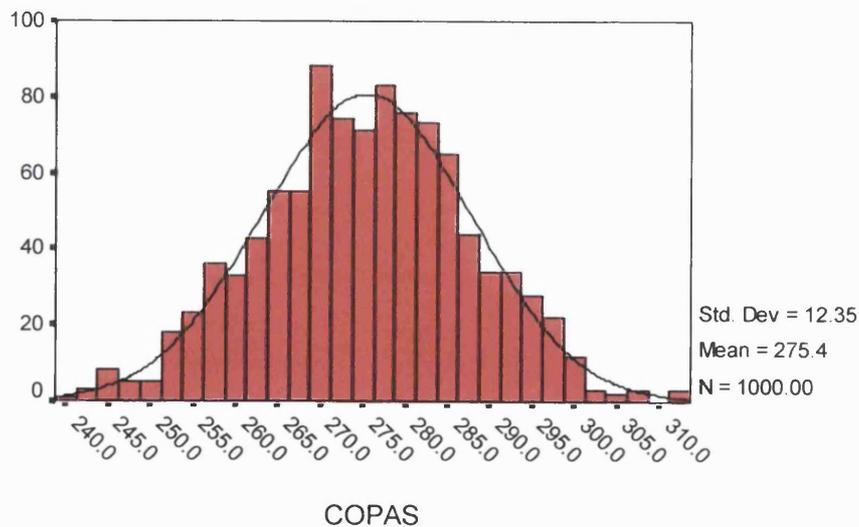


First 5000 Cases.

**Graph 4.2c**

## Histogram of 1000 Bootstrap Copas Values

Model: DL=HCISS+RTS



First 7000 Cases.

**Graph 4.2d**

Graphs 4.2a to 4.2d show the approximate normal distribution of the bootstrap samples for the Copas values using the HCISS + RTS model. A normal distribution curve is superimposed on the histograms for clarity.

**Table 4.2a**

Model: DL=HCISS + RTS. Copas 90% Confidence Interval.  
(Percentile Method)

Data Set Size	Copas 95%:LCI	Copas 95%:UCI	Bootstrap Copas mean	Plug in Estimate Copas	Coverage value
1000	34.62	51.60	43.0	43.02	16.98
3000	110.18	138.87	124.7	125.08	28.69
5000	196.49	233.20	214.7	212.18	36.71
7000	254.81	295.72	275.4	276.11	40.91

Table 4.2 shows that as the data set size increases the Copas coverage values also increase. The table also shows that the bootstrap mean is close in value to the plug in estimate.

**Table 4.2b**

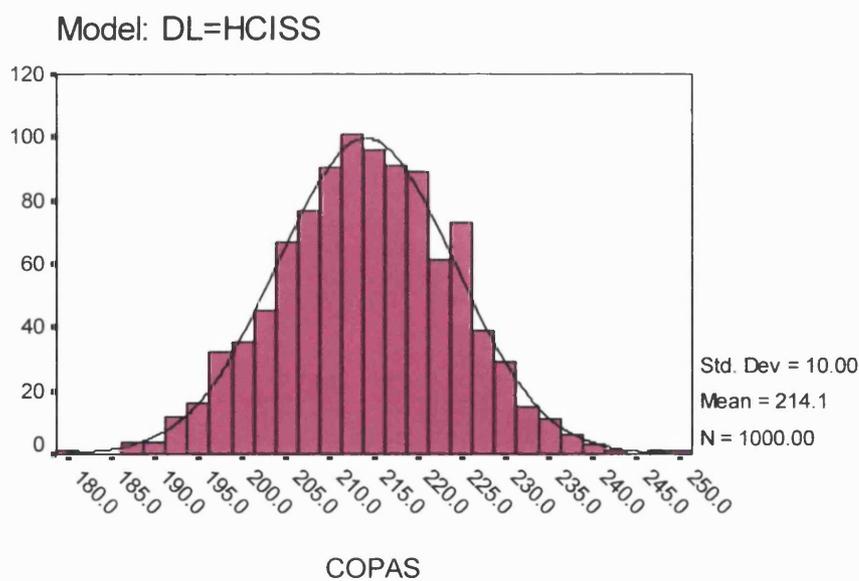
Model: DL=HCISS. Copas 90% Confidence Interval.  
(Percentile Method)

Data Set Size	Copas 95%:UCI Model: HCISS + RTS	Copas 95%:LCI Model: HCISS
1000	51.60	61.71
3000	138.87	207.96
5000	233.20	347.26
7000	295.72	451.79

Table 4.2b shows that the value for the upper confidence interval limit for the HCISS + RTS model is less than the lower confidence interval limit for the HCISS model for all four data set values. This suggests that the coverage value is appropriate.

### Section 4.3

**Histogram of 1000 Bootstrap Copas Values**

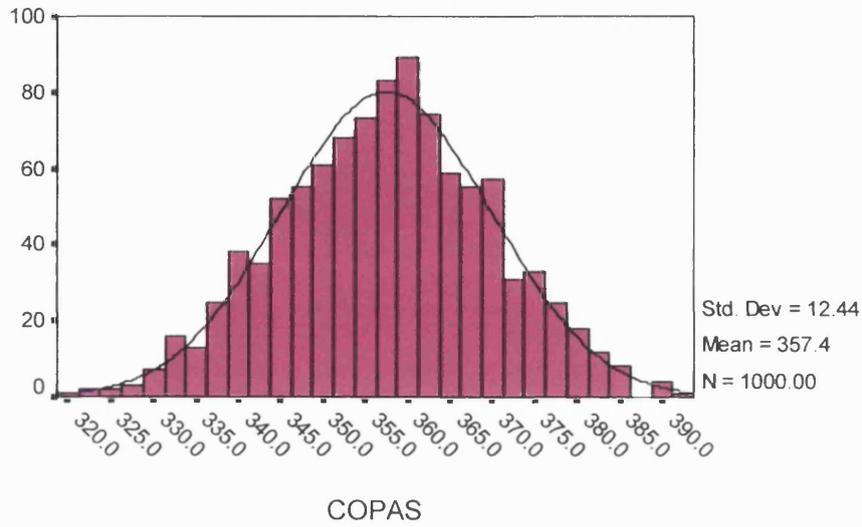


Control (3000 Cases).

**Graph 4.3a**

## Histogram of 1000 Bootstrap Copas Values

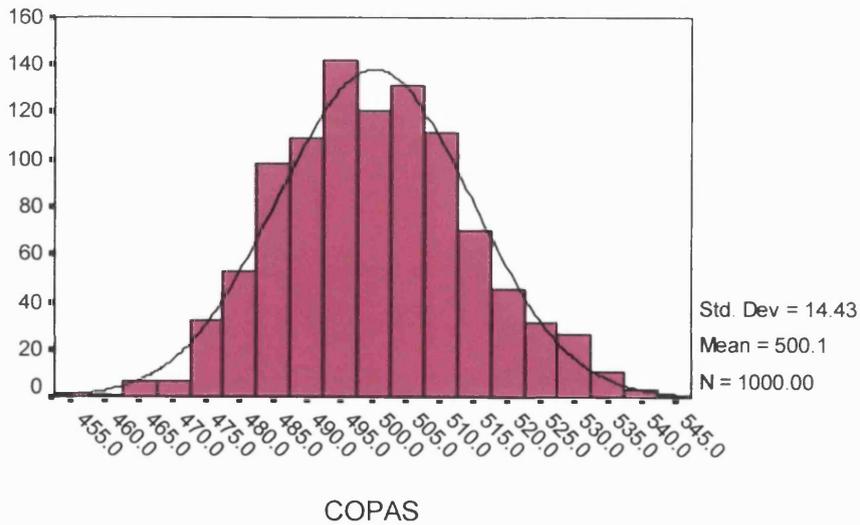
Model: DL=HClSS



**Graph 4.3b**

## Histogram of 1000 Bootstrap Copas Values

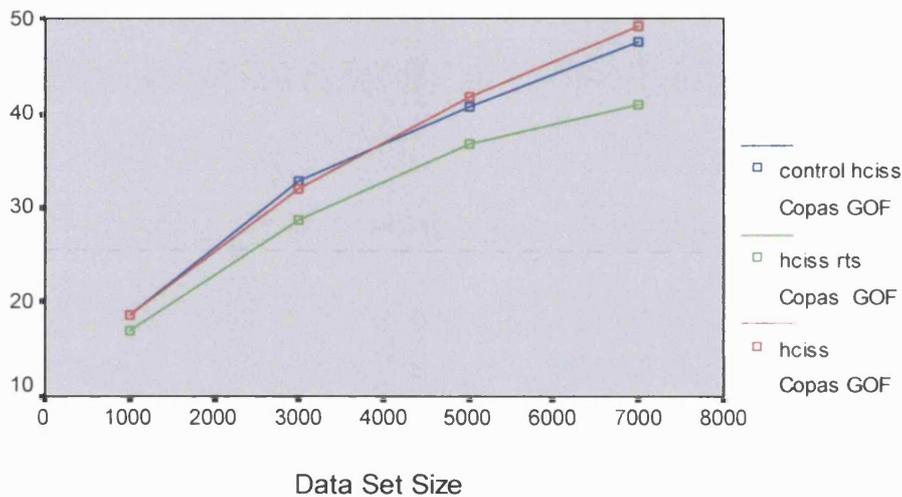
Model: DL=HClSS



**Graph 4.3c**

Graphs 4.3a to 4.2c show the approximate normal distribution of the bootstrap samples for the Copas values using the HCISS control model. A normal distribution curved is superimposed on the histograms for clarity.

90% Coverage Values (Percentile Method) Plotted Against Data Set Size  
Copas GOF Test.



**Graph 4.3d**

Two models were used for the bootstrap Copas confidence interval

1. DL = HCISS:- RED LINE
2. DL = HCISS + RTS:- GREEN LINE

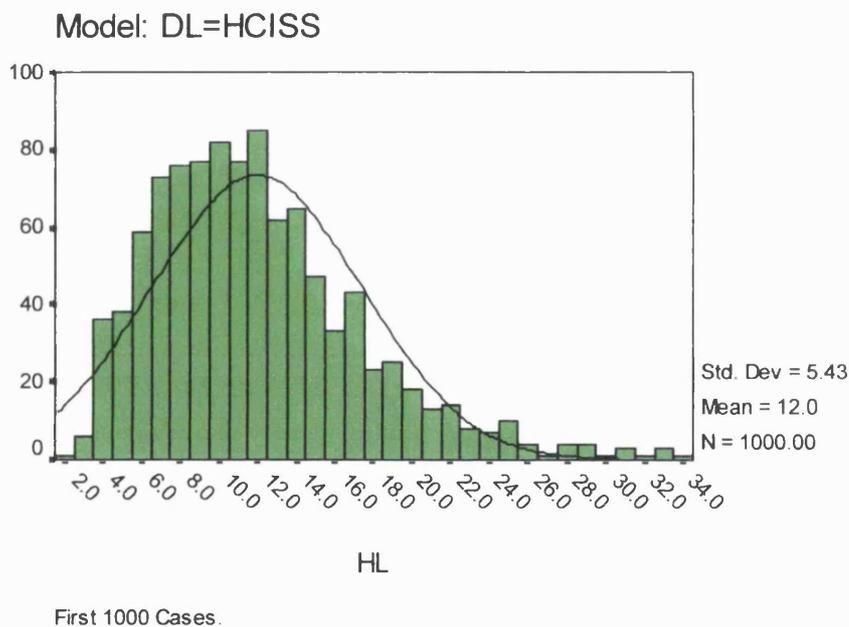
A control was used for the Copas model DL = HCISS, by sequentially repeating the first 1000 cases:- BLUE LINE

Graph 4.3d shows an increase in the Copas 90% coverage values for both models with increase in data set size. The graph also

shows that the Copas coverage values (HCISS model) using the control data set are very similar to the HCISS model indicating that the increase in coverage values is due principally to the size of the data set rather than an increased variability in the composition of the data set.

