University of Glasgow
Faculty of Medicine
Department of Public Health

AN EVALUATION OF INTENSIVE CARE SEVERITY OF ILLNESS SCORING MODELS.

by Brian Mark Livingston BA MSc RGN

Thesis submitted to the Faculty of Medicine, University of Glasgow
for the degree of Doctor of Philosophy, October, 1999.

Declaration

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Glasgow, Faculty of Medicine, Department of Public Health. Unless stated otherwise, the work is that of the author.

The Evaluation of Intensive Care Severity of Illness Scoring Models

Contents

Abstract

Objectives: To evaluate the accuracy of the four main Intensive Care severity of illness scoring models using a large Scottish database, and to investigate different strategies for improving their accuracy in a Scottish setting.

Method: Twenty two out of 25 general adult Intensive Care Units in Scotland collected data for two and half years to allow calculation of Acute Physiology and Chronic Health Evaluation (APACHE) version II and III, Simplified Acute Physiology Score (SAPS) version II, Mortality Probability Model (MPM) version II (calculated on admission and at 24 hours). The models' Goodness of Fit (discrimination and calibration) and performances in subgroups (Uniformity of Fit) was evaluated using Receiver Operating Characteristic Curves, Hosmer-Lemeshow Goodness of Fit test, Chi Squared test and Confidence Intervals. Three of the Models (APACHE II, SAPS II, and MPM II) were customised with Scottish data using logistic regression techniques. Data quality was monitored by rescoring 10% of records over the first nine months of the project, with only the MPM_0 model showing any bias in the probabilities because of data collection errors.

Results: All models had good discrimination but poor calibration. However, the SAPS II and APACHE II models appeared to have better calibration than other models. After excluding patients who where discharged before the first 24 hours there was a deteriotion in both calibration and discrimination in all but the MPM_{24} model. APACHE III, SAPS II and MPM_{24} had superior discrimination to other models in this cohort. Analysis of the Uniformity of Fit showed significant differences between observed mortality and that estimated by the model in important subgroups of patients.

Substituting a pre-sedated Glasgow Coma Score, when patients were sedated over the first 24 hours, improved the discrimination and calibration in patients with altered scores in both the APACHE II and APACHE III models. The performance of the APACHE III model improved, but despite improved discrimination the calibration of the APACHE II model deteriorated.

After customisation all new models showed significant improvements in accuracy. However, only the new SAPS II models had no significant differences between the observed mortality and the estimated mortality using

the Hosmer-Lemeshow Goodness of Fit test. All models, except the new APACHE II model, showed significant differences in important subgroups.

Conclusions: Questions remain about the accuracy of these models even after customisation. Further research is needed to investigate variations in Intensive Care Units and the relationship to clinical effectiveness. However, where case mix adjustment is needed the new customised models remain the most accurate means of doing this in Scottish data.

List of abbreviations used in this thesis

A&E- Accident and Emergency
AIS- Abbreviated Injury Scale
AMS- APACHE Medical Systems
APACHE- Acute Physiology and Chronic Health Evaluation
APS- Acute Physiology Score
CI- Confidence Intervals
CMAAS- Committee on Medical Aspects of Automotive Safety
CRAG- Clinical Resource and Audit Group
DRG- Diagnostic Related Group
GCS- Glasgow Coma Score
GOF- Goodness Of Fit
RGO- Registrar General's Office
HDU- High Dependency Unit
ICS- Intensive Care Society
ICU- Intensive Care Unit
ISD- Information and Statistics Division
MPM- Mortality Probability Model
NHS- National Health Service
ROC- Receiver Operating Characteristic
SAPS- Simplified Acute Physiology Score
SD- Standard Deviations
SICS- Scottish Intensive Care Society
SMR- Scottish Morbidity Record
TISS- Therapeutic Intervention Scoring System
UK- United Kingdom
US- United States of America

Chapter 1- General Introduction

Contents:

1.1 Introduction

As the proportion of health care resources spent on Intensive Care increases, so too does the demand for accurate measures of quality of care. Also, with the emergence of the purchaser and provider split in the 1990s, there has been a focus on clinical effectiveness. More recently the White paper "Designed to Care" published in 1998 has highlighted the importance of clinical audit and the principles of clinical governance (1).

A report into Intensive Care by the Kings Fund in 1989 tried to answer five questions (2):

1. Is there scientific evidence that Intensive Care Units (ICU) cause a decrease in mortality and morbidity?
2. What criteria should be set for admission and discharge to ICUs?
3. Which classes of patients are likely to benefit most from which procedures that are carried out in ICU?
4. For what extra cost is therapeutic benefit gained by using Intensive Care?
5. What scale of provision is needed in the NHS?

They concluded that:

> ".....the lack of data in the United Kingdom would make it impossible to answer the questions posed."

and

> ".....the absence of data on workload, outcome and costs and the heterogeneity of ICUs, make it evident that any recommendation about future provision must be highly speculative."

They also concluded that:

> "Each unit should collect and evaluate data on clinical outcome and costs, both in general and for the care of individual patients."

It was against this background that the Scottish Intensive Care Society (SICS) applied to, and received money from, the Clinical Resource and Audit Group (CRAG) of the Scottish Office for a three year national audit of Intensive Care. The funding also included the monies for this research into the accuracy of existing severity of illness models.

The Objectives of this PhD were:

- To test and assess the accuracy, in a Scottish setting, of the five severity of illness models most commonly in use. These being the Simplified Acute Physiology Score (SAPS) version II, Mortality Probability Model (MPM) version II, Acute Physiology and Chronic Health Evaluation (APACHE) version II and III (3-6).
- To assess the applicability of severity models in a Scottish setting and therefore assess the portability of these systems out of their original development setting.
- To assess, where applicable, different strategies for improving the accuracy of these models within the Scottish Intensive Care population.
- To improve, where possible, the accuracy of these models within the Scottish Intensive Care population.

1.2 Layout of this thesis

The thesis contains 9 chapters. All figures and tables appear as close to the text as possible and are numbered sequentially according to the chapter in which they appear, for example, Chapter 6 has figures 6.1- 6.9. Each chapter has a coloured first page with contents. References and appendices appear at the end of the thesis with appendices running from Appendix 1 to Appendix 3.

1.3 The Chapters

Chapter 2 reviews the literature and history of Intensive Care scoring models. The chapter also discusses the appropriate statistics by which the performance of scoring models are measured. It outlines the reasons the study was undertaken.

Chapter 3 describes the units and patients used in this study. It also describes the methods used to collect data and the statistics and analysis used on the data.

Chapter 4 describes the methods used to ensure data were consistent and accurate. It highlights areas of uncertainty where it was not possible to guarantee data quality. It describes the methods and presents analysis of the re-scoring of 10% of records over a nine month period. It discusses the analysis of this data making conclusions on the useability of the data.

Chapter 5 describes the ICUs and patients in the study. It also presents the results from the analysis of the overall performance of each of the models. It discusses the implications of the results of this overall analysis.

Chapter 6 presents analysis of data assessing the uniformity of fit of the models. The results are discussed and conclusions made on the analysis

Chapter 7 presents methods and results of analysis of the performance of the APACHE models when a pre-sedated Glasgow Coma Score (GCS) is used if patients are sedated for the first twenty four hours. The chapter discusses the results and the implications for the models.

Chapter 8 describes the methods and analysis of results from a process of customisation. Implications from the results are discussed and conclusions on their accuracy are made.

Chapter 9 describes the methods and analysis of results from a process of customisation based on a temporal split in the data. Implications from the results are discussed.

Chapter 10 discusses the results from this study and the implication for the future of these models. It discusses possible reasons for inaccuracies in the models and possible future uses for Intensive Care scoring models.

1.4 Acknowledgements

Many people have helped me in the process of carrying out this research and contributed to the completion of this thesis. In particular I would like to thank:

Ray Jones for allowing me to undertake this research and for all his support during the project.

Cameron Howie for his support throughout this project.

Brian Millar for his support and advice during the project.

Harper Gilmour for advice on statistics

Clinical Resource and Audit Group for funding the project.

The nursing and medical staff for collecting the data without which the research would not have been possible (Participating units and the co-ordinating consultants are listed in Appendix 1).

Fiona MacKirdy for carrying out the validation for the Audit and also proof reading this thesis.

Finally Sarah, Zoe and Orla whose support and sacrifice allowed me to undertake this research.

1.5 Publications and presentation of work

The work described in this thesis has either been published, submitted or is in a stage of advanced preparation for publication, as follows:

Published or in press

Livingston B, MacKirdy F, Howie J, Jones R, Norrie J. Assessment of the performance of five Intensive Care Scoring Models within a large Scottish database. Critical Care Medicine (In Press).

Livingston B, Mackenzie S, MacKirdy F, Howie J. Should the pre-sedation GCS be used when calculating APACHE scores for sedated patients . Critical Care Medicine 2000; 28: 389-394.

Mackenzie S, Livingston B, MacKirdy F, Howie J. The use of pre-sedated GCS improves the performance of APACHE III. Intensive Care Society December 1997 (Abstract, Intensive Care Society Autumn meeting November 1997) Clinical Intensive Care 1997; 8(2):9.

Livingston B, MacKirdy F, Millar B, Howie J, Jones R. An assessment of Intensive Care scoring systems within a national audit. European Congress of Intensive Care Medicine 1996;187-191. (Abstract)

MacKirdy F, Livingston B, Howie J, Millar B. The effects of errors in data collection on mortality prediction generation. European Congress of Intensive Care Medicine 1996;193-197. (Abstract)

In preparation

Livingston B, Mackenzie S, Blatchford O, Knill Jones R, MacKirdy F, Howie J. Customisation of four Intensive Care scoring models using data from a large Scottish database. Paper in preparation.

1.6 Contributors to this thesis

The Author contributed to the design of the project, was responsible for co-ordinating the collection, analysis and interpretation of all the data in the thesis, and wrote all the chapters of this thesis with the exception of Chapter 7.

Fiona MacKirdy contributed to the design of the data validation, was responsible for the collecton of the data for Chapter 4, and advised on its interpretation.

Cameron Howie had the original idea for the project, contributed to the design of the project, was responsible for the overall management of the project and advised on the interpretation of the data.

Simon Mackenzie had the original idea for the analysis in Chapter 7 and jointly wrote the paper with the author on which Chapter 7 was based.

Chapter 2- Literature Review

Contents:

2.1 Intensive Care development and definition

The development of Intensive Care can be traced to the emergence of endotracheal intubation with positive pressure ventilation. In 1952 a Danish epidemiologist reported reduced mortality rate in patients with respiratory failure as a result of endotracheal intubation and ventilation (7). Since then Intensive Care has developed into an important element of medicine. It consumes considerable resources, i.e. an estimated 1-2% of the hospital budget in the UK and a considerably larger 22% in the United States of America (US)(8,9). Many countries now treat Intensive Care as a separate specialty. However, in the UK it still remains a subspecialty of either Anaesthesia or Surgery.

Some have pointed to post-operative recovery rooms as being the earliest origins of Intensive Care with Florence Nightingale saying as early as 1852:

> "It is valuable to have one place in the hospital where post-operative and other patients needing close attention can be watched."

As well as recovery units, there have been a number of areas where specialist units have developed, i.e. trauma, burns, cardiac and neurosurgical ICUs. Intensive Care is now an integral part of hospital medicine.

There are a number of definitions of Intensive Care, a report from the King's Fund panel defined it as (2):

> "a service for patients with potentially recoverable diseases who can benefit from more detailed observation and treatment than is generally available in the standard wards and departments".

This has been criticised as being too broad for clinical practice (10). The Intensive Care Society (ICS) in a report from the Standards Subcommittee (11) defined Intensive Care as:

> ".....the highest level of continuing patient care and treatment... It has as its primary objective the recovery of the patient at least to the stage of return to an intermediate care ward"
>
> and
>
> "...involves continuing supervision, care and treatment by doctors, nurses, physiotherapists, technicians, dieticians and others."

In 1990 the ICS (12) further defined it by saying it:

> "...is usually reserved for patients with potential or established organ failure..."
>
> and
>
> "...requires a multi-disciplinary team approach and the highest possible standards of nursing and medical care. A nurse/patient ratio of 1:1 should be the minimum and the services of a full-time medical resident are essential."