#### Section 4.4

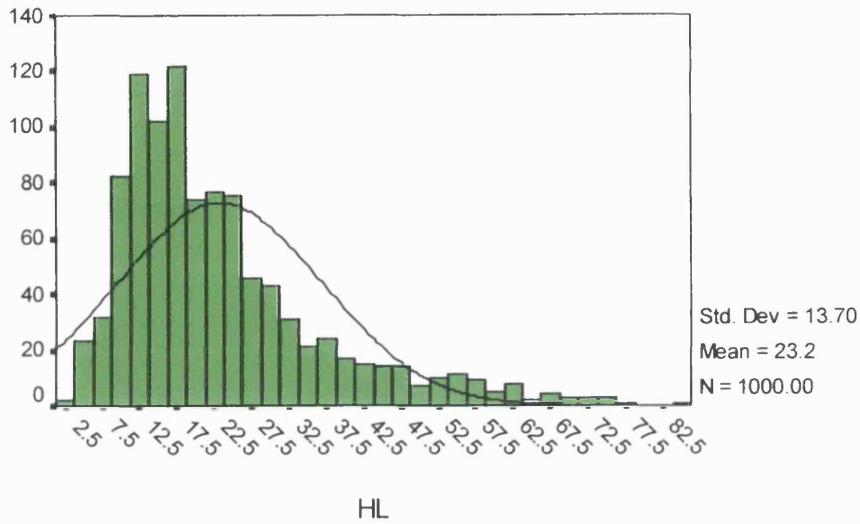
#### Histogram of 1000 Bootstrap HL Values



Graph 4.4a

### Histogram of 1000 Bootstrap HL Values

Model: DL=HCCISS

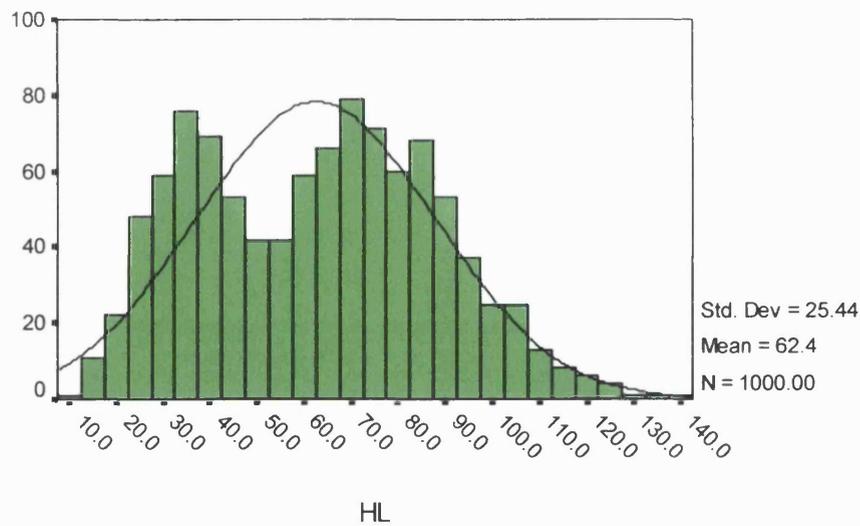


First 3000 Cases.

**Graph 4.4b**

### Histogram of 1000 Bootstrap HL Values

Model: DL=HCCISS

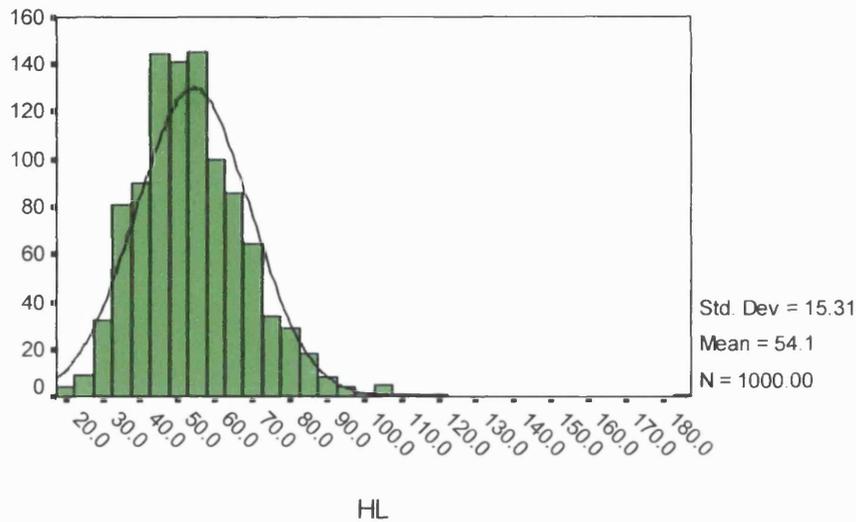


First 5000 Cases.

**Graph 4.4c**

## Histogram of 1000 Bootstrap HL Values

Model: DL=HCCISS



First 7000 Cases.

**Graph 4.4d**

Graphs 4.4a and 4.4b show a non-normal distribution of the bootstrap samples for the HL values using the HCCISS model, both histograms have a long left hand tail. Graph 4.4c also demonstrates a non-normal distribution this time with two peaks. Graph 4.4d shows a distribution which resembles that of a normal distribution but which is fact too narrow (c.f. graph 4.3c).

**Table 4.4**

Model: DL=HCCISS. HL 90% Confidence Interval.

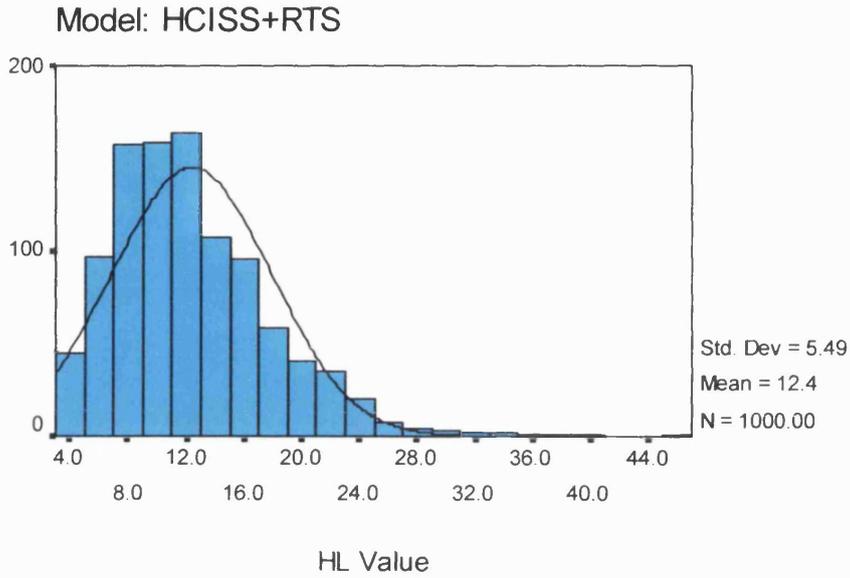
(Percentile Method)

Data Set Size	HL: LCI 95%	HL: UCI 95%	Bootstrap HL mean	Plug in Estimate (HL) value	Coverage value
1000	4.71	22.14	12.00	6.8360	17.43
3000	8.37	53.18	23.20	12.9723	44.81
5000	24.17	104.23	62.40	71.0356	80.06
7000	32.84	81.03	54.10	50.0912	48.19

Table 4.4 shows that the coverage values increase for the first three data set points and then decreases for the last point. Note also the larger differences between the plug in mean and the bootstrap mean compared to the Copas results. Both results indicate that the bootstrap method is not an accurate way of deriving confidence intervals for the HL statistic.

## Section 4.5

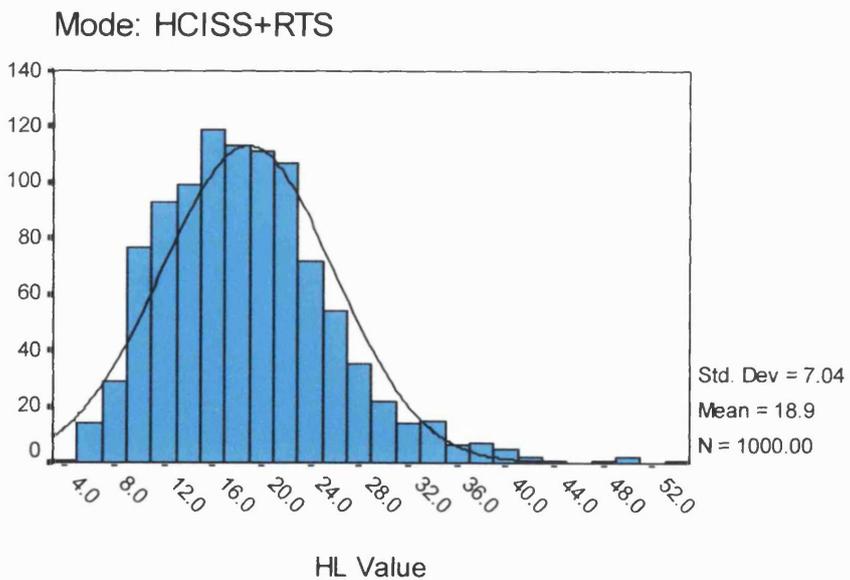
### Histogram of 1000 Bootstrap HL Values



First 1000 Cases.

**Graph 4.5a**

### Histogram of 1000 Bootstrap HL Values

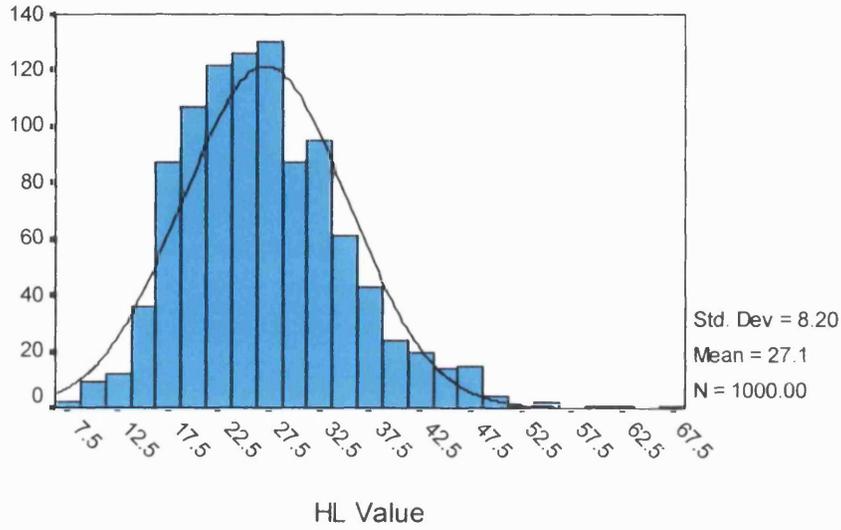


First 3000 Cases

**Graph 4.5b**

### Histogram of 1000 Bootstrap HL Values

Model: HClSS+RTS

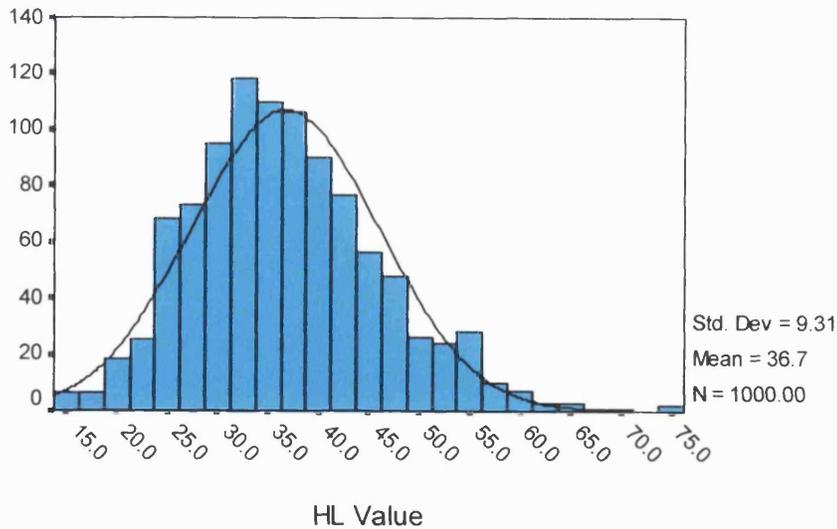


First 5000 Cases.

**Graph 4.5c**

### Histogram of 1000 Bootstrap HL Values

Model: HClSS+RTS



First 7000 Cases.

**Graph 4.5d**

Graph 4.5a shows that the distribution of the HL values has a long right hand tail. Graph 4.5b shows that the HL distribution for the first 3000 cases is too wide. Graphs 4.5c and 4.5d shows a reasonable approximation to a normal distribution.

**Table 4.5a**

Model: DL=HCISS + RTS. HL 90% Confidence Interval.  
(Percentile Method)

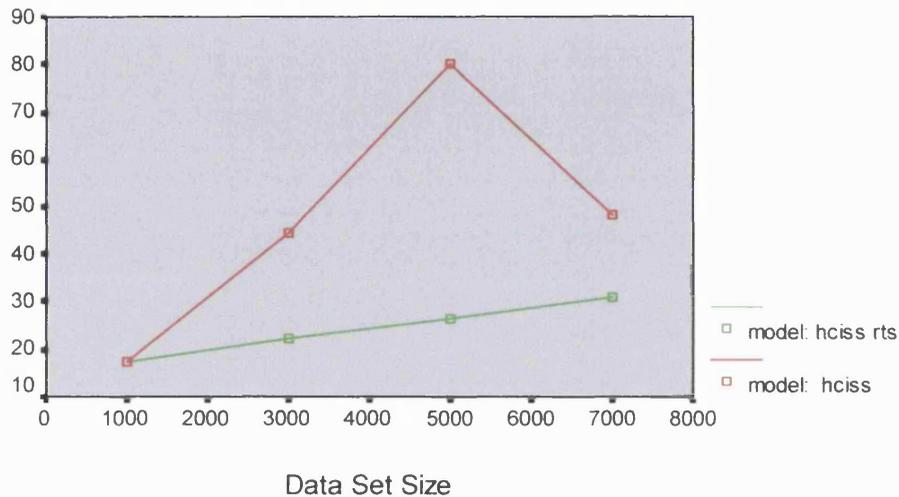
Data Set Size	HL: LCI 95%	HL: UCI 95%	Bootstrap HL mean	Plug in Estimate (HL)	Coverage value
1000	5.08	22.32	12.40	7.2995	17.24
3000	9.24	31.45	18.90	15.0797	22.21
5000	15.65	42.15	27.10	21.5701	26.50
7000	23.12	53.99	36.70	32.4852	30.87

Table 4.5 shows that as the data set size increases the coverage values also increase. Note also the relatively large difference in the plug in estimate and the bootstrap mean. The latter finding again indicates that the bootstrap method is not an accurate way of deriving confidence intervals for the HL statistic.