The ICS defined an ICU as:

> "a specially designed ward where facilities for the critically ill are concentrated and where the level of care and supervision is appreciably greater than on an ordinary ward."(11)

Rowan K. (13) argues that most definitions highlight two distinct tasks firstly:

> "...the prevention of, or early detection of and rapid response to, life-threatening complications in patients who are judged to be at risk of becoming critically ill."

and the second is:

> "the substitution, temporarily, for the physiological function of one or more organ systems that have failed, until the crisis is past."

Although there is a general consensus as to what Intensive Care is, there is less certainty and a lack of evidence that it provides significant improvements in the outcome of those patients receiving it. There is general agreement that the development of Intensive Care was not accompanied by adequate evaluation. The Kings Fund panel in 1989 asked:

> "Is there scientific evidence that ICUs cause a decrease in mortality and morbidity?"

They concluded that there was not enough data to answer this question (2).

Some studies have questioned the effectiveness of Intensive Care. Hook et al have reported similar mortality in patients with pneumococcal bacteremia receiving Intensive Care and those receiving alternative treatment (14). However, this research is contradicted by Rogers et al (15) who showed a reduction in mortality in patients with acute respiratory failure following the introduction of an ICU. A study of provision of Intensive Care in England and Wales reported similar mortality rates in those receiving

Intensive Care and those patients who were refused admission (16). This study did not analyse the nature of those patients who were refused admission to the ICUs.

In the early years of Intensive Care it would have been possible to have conducted randomised controlled trials to evaluate the effectiveness of Intensive Care and the treatments being administered. However, given the widespread use and acceptability of Intensive Care and ICUs it would now be unethical to evaluate treatment by refusing entry to ICUs to some people. This has lead to increasing pressure for clinicians to demonstrate the effectiveness of the treatment that they are giving.

2.2 Audit in Intensive Care

Clinical Audit has been described as:

> "systematic, critical analysis of the quality of medical care, including the procedures used for diagnosis and treatment, and the resulting outcome for the patient." (17)

The process of Audit is often thought of in terms of a cycle (Fig 2.1)

There are three things that can be measured i.e. structure, process and outcome. It is generally accepted that structure and process (the way care is provided) both affect outcome. Bion has suggested a more complex structure for Audit in Intensive Care (Fig 2.2) (18). This is based on the review of standards in structure, process and outcome. The Department of Health has published guidelines that have recommended numbers of ICU beds as between 1-2% of the acute beds (18). Guidelines published by the ICS have recommended that an ICU should have a maximum occupancy of 70% to allow for periods of peak demand (11). There are also recommendations around nursing staff numbers. The ICS recommends that there should be a minimum of one nurse per patient (12) with most ICUs implementing this as practice. How this translates into actual numbers depends on the occupancy of each ICU. However, the ICS has given methods for calculating the numbers of nurses required (11). ICUs are also expected to provide 24 hour consultant cover (18). Levels and experience of medical staff could be expected to have an effect on patient outcomes.