## 90% Coverage Values (Percentile Method)

### Plotted Against Data Set Size

HL GOF test.



**Graph 4.5e**

Graph 4.5e compares the coverage values for both models using the HL statistic plotted against increase in data set size. The HCISS model shows an erratic increase in the coverage value as the data set size increases compared to the linear increase with the HCISS + RTS model. Note also that the coverage values for the HCISS model are considerable larger than the coverage values for the HCISS + RTS models. This is in contrast to the Copas coverage values (graph 4.3d) where the difference is less marked.

**Table 4.5b**

HL 90% Confidence Interval: Percentile Method.

A comparison of two models

Data Set Size	HL: LCI 95% Model: HCISS	HL: UCI 95% Model: HCISS + RTS
1000	4.71	22.32
3000	8.37	31.45
5000	24.17	42.15
7000	32.84	53.99

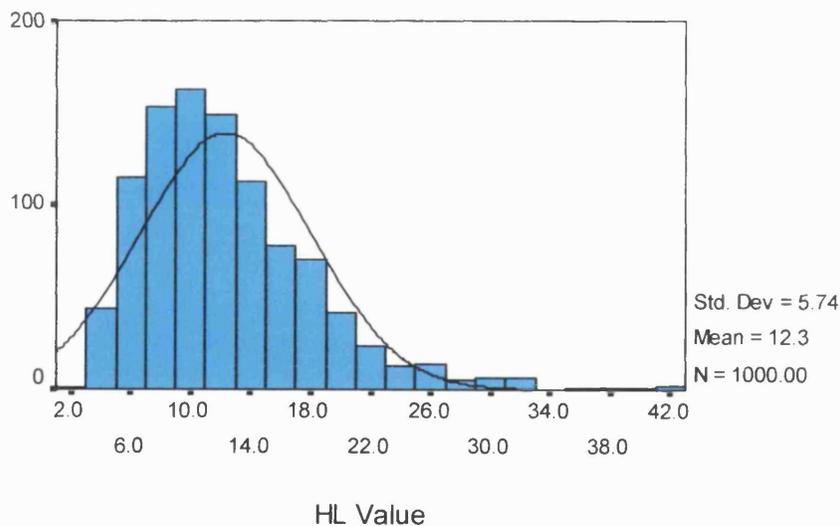
Table 4.5b shows that the lower confidence limit for the HCISS model is lower than the upper confidence limit for the HCISS + RTS model. This indicates that the bootstrap confidence intervals are too wide and hence are unable to separate between the two models.

## Section 4.6

In view of the erratic results of the HL coverage values for the HCISS model a second set of 1000 bootstrap HL cases were generated by running the bootstrap program 2000 times. Cases 1001 to 2000 were used to produce the second set of 1000 bootstrap results.

### Histogram of 1000 Bootstrap HL Values

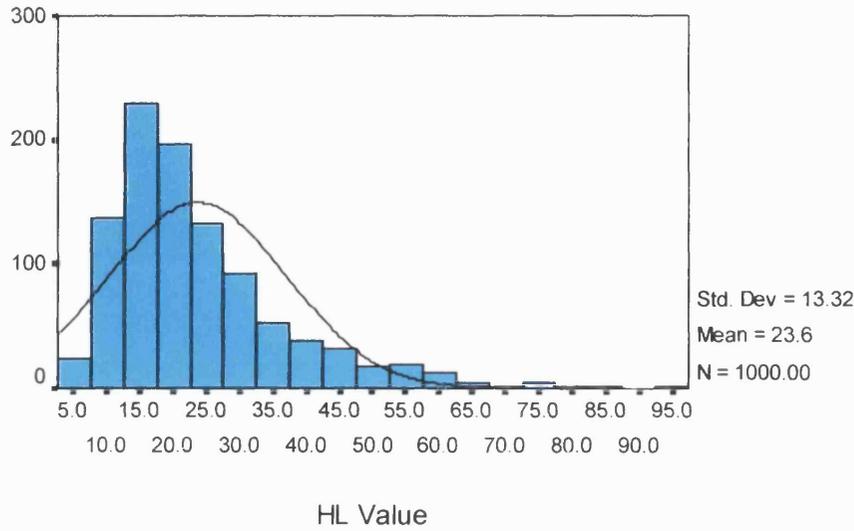
Model: DL = HCISS. First 1000 cases.



**Graph 4.6a**

### Histogram of 1000 Bootstrap Values

Model: DL=HCISS. First 3000 cases.

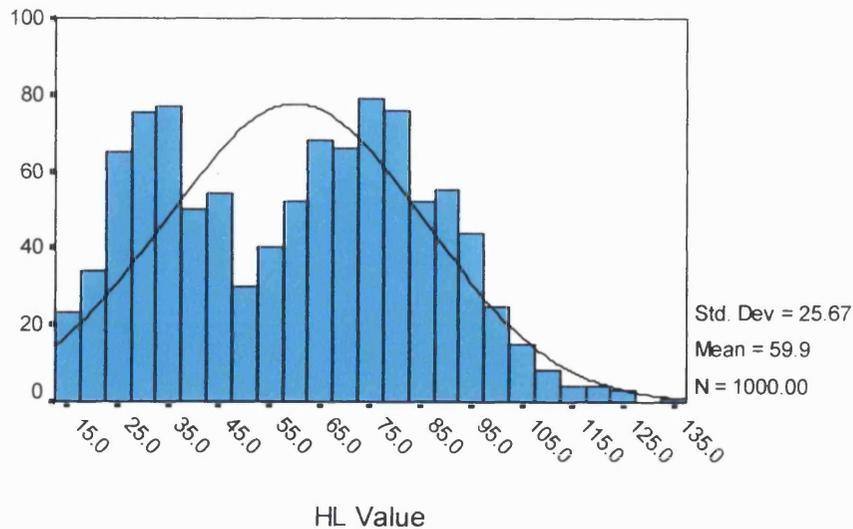


Second 1000 bootstrap results.

**Graph 4.6b**

### Histogram of 1000 Bootstrap Values

Model: DL=HCISS. First 5000 cases.

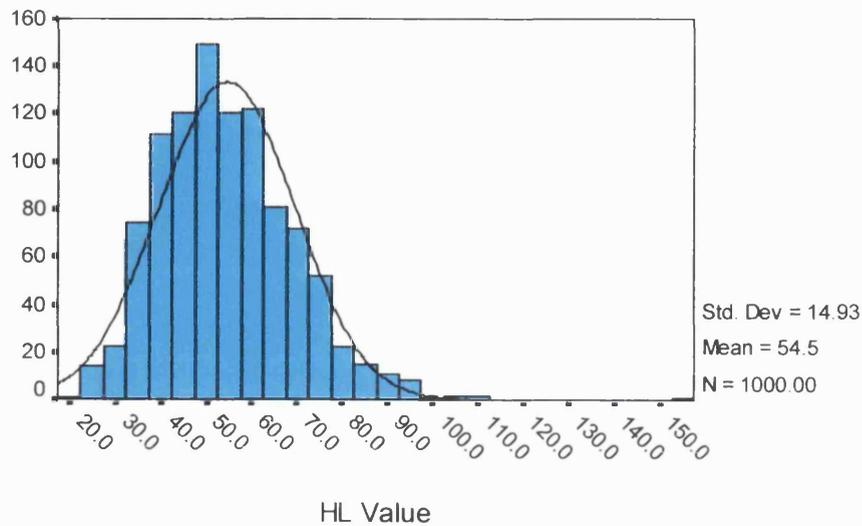


Second 1000 bootstrap results.

**Graph 4.6c**

## Histogram of 1000 Bootstrap Values.

Model: DL=HCISS. First 7000 cases.



**Graph 4.6d**

**Table 4.6**

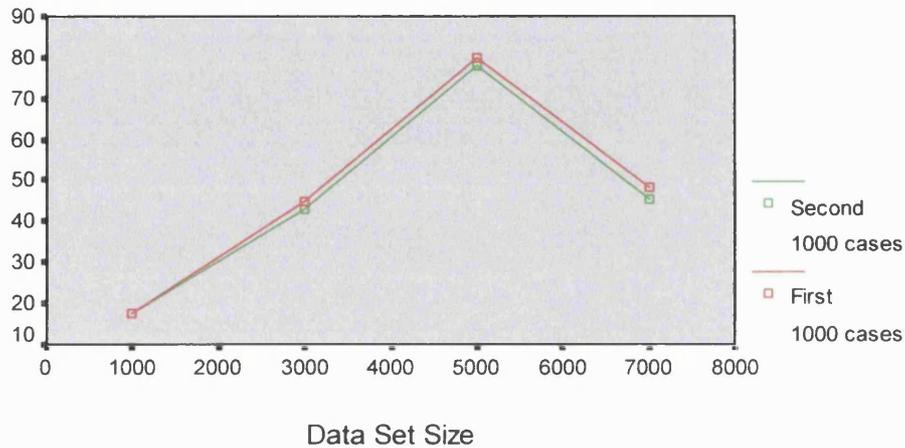
Model: DL=HCISS. HL 90% Confidence Interval.  
Percentile Method. Second 1000 bootstrap samples

Data Set Size	HL: LCI 95%	HL: UCI 95%	Bootstrap HL mean	Plug in Estimate (HL)	Coverage value
1000	5.11	22.67	12.30	6.8360	17.56
3000	8.66	51.94	23.60	12.9723	43.28
5000	21.50	99.52	59.90	71.0356	78.02
7000	33.71	79.12	54.50	50.0912	45.41

## 90% Coverage Values (Percentile Method)

### Plotted Against Data Set Size

#### First and Second 1000 Bootstrap Results



Model: DL=HCISS

### Graph 4.6e

Graphs 4.6a to 4.6d show similar distributions to graphs 4.4a to 4.4d. Graph 4.6e shows that the coverage values for the HL first 1000 bootstrap cases (HCISS model) is almost identical to the second 1000 cases.

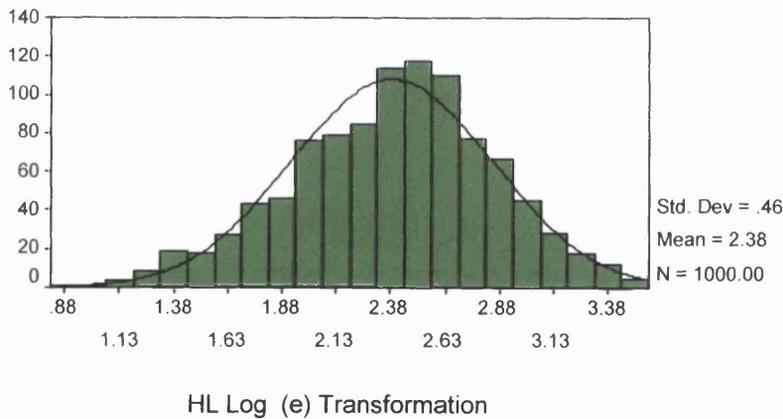
## Section 4.7: Log Transformation of HL Bootstrap Values.

**Model: DL= HCISS.**

The results from section 4.4 showed that the bootstrap sample for the HCISS model were poorly approximated to a normal distribution for all four data set values. A logarithmic transformation to base e (natural logarithm) and base 10 was made. The confidence interval points were defined as being the 50<sup>th</sup> and the 950<sup>th</sup> smallest transformed values. The HL confidence interval points (and the bootstrap mean) were obtained by using the HL value which corresponded to the 50<sup>th</sup> and 950<sup>th</sup> smallest transformed values. The 1000 bootstrap samples were those used in section 4.4

Histogram of 1000 Bootstrap HL Values  
With Log Transformation (Base e)

Model: DL=HCISS

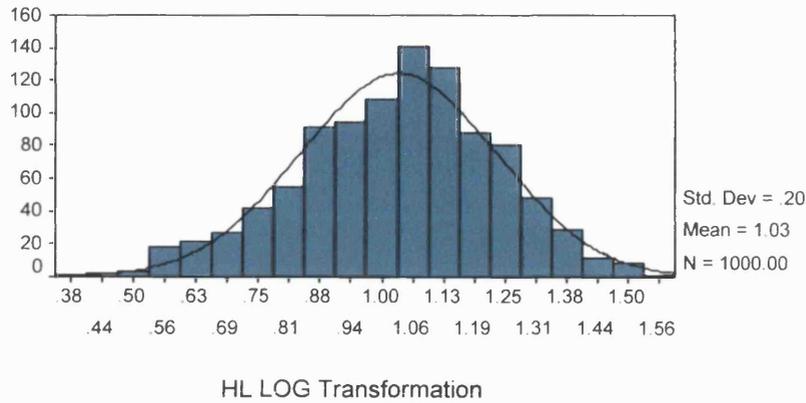


First 1000 cases

**Graph 4.7a**

Histogram of 1000 Bootstrap HL Values  
With Log Transformation (Base 10)

Model: DL=HCCI

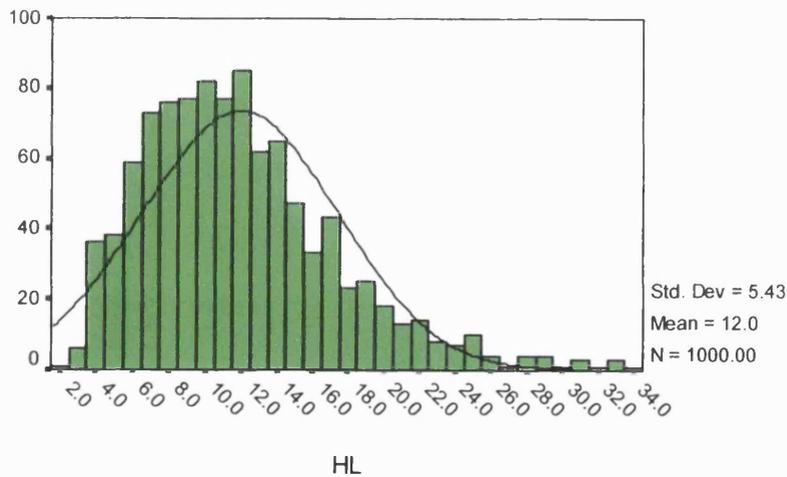


First 1000 cases

**Graph 4.7b**

Histogram of 1000 Bootstrap HL Values

Model: DL=HCCI



First 1000 Cases.