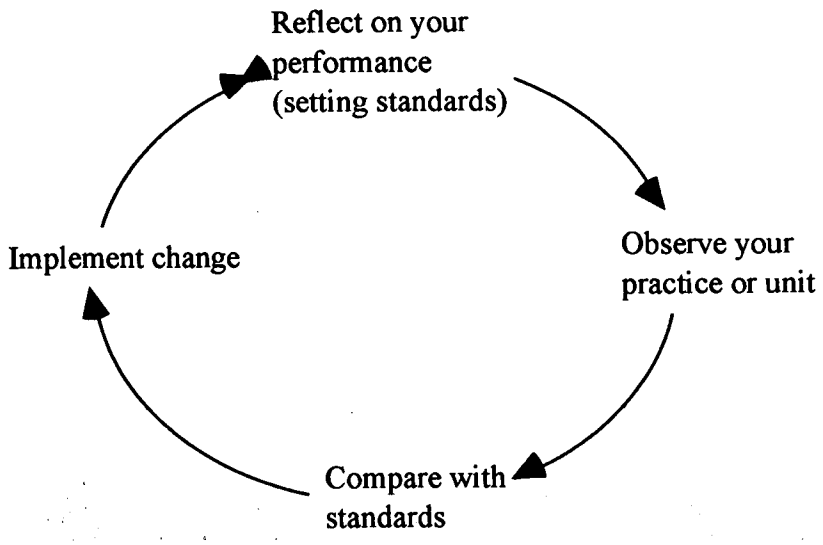


Fig 2.1 Audit cycle

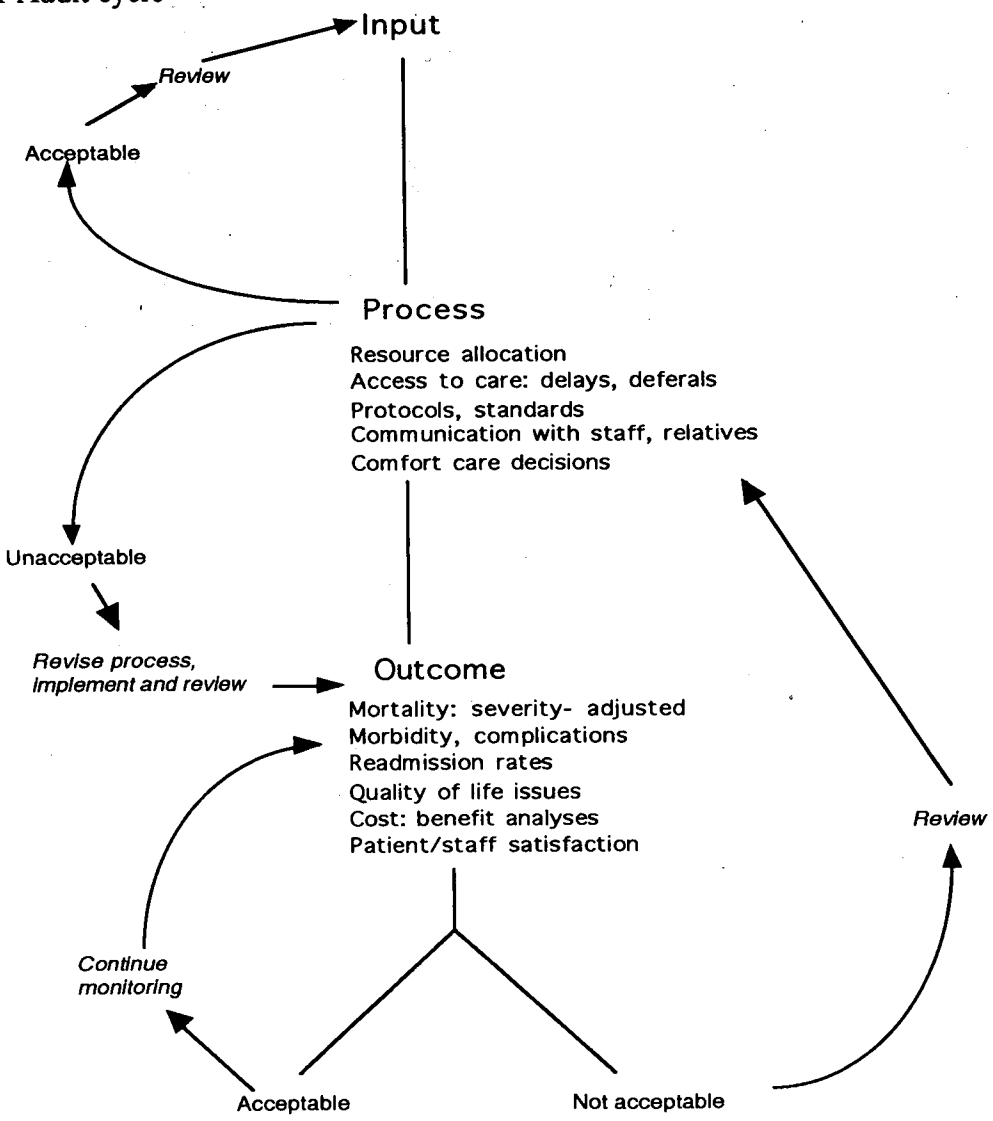


Fig 2.2 A structure for Intensive Care Audit.

However, Audit in Intensive Care has tended to focus on outcome. The majority of this focus has been on mortality. Intensive care is unusual because by its very nature it is concerned with the care of the most sick patients, making death a fairly common event. If process affects outcome then it should be possible to relate it to mortality. However, this has been avoided in the past as it is difficult to disentangle all the factors that might be involved. Although it is clear that process is an important aspect of medical care it has been argued by Bion that:

> "....given a choice between a high quality process of care with a poor outcome, or poor process with a good outcome, most people would choose the latter." (18)

However, the use of mortality as an outcome measure is not as straightforward as it would seem. There is not necessarily a direct relationship between the effectiveness of care and mortality. A number of studies have shown a wide distribution of mortality in ICUs (19,20). The ICUs in the UK APACHE II study had a hospital mortality that ranged from 21% - 45%. The mortality rate in the ICUs in this study ranged from 23%-42% (Table 5.1). It is not appropriate to compare mortality rates in ICU as a measure of quality of care. The mortality rate in an ICU will be related to the type of patient that is admitted. The reason for admission to an ICU is usually to support one or more organ system. The causes of organ failure can be due to a number of underlying disease processes. The expected survival from individual diseases is very different, for example an acute but severe asthma attack may lead a patient to be admitted to an ICU for respiratory support. Similarly someone could be admitted with respiratory failure due to chronic emphysema, but the expected outcome from both these cases will be considerably different. Bion points to various determinants of outcome from critical illness (18).

1. Severity of illness
2. The type of disease process (diagnosis)
3. Physiological reserve (previous health and age)
4. Specificity and timing of treatment.

Rowan pointed out that there will always be unquantifiable factors like "will to live" that will have an affect on outcome (13).

The success of Audit depends on the ability to set measurable standards. As mortality is affected by a number of factors it is clear that some sort of case mix adjustment is needed before legitimately using it as an outcome measure. Increasingly, Intensive Care severity of illness models are being used to adjust for case mix and to allow the setting of standards.

2.3 Statistical issues in severity of illness modelling.

There are a number of statistical issues to be considered when assessing the performance of Intensive Care scoring models. The problems are largely because the scoring models generate probabilities that are on a continuous scale, however the outcome measure (hospital mortality) that is used to assess the accuracy of these models, is dichotomous. The two areas of a model's accuracy that can be measured are discrimination and calibration. Discrimination is the ability of the model to separate survivors and non-survivors. Calibration is the evaluation of the degree of correspondence between estimated mortalities and observed mortalities. Descriptions of discrimination and calibration and how they are measured are described in Chapter 3.

2.3.1 Calibration verses discrimination

The importance of discrimination and calibration will depend on the intended use of a model. Discrimination becomes more important if you are interested in individual mortality probabilities and their use to aid treatment decisions. However, as this study is interested in these models usage in measuring performance of different ICUs, calibration becomes more important. There is perhaps, some argument as to how much emphasis should be placed on the calibration of the models. Some have argued that discrimination should not be considered until a model can show acceptable calibration (21). Wagner, on the other hand, while acknowledging a model must have adequate calibration argues that it must also have good discrimination (personal communication). While accepting that a model is unlikely to be used appropriately without good calibration, for a model to identify those who are likely to live and die it must have good discrimination. If these models are to be used to assess performance in different ICUs they must clearly be able to identify those at high and low risk. It would be possible for a model with virtually no ability to discriminate, to calibrate well, especially if the probabilities were focused around the mortality level of the ICU population (although this is unlikely).

2.3.2 Changing measures of accuracy

The accepted tests for accuracy have changed as these models have been developed and a discussion of these developments takes place later in the chapter. When the first generation of the models examined in this thesis appeared, they reported only discrimination using crude classification tables. As the development of models continued the use of ROC curves became a standard measure of discrimination. Calibration curves were used initially as an indication of calibration. However, the development of the Hosmer-Lemeshow GOF test enabled this more robust measurement of calibration to be

adopted for most studies. A consensus conference into mortality probability models concluded that the performance of models should be determined by both ROC curves and the Hosmer-Lemeshow GOF tests (22). More recently, since the start of this study, some analysis of subgroups has been used to assess the accuracy of mortality probability models. When discussing the development of Intensive Care severity models, research reported in the literature will be described by the statistics originally used in the research.

2.4 The development and history of Intensive Care severity of illness modelling.

2.4.1 Critical Care Scoring Models

There are a large number of scoring models in medicine today, all designed to help decision making in some way. These vary from simple subjective models to more objective complex models based on statistical methods. The relatively low cost of computing has been partly responsible for the increase in the number of models. The use of computers to aid clinical decision making was demonstrated as early as the 1950s (23). Speglhalter and Knill-Jones have reviewed the different approaches to modelling in clinical decision making (23).

A major factor in America contributing to the large rise in use of models is the way health care is funded. Hospitals in the US have increasingly been encouraged to demonstrate both their clinical and cost effectiveness. Diagnostic Related Groups (DRGs) were an early attempt to adjust for case mix, with patients being allocated into related groups more appropriate for payment (24). However, DRGs take no account of severity of illness. A number of models have been designed to address this problem and have been designed to adjust for hospital case mix (25-29).

As there are such a large number of models designed to adjust for case mix, this review is based largely on those associated with critical illness.

2.4.1.1 Trauma and associated models

There have been a variety of models developed for trauma patients from the 1970s. These have tended to be either anatomical or physiological (13). The first of the anatomical models, the Abbreviated Injury Scale (AIS), was developed by the Committee on Medical Aspects of Automotive Safety (CMAAS) (30). There have been a series of anatomical models since then (31-34).

The first physiological trauma model to be developed was the Trauma Index in 1971 (35). There have been a series of new models since then (36-39) with the most recent being the ASCOT model (40)

2.4.1.2 Other models

There have been a number of other models in other areas. In 1982 Durocher et al devised a pneumonia scoring index which used 12 physiological variables to generate a score to indicate severity of illness (41). In 1983 a sepsis score was proposed by Elebute and Stoner using 25 physiological variables (42). A model for indicating severity in patients with cirrhosis, the Child-Turcotte criteria in medically treated cirrhosis, used physiological variables to indicate the severity of the condition (43). There have also been a series of models developed for use on Coronary Care Patients (44)

2.4.1.3 Glasgow Coma Score

One of the most successful and most widely used scoring systems in critically ill patients is the Glasgow Coma Score (GCS). The Score proposed by Teasdale and Jennett in 1974, uses three variables to assess the level of neurological impairment (45). These three variables i.e. eye opening, motor and verbal responses, are scored as follows:

Eye opening	Motor	Verbal
4. Spontaneously	6. Obeys commands	5. Oriented
3. To command	5. Localises pain	4. Confused
2. To Pain	4. Flexion to pain	3. Words only
1. None	3. Abnormal flexion	2. Sounds only
	2. Extension	1. None
	1. None	

Lower scores are an indication of the severity on the coma scale. It is widely accepted that the GCS is a good measure of neurological impairment. It has been demonstrated that increased levels of severity, as measured by the GCS, can be related to increasing mortality (13,46). A study in Scotland, the Netherlands and the US used the GCS to adjust for the severity of head injury to allow comparisons of the treatment they were receiving (47). Despite a wide difference in severity of injury no difference was found in the outcome of the patients. While generally accepted as the best measure of the neurological impairment, there are still questions about its use in Intensive Care (46,48,49). These tend to focus on the fact that patients are often heavily sedated in

Intensive Care, which could give them a much lower GCS than they would have had otherwise.

The models described above have had varying rates of success. It is clear that some fulfil the function of adjusting for severity in specific illness. However, they are all specific to certain types of disease or injury and do not adjust for case mix in the heterogeneous populations of ICUs.

2.4.2 The development of scoring models in Intensive Care

2.4.2.1 Therapeutic Intervention Scoring System

One of the earliest models to attempt to adjust for case mix was the Therapeutic Intervention Scoring System (TISS) (50). The basis for the model was the idea that the level of intervention would act as a proxy for severity of illness. The model had a list of 57 variables, each with a weighting somewhere between 1 and 4, depending on the indication of severity. The types of variables included in the model ranged from the type of monitoring that the patient was receiving, to more invasive procedures like dialysis, or insertion and use of a pulmonary artery catheter. The premise was the greater the level of intervention the greater the level of severity of illness.

However, the problem with this model is that intervention is not always associated with severity of illness, as the amount of intervention is dependant on the presenting disease. Clinicians may also vary in their practice, with some favouring more aggressive techniques, which would lead to higher scores than a clinician who favours a less interventionist approach.

TISS has become widely used in ICUs but as a surrogate for workload and as a tool for assessing cost (51-54). The model has since been updated by Keene and Cullen in 1983 (51). There are also many other ammended versions in use by other groups. As there are now a number of different versions of TISS, all with slightly different variable lists, there is a problem with comparing TISS scores on different units. As many units collect data for different versions of the model it makes it difficult to compare TISS scores. There has been considerable advances in the treatment and technology used in Intensive Care since the last update of the model, however, an ICS working group is presently updating and standardising the model. TISS has been used for a number of purposes:

- Comparing patients (55,56)
- Quantifying low and high risk patients (56,57)
- Monitoring low and high risk patients (56-58)

- Monitoring Nursing Workload (51,52)
- Current utilisation of beds and calculating future bed provision. (51,59,60)
- Resource allocation (53,54)
- Outcome comparisons in same unit over time and between different ICUs (59,61)

There are a number of weaknesses in the TISS model, as it is limited in the interventions and care it measures. For example patients will often need considerable care even when they are less sick, recovering patients may be confused and require a lot of nursing attention but this will not be reflected in the TISS score. Dying patients may have reduced interventions but increased workload for the staff.

2.4.2.2 Acute Physiology and Chronic Health Evaluation

The development of the first Acute Physiology and Chronic Health Evaluation (APACHE) model in 1981 (62) saw the beginning of the present generation of severity of illness models. These were generic models that attempted to measure severity of illness directly, on all patients entering ICU.

Variables for the model were selected by a panel of seven "experts" representing three specialities (Anaesthesiology, Medicine, Surgery), two of whom were the authors. The panel chose 34 variables that were considered to be helpful in estimating severity of illness, based on evidence found in the literature and on the opinions of the seven experts.

Continuous variables were grouped into categories, with one category representing normal and subsequent categories representing more abnormal values, with the weights for most variables from 0-4 (some variables could only get 1 or 2 points). The combination of points for each variable then produced the Acute Physiology Score (APS), which represented a measurement of severity of illness. The model also included a chronic health evaluation, ranging from A to D, with A representing prior good health, and D representing severe chronic illness prior to admission to the ICU.