**Graph 4.4a**

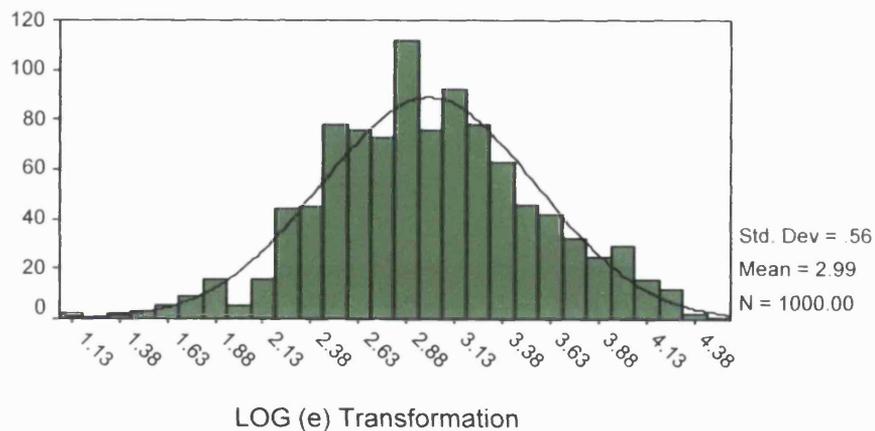
A comparison of graph 4.7a to graphs 4.7b shows that very similar distributions are produced using either log base e or log base 10.

Comparison of graph 4.7a to graph 4.4a above shows that a logarithmic transformation improves the distribution of the HL bootstrap samples.

**Data Set Value: 3000**

Histogram of 1000 Bootstrap HL Values  
With Log Transformation (Base e)

Model: DL=HCCISS

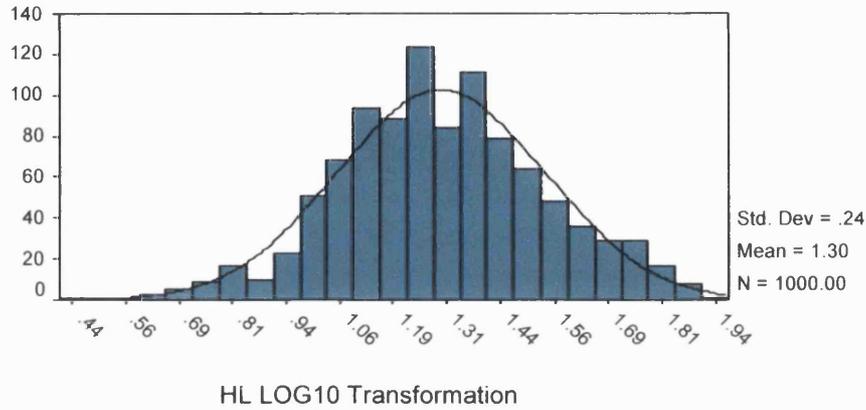


First 3000 cases

**Graph 4.7c**

Histogram of 1000 Bootstrap HL Values  
With Log Transformation (Base 10)

Model: DL=HCCISS

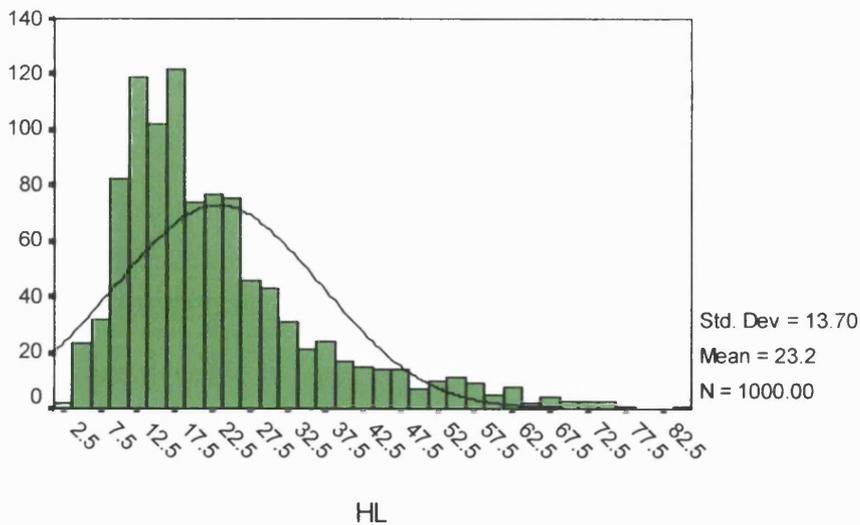


First 3000 cases

**Graph 4.7d**

Histogram of 1000 Bootstrap HL Values

Model: DL=HCCISS



First 3000 Cases.

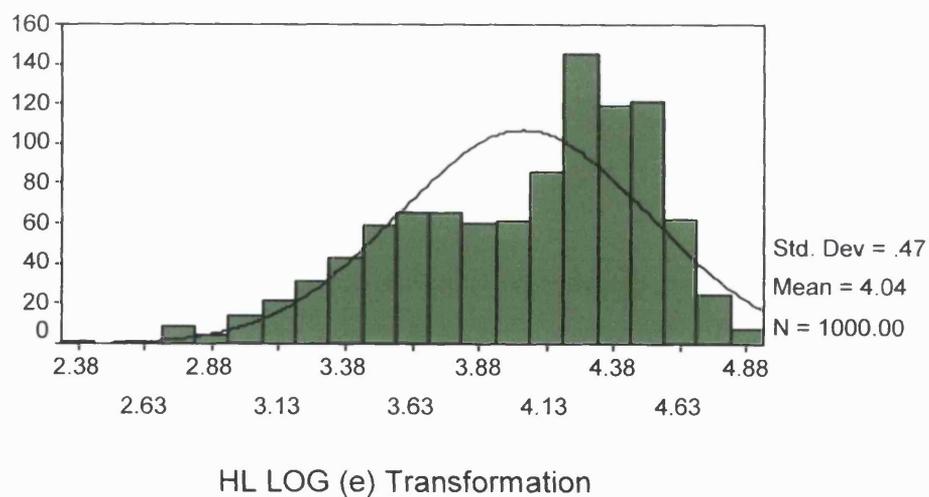
**Graph 4.4b**

A comparison of graph 4.7c to graphs 4.7d shows that very similar distributions are produced using either log base e or log base 10. Comparison of graph 4.7c to graph 4.4b above shows that a logarithmic transformation improves the distribution of the HL bootstrap samples.

**Data Set Value: 5000**

### Histogram of 1000 Bootstrap HL Values with Log Transformation (Base e)

Model: DL=HCISS

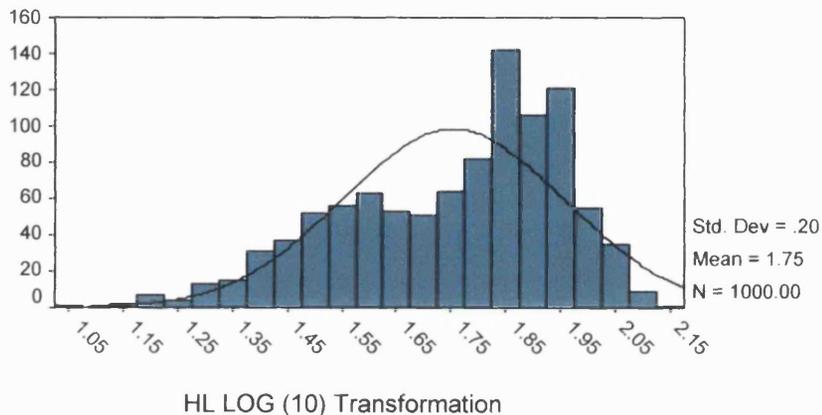


First 5000 cases

**Graph 4.7e**

### Histogram of 1000 Bootstrap HL Values with Log Transformation (Base 10)

Model: DL=HCCI

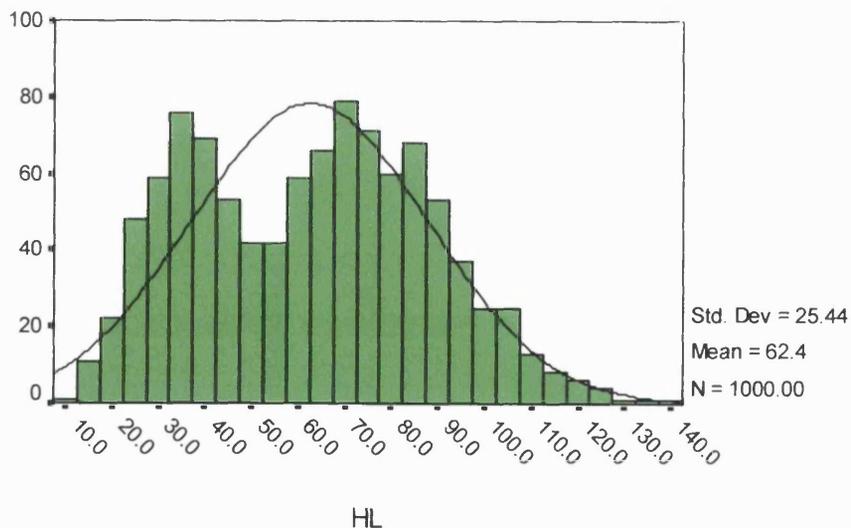


First 5000 cases

**Graph 4.7f**

### Histogram of 1000 Bootstrap HL Values

Model: DL=HCCI



First 5000 Cases.

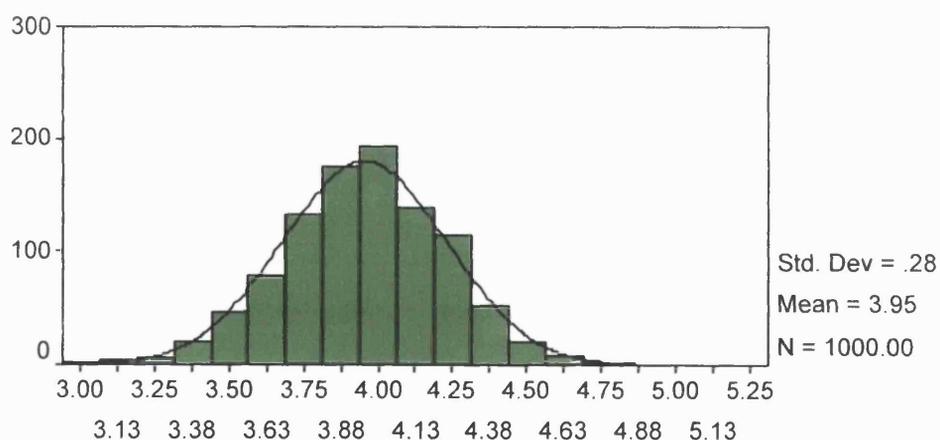
**Graph 4.4c**

A comparison of graph 4.7e to graphs 4.7f shows that very similar distributions are produced using either log base e or log base 10. Comparison of graph 4.7e to graph 4.4c above shows that a logarithmic transformation results in a slight improvement of the distribution of the HL bootstrap samples.

**Data Set Value: 7000**

### Histogram of 1000 Bootstrap HL Values with Log Transformation (Base e)

Model: DL=HCCIIS



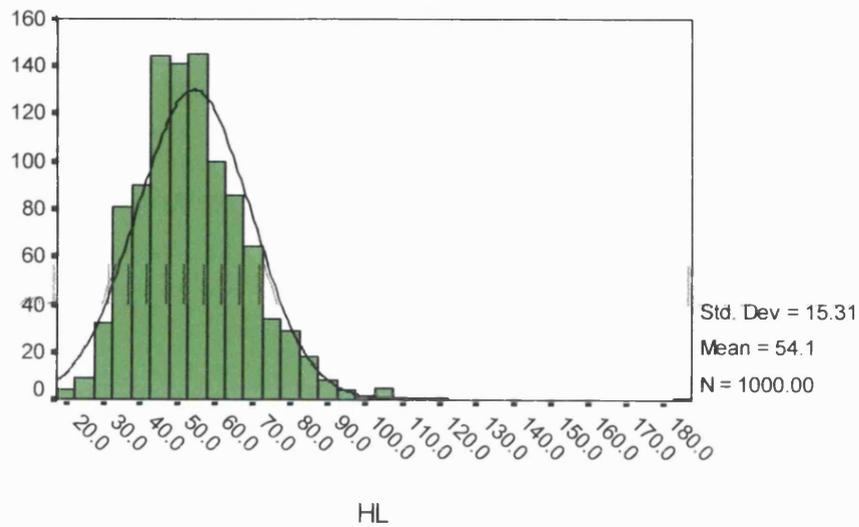
HL LOG (e) Transformation

First 7000 cases

**Graph 4.7g**

## Histogram of 1000 Bootstrap HL Values

Model: DL=HCISS



First 7000 Cases.