The first 32 hours after admission to the ICU was used as the time period for data collection. Variables not measured or unavailable from within the first 32 hours were presumed to be normal. Patients admitted with myocardial infarction, burns patients and patients who were in the unit for less than 16 hours were excluded

Data were collected on 582 ICU admissions over an 8 month period. The authors then used multiple logistic regression and probit regression to examine the relationship between the APS and survival. The authors added other variables to the logistic

regression, age, sex, diagnostic system, chronic health and the patient's surgical status (operative or non operative). The authors reported a significant relationship between the APS and mortality. They also reported that the chronic health evaluation was only significant when it was at its most severe. The authors concluded that although the model was not significantly accurate for individual prediction it was an:

> ."....objective means for describing patient characteristics and estimating severity of illness over a broad range of ICU patients".

They also concluded that the model would be able to assess the similarity of severity of illness in non-randomised studies.

Application of the APACHE model: The APACHE model has been tested in a number of studies. A study of 700 consecutive admissions from five US surgical and medical ICUs that were not in the original study, demonstrated similar accuracy to that found in the original study (63). All five ICUs showed an average predicted mortality that was close to the observed mortality. The study was carried out by the original developers of the APACHE model and only included information from the first 24 hours rather than the first 32 hours. The model has also been applied to data from a number of European countries (France, Spain and Ireland) (64,65). These studies reported a significant relationship between the APS and hospital mortality. Data from a further 833 admissions were also analysed (66). Wagner et al reported that the APS from the APACHE model was

> "..... strongly and significantly associated with outcome within a number of specific cardiovascular, neurologic, respiratory and gastrointestinal diagnosis."

The authors argued that after validation the model could be used in other institutions.

The APACHE model has now been used to adjust for case-mix in a number of studies (54,59,67).

The APACHE model however never gained widespread usage, partly because of the large number of variables and considerable effort required to collect the data. There have also been a number of criticisms of the APACHE model (68):

- The variables were chosen and weighted by subjective methods.
- The developers recognised that, depending on which measurements are made, it is possible for the severity to be underestimated.
- Transfer from another hospital might lead to an underestimation of the model.

- As measures were done over the first 24-32 hours, it is not possible to rule out the effect of treatment causing an underestimation of severity of illness.
- The model was large and required substantial effort to collect data on a routine basis.

2.4.2.3 APACHE II model

Knaus et al updated the model in 1985 (5). The new APACHE II model had a reduced number of variables, with a shorter data collection period reduced from 32 hours to 24 hours. The number of physiological variables in the model was reduced, using clinical judgement, from 34 to 12. The variable weights were changed on the GCS, creatinine and blood gas variables, with the remaining variables maintaining their original weightings. The APS in the new model also included variables for age and a new chronic health evaluation. The model also included an admitting diagnosis, with 34 separately weighted

Table 2.1 Variables in the APACHE II model

Acute Physiology Score Variables
Temperature
Mean arterial pressure
Heart rate
Respiratory rate
Oxygenation
Arterial pH
Serum sodium
Serum potassium
Serum creatinine
Heamatocrit
White blood count
GCS
Age
Chronic health
Variables in equation
Admitting diagnosis
Emergency surgery
Acute Physiology Score

GCS, Glasgow Coma Score; APACHE, Acute Physiology and Chronic Health Evaluation.

diagnoses and a variable for indicating whether the patient was an emergency surgical patient. The APS, diagnostic category and emergency operative status were entered into a

logistic regression model with the APS, emergency surgery variable, and the diagnosis. With the coefficients produced from the logistic regression it was possible to calculate probabilities of hospital mortality. There was considerable variation in the crude mortality rates in different diagnostic catogories. The authors used the equal performance of these groups once the model was applied as an argument for the inclusion of the diagnostic category in the model. The authors reported an ROC curve value of 0.86 and good association between rising APACHE II scores and rising hospital mortality.

Application of APACHE II: The APACHE II model has been tested in a number of studies both in the US and internationally. The performance has been variable with studies reporting very different results. A number of researchers have reported adequate performance claiming that the model is applicable within the population tested (69-73). However, the overall picture varies with some studies reporting poor performance and many reporting negative results in different types of patients. There have been poor performances reported on patients with acute myocardial infarction, pulmonary oedema, trauma patients, post operative patients and patients receiving total parenteral nutrition (74-78). This picture is confusing and inconsistent as some studies report good performance in the same patients (79,80).

One study in the UK is the most comprehensive analysis of the model in the UK (UK APACHE II Study) (13,20,73) . The study involved 26 general ICUs from Great Britain and Ireland, with 8796 admissions to the ICUs. The study, by the ICS, reported an overall GOF and predictive ability that was good while Rowan et al reported that they had only a slightly inferior performance to that in the original model. They reported an area under the ROC curve of 0.83, compared with 0.86 in the original study. They reported a Chi Squared value of 79.81 (8 degrees of freedom). However, although the study reported that this was a reasonable GOF ,they did not report that the Chi Squared value quoted demonstrates a significant difference ($p<0.001$) between observed mortality and predicted mortality. This result may be partly due to the large numbers in the study. Rowan et al did report significant differences between observed mortality and that predicted by the model in operative patients with cardiovascular, neurological and respiratory diagnoses. They also reported significant differences in non operative patients with neurological diagnoses.

There are a number of smaller UK studies of APACHE II (81,82) also showing variable performance. Beck et al in 1997 reported poor performance of the APACHE II model in a small number of patients from one English ICU, with an overestimation of mortality and significant goodness of fit test ($P<0.001$) (83). A large study (12,762) involving 24 ICUs in the South West Thames region, though not specifically reporting the

performance of the model, showed poor goodness of fit of the model with a Chi Squared value of 180.9 ($P<0.001$) (84). The overall picture remains varied and inconsistent

The APACHE II model, despite this inconsistent picture, has become the model most widely used in both the US and the UK. The model has been applied to adjust for case mix for a number of purposes, using the model to:

- demonstrate the level of severity of illness in research studies. (85)
- compare units. (20,86)
- evaluate the efficacy of new treatments (85,87)
- evaluate cost effectiveness (88)
- evaluate the outcome of patients (72,86)
- predict individual mortality (89-91)

There are a number of criticisms of the APACHE II model. The model was developed and validated on the same patients. As the same patients have been used to develop the logistic regression coefficients as have been used to test the model, it would be expected that the model would appear to perform well in that data set. It is now widely accepted that models must be validated outside the data set in which they have been developed (22). The weights for the variables that make up the APS are derived subjectively and may not represent the actual relationship the variable has to mortality. The model again treated missing results as normal and like the original APACHE model could be affected by the differences in data collection. The effect of treatment over the first 24 hours could not be ruled out. The start of data collection and where treatment begins could lead to possible underestimation of mortality (lead time bias).

2.4.2.4 Development of APACHE III

In 1991 Knaus et al proposed a new APACHE model (6). The new model was developed using data collected from 41 ICUs from across the US. They collected data on 17,440 patients from both medical and surgical ICUs. The new model involved some fundamental changes from the APACHE II model. The authors compiled a list of 212 disease categories allowing the choice of a diagnosis to indicate the primary reason for admission. These diagnoses were classified into medical and surgical groups and by which major organ system was involved. The model also included variables indicating the patient's source of admission and length of time at this source. The authors used 20 physiological variables based on past experience and evidence in the literature.

Table 2.2 Variables in the APACHE III Acute Physiology Score

Acute Physiology Score Variables
Pulse
Mean blood pressure
Temperature
Respiratory rate
Oxygenation
Heamatocrit
White blood count
Creatinine
Urine output
Urea
Sodium
Albumin
Bilirubin
Glucose
Acid base
GCS
Age

GCS, Glasgow Coma Score.

The data from all 40 ICUs were combined and patients were then randomly split into a development and validation set. Weights for physiological variables were derived using multiple logistic regression using a mixture of categorical and continuous variables. The authors investigated and found significant interactions between some variables. Where such interactions were present, combined variables were created and weighted accordingly. Weights for the physiological variables were derived from a development set and then tested on the validation set.

Weights were derived for chronic health variables and five year age categories from the overall database. Weights for all variables were then converted to integer points to form the new APS. Weights for the diagnostic categories were generated from the overall sample. However, unlike previous models, the multiple logistic regression equation, and the coefficients generated were not published and are now proprietorial, with the rights to use the APACHE III model being held by a commercial company (APACHE Medical Systems Inc.). The authors also investigated the possibility of using the model as a basis for predictions over time.

The authors reported improved performance from that demonstrated by the APACHE II model. The APACHE III model had an improved discrimination with a reported area under the ROC curve of 0.90 compared with 0.85 in the APACHE II model. The overall correct classification at 0.50 predicted risk in the APACHE III model is 88.2% compared with 85.5% in the APACHE II model. However, as calibration curves were the only tool used for measuring the calibration, it is not possible to fully assess this aspect of the model. The authors reported that the model reduces the amount of unexplained variation in hospital mortality due to previously unmeasured patient characteristics. The authors imply that the addition of new variables and their change in approach to the weighting of physical variables has improved the model's ability to adjust for case mix. The authors claim that the model can be used to ensure that there is equal severity of illness in different groups in research trials. They also report that the new model accounts for a considerable proportion of the difference in hospital mortality in different ICUs (R^2=0.90). They argue that the majority of the difference between ICUs can therefore be attributed to patient characteristics rather than treatment. They also go on to claim that the difference between observed mortality is one measure of quality of care, but bed availability and screening for admission can vary in different hospitals which effects the legitimacy of this analysis.

The authors, in 1994, continued to further extend the model to provide daily estimates of mortality (92). The authors reported good discrimination over the first 7 days of Intensive Care, with areas under the ROC curves ranging from 0.90 on the first day to 0.84 for patients still in the ICU on the seventh day. They concluded that daily estimates might help improve outcome from clinical decision making in Intensive Care.

Application of APACHE III: The APACHE III model has been used in a number of studies (57,93-99). However, the model has not enjoyed such wide spread usage as the APACHE II model. This is almost certainly because in order to use the model, studies would have to pay a licence fee for both the usage of the software and the APACHE III equations, although the equations are available free on request to researchers. The reported studies investigating the performance of the APACHE III model have recorded differing performance. Other studies have reported similar results to the original model (100,101). However, a Brazilian study showed a considerable underestimation of mortality in 1,734 patients despite reasonable discrimination (area under the ROC curve 0.82) (102). The study further reported that this discrepancy was, in part, associated with the level of technology available to the different units (103).
A small study in a UK ICU reported that the APACHE III model had reasonable discrimination (area under the ROC curve 0.847) (83). However, the model showed poor calibration with a consistent underestimation of mortality, reflected in both the

calibration curve and the Hosmer-Lemeshow GOF test. The study also reported poor calibration in different diagnostic categories, but the results from the analysis were not consistent as some diagnostic groups showed poorer fit than others. The numbers in this study were relatively small (n=1,144) and it is therefore hard to draw any firm conclusions. There has been a large UK study of the APACHE III model in 17 ICUs in the South West Thames area. The study reported results on 12,793 patients (84,104,105). The results reported better discrimination than that reported in the original paper. But the study showed poor calibration and reported a significant difference between observed mortality and that predicted by the model, with a significant Hosmer-Lemeshow test (X^2=543.12; P<0.001).

The APACHE III model has been used in a number of studies to:

- assess the cost and outcome in different ICUs (96,97).
- make comparisons of outcome in different ICUs (98)
- predict the likelihood of life support in the following 24 hours (95)
- examine process in ICUs and compare with case mix adjusted mortality (94).
- adjust for case mix in a study assessing the effectiveness of right heart catheterisation (93).

The new model was developed with less subjectivity, with weights being generated using statistical methods. Although the developers from the original studies have demonstrated improved performance in the APACHE III model there have been a number of criticisms of the model. Some of the weights in the new model were based on data from a development cohort and then tested on an independent cohort. However, the results from the overall model were from analysis of the total data set. There is general consensus that models must be validated and tested on independent cohorts (22). To use the model, ICUs must pay for both software and for the use of the equation, thus making access to validate the model difficult. The authors of the original study have reported calibration curves but have not reported Hosmer-Lemeshow GOF tests which is the standard measure of GOF (22) as indicated by a consensus conference on mortality prediction.