**Graph 4.4d**

Comparison of graph 4.7g to graph 4.4d above shows that a logarithmic transformation results in an improvement of the distribution of the HL bootstrap samples.

**Table 4.7a**

Model: DL=HCISS. Log e Transformation.

HL 90% Confidence Interval. (Percentile Method)

Data Set Size	HL: LCI Point	Bootstrap Mean	HL: UCI Point
1000	4.7	10.8	22.1
3000	8.4	19.9	53.2
5000	24.2	56.6	104.2
7000	32.8	51.7	81.0

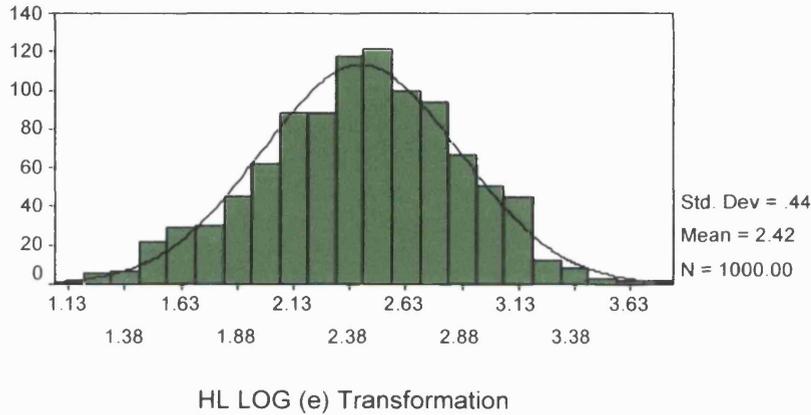
**Section 8. Log Transformation of HL Bootstrap Values.**

**Model: DL= HCISS+RTS.**

The results from section 4.5 showed that the bootstrap samples for the HCISS + RTS model were reasonably approximated to a normal distribution for all four data set values. To calculate the corresponding confidence intervals for the HCISS + RTS model a logarithmic transformation to base e (natural logarithm) was made. The confidence interval points were defined as being the 50<sup>th</sup> and the 950<sup>th</sup> smallest transformed values. The HL confidence interval points (and the bootstrap mean) were obtained by using the HL value which corresponded to the 50<sup>th</sup> and 950<sup>th</sup> smallest transformed values. The 1000 bootstrap samples were those used in section 4.5.

### Histogram of 1000 Bootstrap HL Values with Log Transformation (Base e)

Model: DL=HCISS+RTS

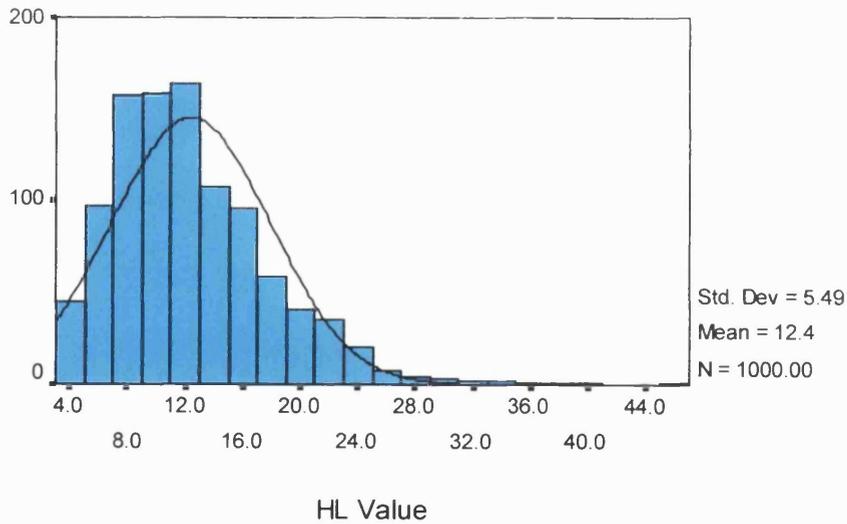


First 1000 cases

**Graph 4.8a**

### Histogram of 1000 Bootstrap HL Values

Model: HCISS+RTS



First 1000 Cases.

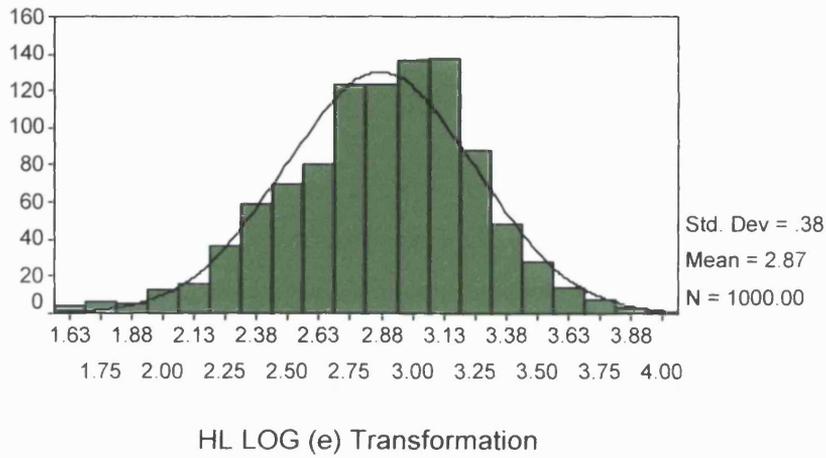
**Graph 4.5a**

Comparison of graph 4.8a to graph 4.5a above shows that a logarithmic transformation results in an improvement of the distribution of the HL bootstrap samples.

**Data Set Value: 3000**

### Histogram of 1000 Bootstrap HL Values with Log Transformation (Base e)

Model: DL=HCCISS+RTS

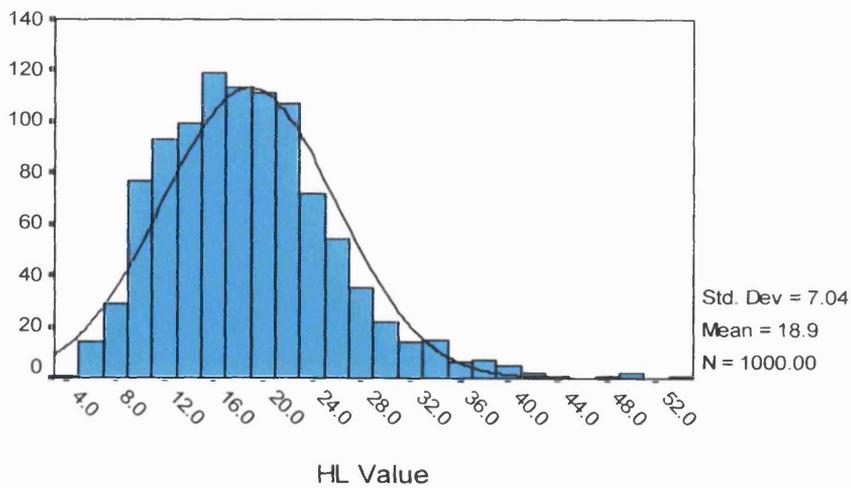


First 3000 cases

**Graph 4.8b**

### Histogram of 1000 Bootstrap HL Values

Mode: HCCISS+RTS



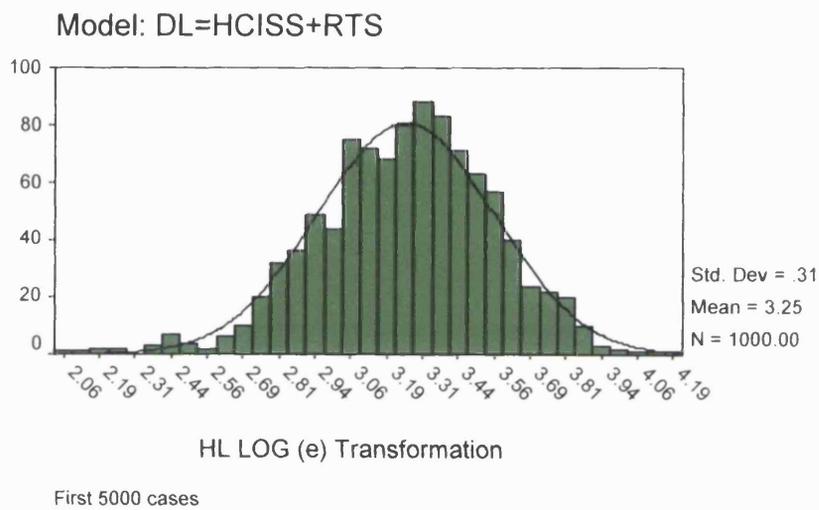
First 3000 Cases

**Graph 4.5b**

Comparison of graph 4.8b to graph 4.5b above shows that a logarithmic transformation results in an improvement of the distribution of the HL bootstrap samples.

**Data Set Value: 5000**

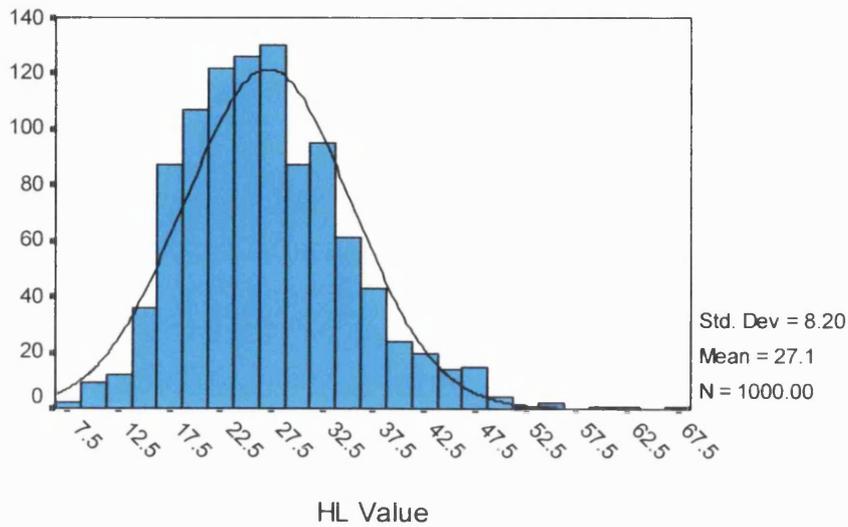
**Histogram of 1000 Bootstrap HL Values  
with Log Transformation (Base e)**



**Graph 4.8c**

## Histogram of 1000 Bootstrap HL Values

Model: HClSS+RTS



First 5000 Cases.

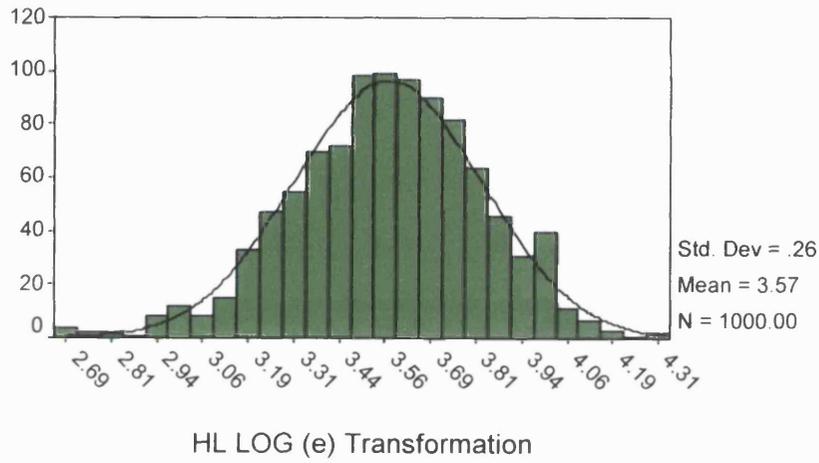
### Graph 4.5c

Comparison of graph 4.8c to graph 4.5c above shows that a logarithmic transformation results in an improvement of the distribution of the HL bootstrap samples.

**Data Set Value: 7000**

### Histogram of 1000 Bootstrap HL Values with Log Transformation (Base e)

Model: DL=HCCISS+RTS

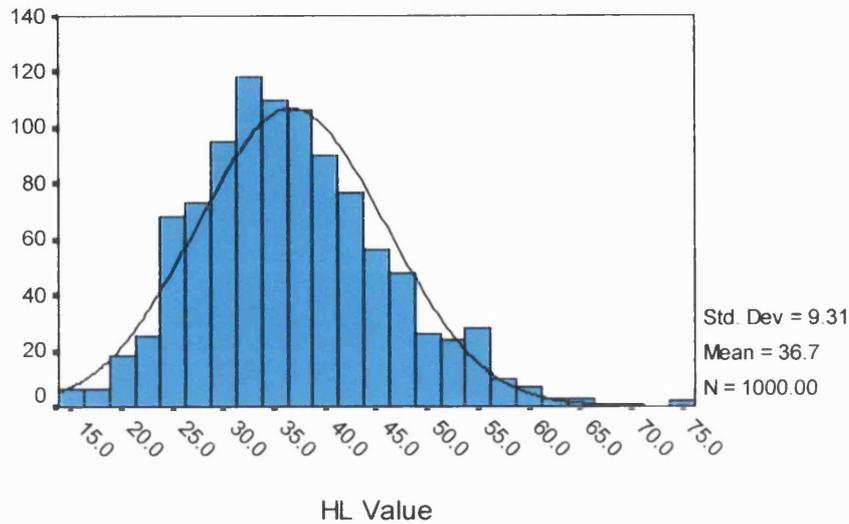


First 7000 cases

**Graph 4.8d**

## Histogram of 1000 Bootstrap HL Values

Model: HCISS+RTS



First 7000 Cases.

### Graph 4.5d

Comparison of graph 4.8d to graph 4.5d above shows that a logarithmic transformation results in no improvement in the distribution of the HL bootstrap samples.

### Table 4.7a

Model: DL=HCISS. Log e Transformation.

HL 90% Confidence Interval. (Percentile Method)

Data Set Size	HL: LCI Point	Bootstrap Mean	HL: UCI Point
1000	4.7	10.8	22.1
3000	8.4	19.9	53.2
5000	24.2	56.6	104.2
7000	32.8	51.7	81.0

**Table 4.8a**

Model: DL=HCISS + RTS. Log e Transformation

HL 90% Confidence Interval. (Percentile Method)

<u>Data Set Size</u>	<u>HL: LCI Point</u>	<u>Bootstrap Mean</u>	<u>HL: UCI Point</u>
1000	5.1	11.2	22.3
3000	9.2	17.6	31.4
5000	15.7	25.7	42.2
7000	23.1	35.4	54.0

Table 4.7a and table 4.8a shows the 10% confidence intervals for the HCISS model and the HCISS + RTS model following logarithmic transformation with base e. A comparison of the two tables shows that despite the logarithmic transformation the confidence intervals are unable to distinguish between the two models.

## **Section 5: Discussion**

The results of this study demonstrate that the bootstrap percentile method produced confidence intervals which can be applied to the Copas goodness of fit test. This conclusion is based upon three facts. Firstly, that histograms of Copas bootstrap samples approach a normal distribution over a range of data set sizes. Secondly, the Copas confidence intervals were able to distinguish between the HCISS model and the HCISS + RTS model. Thirdly, the small difference between the bootstrap mean and the plug in estimate. In contrast the HL bootstrap percentile method produced confidence intervals which were unable to distinguish between the HCISS model and the HCISS + RTS model. This was due to the fact that the histograms of HL bootstrap samples had distributions which were frequently too wide and which did not approach normal distributions. Although the distributions were improved with logarithmic transformations the confidence intervals were still found to be too large and were not able to distinguish between the two models. The Copas test was included in the study because simulation studies in chapter 7 suggests that the Copas test may be a viable alternative to the HL test. The HL test was included because it is the standard test used in assessing the calibration aspect of goodness of fit in logistic regression in the context of trauma scoring. The results of this study confirm the findings of Efron et al (1998) that the bootstrap works well for summation statistics like the Copas test but less well for more complex statistics like the Hosmer Lemeshow test. No previous studies have been identified which have evaluated bootstrap confidence intervals for either the Copas test or the HL test.

The percentile method was used in preference to the other bootstrap methods of calculating confidence intervals mainly because of its simplicity which makes programming more straightforward. For simple summation statistics it can produce accurate coverage intervals. The percentile method is both logical and also intuitively appealing to non-statisticians.

The Copas bootstrap mean values were close to their plug in estimates which suggests that no adjustment by using the BC method was necessary for the bootstrap mean. The BC method as mentioned in the introduction readjusts the bootstrap mean by adding the difference in value between the bootstrap mean from the bootstrap median (the 50 percentile point).

Other methods of bootstrapping confidence intervals were not evaluated because of their complexity. Many of these bootstrap methods are not readily available in computer software packages. Previous work as discussed in the earlier part of this chapter has indicated that these other methods of calculating confidence intervals may not be superior to the percentile method when applied to simple summation statistics such as the Copas test. The poor sampling distribution seen with the HL statistic suggests that other methods of bootstrapping may also produce inaccurate confidence intervals.

In summary the results from this study have shown that the percentile bootstrap method can generate accurate confidence intervals for the Copas test but not for the HL statistic.

## APPENDIX 1

### Program A for bootstrapping the Copas statistic in Logistic Regression.

Note that in program A only 100 output data sets for the mean Copas value (*sum1* through to *sum100*) have been shown in order to save space. To output 1000 bootstrap mean Copas results then *sum101* through to *sum1000* will need to be added to the program shown. The number of bootstrap samples is controlled by `n = .....` Line A. The size of the data set from which sampling is set is determined by `j > 7000` Line B and by `i > 7000` Line C.

Note the syntax; LINE A, LINE B and LINE C should be deleted before using the program.

The *final* data set was exported and saved as an excel file. The *stat* and *Copas* variable columns were copied and then pasted into program B, immediately below the *datalines* statement; line of x's.

```
data test;
  input ds hciss regcs resbp rerr rts num ;
datalines;
.00 26.00 3.00 4.00 4.00 11.00 1.00
1.00 5.00 .00 4.00 4.00 8.00 2.00
1.00 13.00 4.00 4.00 4.00 12.00 3.00
.00 25.00 .00 4.00 .00 4.00 4.00
1.00 16.00 4.00 4.00 4.00 12.00 5.00
1.00 10.00 4.00 4.00 4.00 12.00 6.00
.00 43.00 3.00 4.00 4.00 11.00 7.00
1.00 9.00 4.00 4.00 4.00 12.00 8.00
1.00 9.00 4.00 4.00 4.00 12.00 9.00
1.00 10.00 4.00 4.00 4.00 12.00 10.00
.00 17.00 .00 .00 .00 .00 11.00
;
run;

%macro repeat (n=100);
%local i;
%do i=1 %to &n;
data test&i;
  set test point=selection nobs=n;
  k=&i*12345;
selection=int(ranuni(k)*n)+1;
j+1;
if j >7000 then stop;
run ;
LINE A
LINE B
```

```

*close listing destination. i.e. output window;
ods listing close;
proc logistic data=work.test&i des;
  model ds = hciss / selection = backward;
  output out=lout&i p=pred;
run;
data work.lout&i;
  set lout&i;
  copas = (ds-pred)**2;
run;

*re-open listing destination;
ods listing;
proc means data=work.lout&i;
var copas;
output out=sum&i;
  run;
  %end;
%mend;

options symbolgen mprint mlogic;
%repeat;

data final;
  set sum1 sum2 sum3 sum4 sum5 sum6 sum7 sum8 sum9 sum10
  sum11 sum12 sum13 sum14 sum15 sum16 sum17 sum18 sum19 sum20
  sum21 sum22 sum23 sum24 sum25 sum26 sum27 sum28 sum29 sum30
  sum31 sum32 sum33 sum34 sum35 sum36 sum37 sum38 sum39 sum40
  sum41 sum42 sum43 sum44 sum45 sum46 sum47 sum48 sum49 sum50
  sum51 sum52 sum53 sum54 sum55 sum56 sum57 sum58 sum59 sum60
  sum61 sum62 sum63 sum64 sum65 sum66 sum67 sum68 sum69 sum70
  sum71 sum72 sum73 sum74 sum75 sum76 sum77 sum78 sum79 sum80
  sum81 sum82 sum83 sum84 sum85 sum86 sum87 sum88 sum89 sum90
  sum91 sum92 sum93 sum94 sum95 sum96 sum97 sum98 sum99 sum100;
run;

data test1;
selection=int(ranuni(12345)*n)+1;
set test point=selection nobs=n;
i+1;
  if i >7000 then stop;           LINE C
  drop i;
run;

```

## APPENDIX 2

### Program B for bootstrapping the Copas statistic in Logistic Regression.

This program was written so as to strip out the four unwanted statistics:- N, Max, Min and STD using an *If then delete* statement. The mean Copas result (written as *copas1* in the program) was then converted into the actual Copas value for the bootstrap data set by multiplying the *copas1* value by the data set size value. The Copas variable column in the final data set represents the Copas value for each bootstrap sample.

```
data final;
  input stat $ copas1;
  if stat = 'N' then gof=1;
  if stat = 'MAX' then gof=2;
  if stat = 'MIN' then gof=3;
  if stat = 'MEAN' then gof=4;
  if stat = 'STD' then gof=5;
  if gof ne 4 then delete;
  datalines;
xxxxxxx
;
run;

  data work.final;
  set final;
  copas = copas1*7000;
run;
```

# **CHAPTER 11**

## **A STUDY TO EVALUATE TRAUMA**

### **MODELS USING THE COPAS**

#### **GOODNESS OF FIT TEST**

### **CONTENTS**

	<b>Page Number</b>
<b>Section 1: Introduction</b>	<b>301</b>
<b>Section 2: Aims</b>	<b>301</b>
<b>Section 3: Methodology</b>	<b>302</b>
<b>Section 4: Results</b>	<b>304</b>
<b>Section 5: Discussion</b>	<b>306</b>

## **Section 1: Introduction**

Previous chapters have demonstrated that the Copas goodness of fit test may be a useful alternative to the Hosmer Lemeshow test. Simulation studies in chapter 10 have demonstrated that bootstrap confidence intervals using the percentile method can be generated for the Copas statistic. Chapter 4 of this thesis (basic modelling) showed that there was a large difference in the Hosmer Lemeshow value when comparing the HCISS model (HL = 52.8) to the HCISS + coded GCS model (HL = 21.53). Models with additional predictors such as coded RR (respiratory rate), coded SBP (systolic blood pressure) and age only resulted in a modest reduction in the Hosmer Lemeshow value. The statistical significance of these finding without the use of confidence intervals is difficult to assess.

## **Section 2: Aims**

The aim of the study was to evaluate nine trauma scoring models using the Copas goodness of fit test together with bootstrap generated confidence intervals.

## **Section 3: Methodology**

**Section 3.1** The revised USC data set was used.

The entire data set of 7069 cases was used in this study.

### **Section 3.2**

Logistic regression was performed using SAS version 8.

Method of model selection chosen was:- Backward Selection.

The descending option was used so that predicted probabilities were calculated for survival rather than death.

### **Section 3.3**

Nine models were used. ('C' represents the RTS coded version)

1. HCISS
2. HCISS CGCS
3. HCISS CGCS CSBP
4. HCISS CGCS CSBP CRR
5. HCISS CGCS CSBP CRR Age
6. HCISS RTS (unweighted)
7. HCISS RTS (unweighted) AGE
8. HCISS CSBP
9. HCISS CRR

### **Section 3.4**

The method for calculating the Copas statistic is:-

$$\text{Copas} = \sum (ds - \text{pred prob})^2$$

ds = outcome variable: – death or survival

pred prob = predicted probabilities

### **Section 3.5 Bootstrap Generated Confidence Intervals**

Exactly the same method was use to generate the bootstrap confidence intervals as described in chapter 10 (percentile method). 1000 bootstrap samples were generated. The 90% confidence interval was calculated by determining the 50<sup>th</sup> and 950<sup>th</sup> largest bootstrap Copas value for each bootstrap sample.

The Copas value for the actual data set (plug in estimate) was calculated for the nine different models using the same method as that used to generate the bootstrap samples i.e. SAS version 8, model selection:- backward likelihood ratio. The bootstrap mean was calculated as the  $\sum$ bootstrap values/1000. The difference between the plug in estimate and the bootstrap mean was defined as the bias for that parameter using the method previously described in chapter 10 (Efron et al: 1998, page 138).

Bias corrected Copas (BCC) = Copas (plug in value) – Bias

Bias = Bootstrap mean Copas – Copas (plug in value)

### **Section 3.6**

The distribution of the bootstrap samples was evaluated for all nine models.

## Section 4: Results

The distribution of the 1000 bootstrap samples was evaluated. All nine models produced similar distributions which were well approximated to a normal distribution.

**Table 1**

<b>Model</b>	<b>Bootstrap Mean Copas</b>	<b>Plug in Copas</b>
1. HCISS	483.1	484.4
2. HCISS CGCS	309.3	310.0
3. HCISS CGCS CSBP	277.2	277.9
4. HCISS CGCS CSBP CRR	274.8	275.7
5. HCISS CGCS CSBP CRR Age	264.6	265.5
6. HCISS RTS	278.7	279.3
7. HCISS RTS AGE	267.4	268.0
8. HCISS CSBP	348.7	349.3
9. HCISS CRR	322.5	323.1

Table 1 shows the bootstrap mean Copas result and the plug in estimate for all models.

**Table 2**

<b>Model</b>	<b>Bias</b>	<b>Bias Corrected Copas (BCC)</b>
1. HCISS	-1.30	485.70
2. HCISS CGCS	-0.70	310.70
3. HCISS CGCS CSBP	-0.70	278.60
4. HCISS CGCS CSBP CRR	-0.90	276.60
5. HCISS CGCS CSBP CRR Age	-0.90	266.40
6. HCISS RTS	-0.60	279.90
7. HCISS RTS AGE	-0.60	268.60
8. HCISS CSBP	-0.60	349.90
9. HCISS CRR	-0.60	323.70

Table 2 shows that the bias was small for all models.

**Table 3**

<b>Model</b>	<b>5% LCI Value</b>	<b>BCC</b>	<b>95% UCI Value</b>
1. HCISS	457.5	485.7	507.8
2. HCISS CGCS	288.2	310.7	331.1
3. HCISS CGCS CSBP	256.8	278.6	298.0
4. HCISS CGCS CSBP CRR	254.9	276.6	296.5
5. HCISS CGCS CSBP CRR Age	244.5	266.4	286.0
6. HCISS RTS	258.1	279.9	301.1
7. HCISS RTS AGE	246.9	268.6	289.1
8. HCISS CSBP	326.0	349.9	371.8
9. HCISS CRR	298.4	323.7	346.4

Table 3 shows a significant improvement in model fit when a single coded RTS variable is added to the HCISS variable.

## **Section 5: Discussion**

The results of this study firstly demonstrate that the addition of a single coded physiological variable to the HCISS model results in a significant reduction in the bias corrected Copas value (BCC) when judged by the 90% confidence intervals. The greatest reduction in the BCC value with respect to the aforementioned models was seen with model 2 (HCISS + coded GCS).

The smallest bias corrected Copas value (BCC) was seen with model 5 (HCISS + coded GCS + coded SBP + coded RR + Age). The bootstrap confidence intervals however indicate that it was not significantly superior to models 3 (HCISS + coded GCS + coded SBP) and 4 (HCISS + coded GCS + coded SBP + coded RR). The BCC for model 4 was marginally smaller than model 6. The 90% confidence intervals however indicate that the difference was not statistically significant. In other words splitting the RTS into its component parts resulted in a slight improvement in model fit using the Copas test, but it was not significant using bootstrap generated confidence intervals. The addition of the age variable to both of the latter two models (models 5 and model 7) resulted in a slight reduction in the BCC value. This again was not statistically significant.

Previous work by Hannan et al (1999) using the HL statistic suggested that splitting the revised trauma score (RTS) into its component values improved the goodness of fit of the model. As mentioned previously in chapter 7 the HL test and the Copas test measure different aspects of model fit. Despite this limitation the results of this study suggest that the slight improvement in splitting

the RTS into its component parts may not be statistically significant when using the Copas test to assess model fit

# **CHAPTER 12**

## **CONCLUSIONS**

## **SUMMARY OF MAIN FINDINGS AND CONCLUSIONS**

Chapter 2 provides a detailed review of the development of TRISS (Champion, 1990a), ASCOT (Champion, 1990b), ICD-9 based models (Osler, 1996), NISS (Osler, 1997) and mAP (Sacco, 1999). A recent comprehensive study of nine Abbreviated Injury Scale and ICD-9 based models has been performed by Meredith et al (2002). These authors found that ICISS had the best discrimination using ROC analysis. The Anatomical Profile however had the best calibration using the Hosmer Lemeshow test.

Chapter 3 provides a detailed description of the method which was used to prepare the USC data set. Only cases which had a complete set of values for the variables chosen were included. A complete data set made direct comparisons between different models easier due to the fact that many goodness of fit statistics increase as the data set size becomes larger.

Chapter 4 was a comparison of nine models using the following variables; HCISS, components of the RTS and age. Dividing the RTS into its component parts resulted in improved calibration for some models but not all. A recent paper by Osler et al (2002) found that the log odds of death was not linear when plotted against ISS, thus violating one of the fundamental assumptions of logistic regression (Harrell, 1996). These authors calculated the odds using the actual number of deaths and survivors for each ISS value rather than the probabilities generated by the fitted model. The results

from chapter 4 showed that using the revised USC data set the log odds for survival was linear when plotted against HCISS.

Chapter 5 was a study which compared the three different methods of calculating the Hosmer Lemeshow goodness of fit statistic over a range of data set values. No clear trends from the studies were identified with regard to which method gives the least erratic results.

Chapter 6 was another series of studies which addressed one of the limitations of the Hosmer Lemeshow statistic, namely its over sensitivity to change in the covariate pattern of a predictor variable. Reducing the covariate pattern of a variable by recoding can have an appreciable impact on the HL value and may result in over prediction of the model's goodness of fit. The effect was found to be variable dependent and was most pronounced when the predictor variable was reduced to a small number of covariate groupings. The effect was found to be less with models with more than one predictor variable. All three methods of calculating the HL test were found to be sensitive to changes in covariate grouping.

Chapter 7 was a comparative study of six goodness of fit tests using two models (HCISS and HCISS + RTS) over a range of data set values. The deviance statistic (*proc genmod* method) was able to distinguish between the two models for all data set values selected. The Copas test, the Brier test and the unweighted standardized residuals test (USR) were all able to distinguish between the above two models in 92.9% of cases for the smaller

data sets (50-1000) and in 100% of cases for the larger data sets (6050-7000). The HL test was able to correctly distinguish between the two models in 91.2% of cases for the smaller data set (51-1000) and in 100% of cases for the larger data set (6050-7000). The results from this study suggest that the deviance statistic (*proc genmod* method) had the largest power to detect a difference between the HCISS model and the model with HCISS + RTS. The Copas, Brier and USR tests all showed the same power to detect a difference between the two models. All three tests however performed poorly with very small data set sizes. The deviance statistic (-2LL method) although having the greatest power in this study does have two drawbacks. Firstly, there are several ways that the test statistic can be calculated. Secondly, the log-likelihood statistic is also used for model development in the likelihood ratio methods. The Brier test produces results which are very small in magnitude making comparisons between models more difficult. The unweighted standardized residuals test used in this study has not been fully evaluated. The Copas test is therefore the most potential viable alternative to the HL test even though the two tests measure different aspects of goodness of fit.

Chapter 8 was a study to evaluate data splitting as a method of validating a logistic model. Both parts of the study highlighted that data splitting can be an unreliable method of correcting for over-fitting in the logistic model.

Chapter 9 evaluated the effect of increasing the sampling pool on the cross validation value on two trauma scoring models (HCISS and HCISS + RTS). The variability of the cross validation Copas

results which occurred by increasing the sampling pool suggests that the cross validation method is not a reliable way of correcting for over-fitting in trauma scoring modelling. The large number of samples generated in these simulation studies makes the variability unlikely to be due to the sampling size. A control model showed that most of the variability could be explained by the increase in the data set size rather than due to variability of the sampling method.

Chapter 10 was an evaluation of the bootstrap percentile method using the Copas and Hosmer Lemeshow goodness of fit tests. The results from this study demonstrated that the bootstrap percentile method produced confidence intervals which can be applied to the Copas goodness of fit test. This conclusion was based mainly upon two facts. Firstly, histograms of Copas bootstrap samples approached a normal distribution when evaluated against a range of data set sizes. Secondly, the Copas confidence intervals were able to distinguish between the HCISS model and the HCISS + RTS model. In contrast the HL bootstrap percentile method produced confidence intervals which were unable to distinguish between the HCISS model and the HCISS + RTS model. This was due to the fact that the histograms of HL bootstrap samples had distributions which were frequently too wide and which did not approach a normal distribution. Although the distributions were improved with logarithmic transformations the confidence intervals were still found to be too wide and were not able to distinguish between the two models.

Chapter 11 re-evaluated several models previously examined in chapter 4. In particular the effect of splitting the RTS into its component parts was examined using the Copas test with bootstrap confidence intervals. The results from this chapter showed that splitting the RTS into its component parts did result in a slight improvement in model fit but this was not statistically significant when evaluated using bootstrap confidence intervals. The addition of the age variable to both the HCISS + RTS model and HCISS + component RTS model resulted in a slight reduction in the bias corrected Copas value for both models. The difference between the latter two models was again not statistically significant using percentile bootstrap generated confidence intervals.

## SUMMARY

The results from this thesis has highlighted the potential value of the Copas test as an alternative to the Hosmer Lemeshow statistic. The accuracy of this statistic is increased by using bootstrap confidence intervals using the percentile method. Using the programs provided in this thesis the Copas test can be readily used by non-statisticians working in the field of trauma scoring model development. Although many other goodness of fit tests have been proposed, many of them are too complicated for routine use at the present time. Similarly, several other methods of calculating bootstrap confidence intervals have been described. Although many of them have theoretical advantages over the percentile method, the results from chapter 10 have shown that the latter works well for the Copas test. The percentile method is also

intuitively appealing and easy to implement using the program given in this thesis. By using the Copas test with percentile bootstrap generated confidence intervals a further evaluation of ICISS, NISS and mAP should now be possible.

## CHAPTER 13: REFERENCES

Alison PD. (1999)

Binary logit analysis In *Logistic Regression using the SAS system. Theory and Applications.*

SAS Institute Inc., Cary, NC, USA.

Altman DG., Royston. (2000)

What do we mean by validating a prognostic model?

*Stat Med.* 19: 453-473.

Al West T. (2000a)

A comparison of ISS and NISS in the context of physiology in a large national trauma registry.

American Association for the Surgery of Trauma: 60<sup>th</sup> Annual Scientific Meeting. October 2000.

Web Site. [aast.org/00abstracts/00absPoster\\_057.html](http://aast.org/00abstracts/00absPoster_057.html)

Al West T., Rivara FP., Cummings P et al. (2000)

Harborview assessment for risk of mortality: an improved measure of injury severity on the basis of ICD-9- CM.

*J Trauma.* 49(3): 530-540.

American Association for Automotive Medicine.

The abbreviated injury scale (AIS)-1990 revision.

Des Plaines, IL:1990.

Baker SP., O'Neil B., Haddon W et al. (1974)

The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care.

*J Trauma*. 14(3): 187-196.

Baker SP., O'Neil B. (1976)

The injury severity score: an update.

*J Trauma*. 16(11): 882-885

Balogh Z., Offner PJ., Moore EE et al. (2000)

NISS predicts post-injury multiple organ failure better than the ISS.

*J Trauma*. 48(4): 624-628.

Batchelor JS. (2000)

Adult pre-hospital scoring systems: a critical review.

*Trauma*. 2(4): 253-260.

Becalick DC., Coates TJ. (2001)

Comparison of artificial intelligence techniques with UKTRISS for estimating probability of survival after trauma.

*J Trauma*. 51(1): 123-133.

Bertolini G., D'Amico R., Nardi D., et al. (2000)

One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model.

*J Epidemiol and Biostat*. 5(4): 251-253.

Bourbeau R. (1993)

Analyse comparative de la mortalite violente dans les pays

developpees et dans quelques pays en developpement durant la periode 1985-1989.

*World Health Statistics Quarterly*. 46: 4-32.

Brenneman FD., Boulanger BR., McLellan BA et al. (1998)

Measuring Injury Severity: time for a change?

*J Trauma*. 44(4): 580-582.

Brier GW. (1950)

Verification of weather forecasts expressed in terms of probability.

*Monthly Weather Review*. 78: 1-3.

Carpenter J., Bithell J. (2000)

Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians.

*Stat Med*. 19: 1141-1164.

Cayten CG., Stahl WM., Murphy JG et al. (1991)

Limitations of the TRISS method for inter-hospital comparisons: a multi-hospital study.

*J Trauma*. 31(4): 471-482.

Champion HR., Sacco WJ., Hannan DS et al. (1980)

Assessment of injury severity: the Triage Index.

*Crit Care Med*. 8(4): 201-208.

Champion HR., Sacco WJ., Carnazzo AJ et al. (1981)

The Trauma Score.

*Crit Care Med*. 9(9): 672-676.

Champion HR., Sacco WJ., Hunt TK. (1983)  
Trauma severity scoring to predict mortality.  
*World J Surg.* 7(1): 4-11.

Champion HR., Sacco WJ., Copes WS et al. (1989)  
A revision of the trauma score.  
*J Trauma.* 29(5): 623-629.

Champion HR., Copes WS., Sacco WJ et al. (1990a)  
The Major Trauma Outcome Study: establishing national norms for  
trauma care.  
*J Trauma.* 30(11): 1356-1365.

Champion HR., Copes WS., Sacco WJ et al. (1990b)  
A New Characterisation of Injury Severity.  
*J Trauma.* 30(5): 539-546.

Champion HR., Copes WS., Sacco WJ et al. (1996)  
Improved predictions from a Severity Characterisation of Trauma  
(ASCOT) over Trauma and Injury Severity Score (TRISS): results  
of an independent evaluation.  
*J Trauma.* 40(1): 42-49.

Chernick MR. (1999)  
Confidence Sets and Hypothesis Testing In *Bootstrap methods. A  
Practitioner's Guide.*  
New York. John Wiley & Sons Inc.

Clark DE., Ryan LM. (1997)

Modelling injury outcomes using time-to-event methods.

*J Trauma*. 42(6): 1129-1134.

Committee on Medical Aspects of Automotive Safety. (1971)

Rating the severity of tissue damage: I. The Abbreviated Scale.

*JAMA*. 215(2): 277-280.

Copas JB. (1989)

Unweighted sum of squares test for proportions.

*Applied Statistics*. 38(1).71-80.

Copes WS., Champion HR., Sacco WJ et al. (1990)

Progress in characterising anatomic injury.

*J Trauma*. 30(10): 1200-1207.

Cox DR., Snell EJ. (1989)

*The analysis of Binary data*, 2<sup>nd</sup> edition.

Chapman and Hall. London.

Demetriades D., Chan LS., Velmahos G et al. (1998)

TRISS methodology in trauma: the need for alternatives.

*Br J Surgery*. 85(3): 379-384.

Demetriades D., Chan LS., Velmahos G et al. (2001)

TRISS methodology an inappropriate tool for comparing outcomes between trauma centers.

*J Am Coll Surg*. 193(3): 250-254.

Diaconis P., Efron. (1983)

Computer-intensive methods in statistics.

*Sci Amer.* 248: 116-130.

Efron B. (1979)

Bootstrap methods: another look at the jackknife.

*Ann Statist.* 7: 1-26.

Efron B. (1982)

The jackknife, the Bootstrap and other resampling plans.

Volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM. Philadelphia.

Efron B. (1983)

Estimating the error rate of a prediction rule: improvement on cross validation.

*J Am Stat Assoc.* 78: 316-331.

Efron B., Gong G. (1983)

A leisurely look at the bootstrap, the jackknife and cross-validation.

*Am Statist.* 37: 36-48.

Efron B., Tibshirani R. (1986)

Bootstrap methods for standard errors: confidence intervals and other measures of statistical accuracy.

*Statist Sci.* 1: 54-77.

Efron B., Tibshirani R. (1998)

*An introduction to the bootstrap.*

New York. Chapman & Hall. 1993. Reprint 1998

Farrington CP. (1996)

On assessing goodness of fit of generalized linear models to sparse data.

*J Royal Statist Soc, Series B.* 58(2): 349-360.

Geisser S. (1975)

The predictive sample reuse method with applications.

*J Am Statist Assoc.* 70: 320-328.

Gillott AR., Copes WS., Langan E et al. (1992)

TRISS unexpected survivors-a statistical phenomena or a clinical reality.

*J Trauma.* 33(5): 743-748.

Goldberg JL., Gelfand J., Levy PS. (1980)

An evaluation of the Illinois trauma registry.

*Med Care.* 18: 520-531.

Goldberg JL., Goldberg J., Levy PS et al. (1984)

Measuring the severity of injury: the validity of the revised estimated survival probability index.

*J Trauma.* 24(5): 420-427.

Grisoni E., Stallion A., Nance ML et al. (2001)

The New Injury Severity Score and the evaluation of paediatric trauma.

*J Trauma.* 50(6): 1106-1010.

Hall P. (1992)



[www.hems-london.org.uk](http://www.hems-london.org.uk)

Barts and The London   
NHS Trust

Department A&E and Prehospital Care  
Royal London Hospital, Whitechapel  
London E1 1BB

Tel: 020 7943 1303  
Fax: 020 7377 7014  
Pager: 07659 136746  
Email: [t.j.coats@mds.qmw.ac.uk](mailto:t.j.coats@mds.qmw.ac.uk)

**Mr Tim Coats MD FRCS FFAEM**  
Senior Lecturer A&E/Prehospital Care

*The bootstrap and Edgeworth Expansion.*

Springer-Verlag. New York.

Hannan EL., Mendeloff J., Farrell LS et al. (1995a)

Validation of TRISS and ASCOT using a non-MTOS trauma registry.

*J Trauma.* 38(1): 83-88.

Hannan EL., Mendeloff J., Farrell LS et al. (1995b)

Multivariate models for predicting survival of patients for trauma from low falls: the impact of gender and pre-existing conditions.

*J Trauma.* 38(5): 697-704.

Hannan EL., Farrell LS., Gorthy SH et al. (1999)

Predictors of Mortality in Adult Patients with Blunt Injuries in New York State: A comparison of the Trauma and Injury Score (TRISS) and the International Classification of Disease, Ninth Revision-based Injury Severity Score (ICISS).

*J Trauma.* 47(1): 8-14.

Harrell Jr FE., Lee KL., DB Mark. (1996)

Tutorial in Biostatistics. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.

*Stat Med.* 15: 361-387.

Harrell Jr FE. (2002)

*Regression modelling strategies.*

Springer-Verlag. New York.

Hill DA., Lennox AF., Neil MJ et al. (1992)

Evaluation of TRISS as a means of selecting trauma deaths for clinical peer review.

*Aust N Z J Sur.* 62: 204-208.

Hjorth JSU. (1994)

*Computer Intensive Statistical Methods.*

London. Chapman & Hall.

Hollis S., Yates DW., Woodford M et al. (1995)

Standardized comparisons of performance indicators in trauma: a new approach to case-mix variation.

*J Trauma.* 38(5): 763-766.

Hosmer DW., Lemeshow S. (1980)

A goodness of fit test for the multiple logistic regression model.

*Communications in Statistic- THEOR. METH.* A(9)10 1043-1069.

Hosmer DW., Lemeshow S., Klar J. (1988)

Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small.

*Biom J.* 30: 911-924.

Hosmer DW., Lemeshow S. (1989)

Assessing the fit of the model. In *Applied Logistic Regression.*

John Wiley & Sons. New York, 140-145.

Hosmer DW., Hosmer T., Le Cessie S et al. (1997)

A comparison of Goodness-Of- Fit tests for the logistic regression model.

*Stat Med.* 16: 965-980.

Hou LF., Tsai MC. (1996)

Comparison between TRISS and ASCOT methods in Tainan area.

*Kao Hsiung I Hsueh Ko Hsueh Tsa Chih.* 12(12): 691-698.

Husum H., Strada G. (2002)

Injury Severity Score versus New Injury Severity Score for penetrating injuries.

*Prehosp Disaster Med.* 17(1): 27-32.

Karmy-Jones R., Copes WS., Champion HR et al. (1992)

Results of a multi-institutional outcome assessment-Results of a structured peer review of TRISS-designated unexpected outcomes.

*J Trauma.* 32(2): 196-203.

Kirkpatrick JR., Youmans RL. (1971)

An aide in the evaluation of injury victims.

*J Trauma.* 11(8): 711-714.

Kuss O. (2002)

Global goodness-of- fit tests in logistic regression with sparse data.

*Stat Med.* 21: 3789-3801.

Levy PS., Mullner R., Goldberg J et al. (1978)

The estimated survival probability index of trauma severity.

*Health Ser Res.* 13: 28-35.

Levy PS., Goldberg J and Rothrock J. (1982)  
The revised estimated survival probability index of trauma severity.  
*Pub Health Rep.* 97(5): 452-459.

Lane PL., Doig G., Stewart TC et al. (1997)  
Trauma outcome analysis and the development of regional norms.  
*Accid Anal Prev.* 29(1): 53-56.

le Cessie S., van Houwelingen JC. (1991)  
A goodness-of-fit test for binary regression models based on  
smoothing methods.  
*Biometrics.* 47: 1267-1282.

le Cessie S., van Houwelingen JC. (1995)  
Testing the fit of a regression model via scores tests in random  
effects models.  
*Biometrika.* 51: 600-614.

Logistic Regression Analysis Examples.  
In *SPSS Regression Models 9.0.* (1999)  
SPSS Inc. Chicago, IL. 35-62.

MacKenzie EJ., Steinwachs DM., Shankar B. (1989)  
Classifying trauma severity based on hospital discharge diagnoses.  
Validation of an ICD-9CM to AIS-85 conversion table.  
*Med Care.* 27(4): 412-422.

Markle J., Cayten CG., Byrne DW et al. (1992)  
Comparison between TRISS and ASCOT methods in controlling

for injury severity.

*J Trauma.* 33(2): 326-331.

McCarthy PJ. (1976)

The use of balanced half-sample replication in cross validation studies.

*J Am Stat Ass.* 71: 596-604.

McCullagh P. (1985)

On the asymptotic distribution of Pearson's statistic in linear exponential-family models.

*International Statistical Review.* 53(1): 61-67.

Milham FH., Malone M., Blansfield J et al. (1995)

Predictive accuracy of the TRISS survival statistic is improved by a modification that includes admission pH.

*Arch Surg.* 130(3): 307-311.

Milzman DP., Boulanger BR., Rodriguez A et al. (1992)

Pre-existing disease in trauma patients: a predictor of fate independent of age and injury severity score.

*J Trauma.* 32(2). 236-243.

Meredith JW., Evans G., Kilgo PD et al. (2002)

A comparison of the abilities of nine scoring algorithms in predicting mortality.

*J Trauma.* 53(44): 621-628.

Napolitano LM., Ferrer T., McCarter RJ et al. (2000)

Systemic inflammatory response syndrome score at admission independently predicts mortality and length of stay in trauma patients.

*J Trauma.* 49(4): 647-653.

Nagelkerke NJ. (1991)

A note on the general definition of the coefficient of determination.

*Biometrika.* 78: 691-692.

Norris R., Woods R., Harbrecht B et al. (2002)

TRISS unexpected survivors: an outdated standard?

*J Trauma.* 52(2): 229-234.

Ogawa M., Sugimoto T. (1974)

Rating the severity of the injured by ambulance attendants.

*J Trauma.* 14: 934-937.

Osius G., Rojek D. (1992)

Normal goodness of fit tests for multinomial models with large degrees of freedom.

*J Am Stat Ass.* 87(420): 1145-1152.

Osler T., Rutledge R., Deis J et al. (1996)

ICISS; An International Classification of Disease-9 based Injury Severity Score.

*J Trauma.* 41(3): 380-388.

Osler T., Baker S., Long W. (1997)

A modification of the Injury Severity Score that both improves accuracy and simplifies scoring.

*J Trauma.* 43(6): 922-926.

Osler T., Badger M., Rogers F et al. (2002)

A simple mathematical modification of TRISS markedly improves calibration.

*J Trauma.* 53(4): 630-634.

Osterwalder JJ., Riederer M. (2000)

Quality assessment of multiple trauma management by ISS, TRISS or ASCOT?

*Schweiz Med Wochenschr.* 130(14): 499-504.

Picard RR., Berk KN. (1990)

Data splitting.

*Amer Statist.* 44(2): 140-147.

Pigeon JG., Heyse JF. (1999a)

A cautionary note about assessing the fit of logistic regression models.

*J Applied Statistics.* 26: 847-853.

Pigeon JG. (1999b)

An improved goodness of fit statistic for probability prediction models.

*Biometrical Journal.* 41: 71-82.

Pulkstenis E., Robinson TJ. (2002)

Two goodness of fit tests for logistic regression models with continuous covariates.

*Stat Med.* 21: 79-93.

Quenouille MH. (1949)

Approximate tests of correlation in time series.

*J R Statist Soc. B.* 11: 18-84.

Royston P. (1992)

The use of cusums and other techniques in modelling continuous covariates in logistic regression.

*Stat Med.* 11: 1115-1129.

Roumen RM., Redl H., Schlag G et al. (1993)

Scoring systems and blood lactate concentrations in relation to the development of adult respiratory distress syndrome and multiple organ failure in severely traumatized patients.

*J Trauma.* 35(3): 349-355.

Rutledge R., Fakhry S., Baker C et al. (1993)

Injury severity grading in trauma patients; a simplified technique based on ICD-9 coding.

*J Trauma.* 35: 497-507.

Rutledge R., Hoyt DB., Eastman AB et al. (1997)

Comparison of the Injury Severity Score and ICD-9 diagnosis codes as predictors of outcome in injury: analysis of 44,032 patients.

*J Trauma.* 42(3): 477-487.

Rutledge R., Osler T., Emery S et al. (1998)

The end of the Injury Severity Score (ISS) and the Trauma and Injury Severity Score (TRISS): ICISS, an International Classification of Diseases, Ninth Revision-based prediction tool, outperforms both ISS and TRISS as predictors of trauma patient survival, hospital charges and hospital length of stay.

*J Trauma.* 44(1): 41-49.

Ryan TP. (1997)

Logistic Regression In *Modern Regression Methods*

John Wiley & Sons Inc. New York.

SAS Website

SAS macro. Name:- *jackboot*

Title:- Jackknife and Bootstrap Analyses 1995

[www.sas.com/techsupport/macro/jackboot](http://www.sas.com/techsupport/macro/jackboot)

Sacco WJ., Copes WS., Bain LW et al. (1993)

Effect of pre-injury illness on trauma patient survival outcome.

*J Trauma.* 35(4): 538-543.

Sacco WJ., MacKenzie EJ., Champion HR et al. (1999)

Comparison of alternative methods for assessing injury severity based on anatomic descriptors.

*J Trauma.* 47(3): 441-447.

Sauaia A., Moore FA., Moore EE et al. (1994)

Early predictors of post injury multiple organ failure.

*Arch Surg.* 129(1): 39-45.

Sauaia A., Moore FA., Moore EE et al. (1995).  
Epidemiology of trauma deaths: a reassessment.  
*J Trauma*. 38(2): 185-193.

Stephenson SC., Langley JD., Civil ID. (2002)  
Comparing measures of injury severity for use with large databases.  
*J Trauma*. 53(2): 326-332.

Steyerberg EW., Harell Jr FE., Borsboom GJ et al. (2001)  
Internal validation of predictive models: Efficiency of some  
procedures for logistic regression analysis.  
*J Clin Epidemiol*. 54: 774-781.

Stone M. (1974)  
Cross-validated choice and the assessment of statistical  
predictions (with discussion).  
*J Royal Statist Soc. B* 36: 111-113.

Stukel TA. (1988).  
Generalised logistic models.  
*J Amer Statist Assoc*. 83 426-431.

Tukey JW. (1958)  
Bias and confidence in not quite large samples.  
*Ann Math Statist*. Abstract. 29: 614.

White H. (1982)  
Maximum likelihood estimation of misspecified models.  
*Econometrica*. 50(1): 1-25.