2.4.2.5 Development and Application of Simplified Acute Physiology Score

The Simplified Acute Physiology Score (SAPS) model was proposed in an effort to produce a simpler severity of illness model than the APACHE model (106). Le Gall et al used statistical techniques to determine which of the 34 APS variables was significantly associated with mortality. The SAPS model reduced the number of variables to 13 mandatory variables in an effort to reduce the bias produced by missing values. The

authors proposed the model as a means of classifying patients into groups that composed of comparable probabilities of death.

As with the APACHE model, the authors showed good correlation with mortality when tested on 126 patients. The model has been applied in a number of studies with different groups of patients. (74,107-109)

Development of SAPS II: The model was updated in 1993 by Le Gall et al (3). The new model (SAPS II) was based on a large cohort of patients (14,745) taken from 137 ICUs from 12 different countries. The authors split the database into a development set (8,369) to construct the model and a validation set to test the model (4,628). The study collected data on 37 variables and tested each to determine an independent association on hospital mortality. Of the 37 variables 17 were included in the final model. The LOWESS smoothing function was used to determine the categorisation of variables before using multiple logistic regression to determine the weightings for each of the variables and their categories. The development cohort was then used to develop a multiple logistic regression equation that would allow the calculation of a probability of mortality for each patient. The equation included the score but also a log transformation of the score to compensate for the skewed distribution of the score. Results reported for the new model suggested that the performance was good, with an area under the curve of 0.86 and a non significant Hosmer-Lemeshow test. Probabilities of mortality were calculated for the old SAPS model to allow comparisons with the older model. The new model would appear to be a significant improvement, with the area under the ROC curve for the old model of 0.80 compared to 0.86 in the new model.

The authors concluded that the model is

> "....an extremely effective system for estimating the probability of mortality for ICU patients."

However, they also concluded that any future research should concentrate on testing the model in common cohorts, using all available systems. The authors suggest that the system should be used for analysis of aggregated data rather than for individual patients and any result treated with considerable care.

Table 2.3 Variables in SAPS II model

Heart rate
Systolic pressure
Temperature
Pulmonary artery pressure (if ventilated)
Urinary output
Serum urea level
White blood count
Serum potassium
Serum sodium
Serum bicarbonate
Bilirubin
GCS
Chronic health
Type of admission

GCS, Glasgow Coma Score

Application of SAPS II: The new SAPS II model has been more widely used than the original model. There have been a number of studies looking at the performance of the SAPS II model with a variety of results (110-113). Castella et al published results based on the data from the European and American multi-centre study, from which the SAPS II model was developed and initially validated (21). These results showed that the SAPS II model had a better performance than the original model. However, by far the largest and most important evaluation of the SAPS II model has been as part of the EURICUS-1 study (114,115). The model was tested on a database of 16,060 patients admitted consecutively to 89 ICUs. The model had good discrimination with an area under the ROC curve of 0.82. However, the model exhibited poor calibration with a significant Hosmer-Lemeshow test and large variations in the performance of different subgroups of patients.

The Model has been used in a number of studies (41,116,117).

The SAPS models have been criticised for not including a diagnostic element. The argument is that the probability of a patient surviving ICU treatment will be profoundly affected by the underlying diagnosis. Le Gall et al discuss this in the paper and conclude that it is not possible to accurately include a diagnostic weighting in the model (3). They point to the difficulty in choosing one diagnosis in ICU patients who often have very complex underlying disease processes.

2.4.2.6 Development of the Mortality Probability Model

There have been two distinct phases in the development of the Mortality Probability Models (MPM). The first was the development of the original MPM model and the second the development of the MPM II model, which will be described later in this chapter. The development of the original MPM model occurred over a number of years and involved the production of a number of different models.

In 1985 Lemeshow et al proposed two models which were both developed from an original data set of 755 admissions to a single US ICU(118). They collected data on 137 variables at admission and 75 variables at both 24 and 48 hours. Using a stepwise linear discriminant function they determined which variables were significantly associated with mortality at admission and for patients who remained in the ICU after 24 hours. The developers produced two models, one including variables collected before or on admission ($MPM_{0\text{-}of}$) and one using variables from both before and since admission up to an initial 24 hour period (MPM_{24}). Each of the models contained 7 variables. The MPM models required only a yes/no answer to questions. This makes data collection much easier with less likelihood of missing data.

In 1987 a second admission model was proposed containing 3 more variables related to cardio-pulmonary resuscitation ($MPM_{0\text{-}cpr}$) (119). In 1988 Lemeshow et al introduced three new models based on an increased data set of 1,997 patients (120): a new 24 hour model, a model calculated for patients remaining in ICU at 48 hours (MPM_{48}) and a model predicted over time (MPM_{0t}). Underpinning the MPM models was the principle that the probability of mortality would change over time, and that the importance of different variables may change as well. The same principles of univariate testing of the variables for association with mortality, before entering into a multiple logistic regression model, were applied.

Little is published on the performance of these early models. They report discrimination in the form of classification tables for a cut off point of typically 50%. The first admission model $MPM_{0\text{-}ofm}$ had a correct classification of 87% with the original MPM_{24} reporting 85% (118). Teres et al reported a correct classification for the original MPM_0 model of 86% with the new admission model $MPM_{0\text{-}cpr}$ having a classification of 85% (119). These are crude measures of discrimination and give us little picture of the overall performance of these models. Schaefer et al in 1990 validated the MPM_{0}-cpr model on an independent database, describing a significantly higher mean probability of death for those that died than those that survived (121).

The numbers involved in constructing these models are small, with the revised models based on data sets of no more than 1,997 patients. It is also hard to get a picture of how these models correlated with severity. However, there were certain advantages in these new models when compared to their contemporary APACHE model. The models have small numbers of variables and data are therefore easy to collect, with data less likely to be missing. Also, as the data required could be collected in a boolean form (0/1), data collection was made easier. It can be argued that the avoidance of choosing a diagnosis or treatment variable can make the system much easier to use as it avoids the difficulty of having to choose a single precipitating factor for ICU admission. The MPM_0 model, calculated on admission, is unaffected by the treatment received on the ICU.

Although collection of data is easier, one criticism of the model is that to answer some of the questions requires a certain amount of interpretation which may introduce bias into the measurement. Another is that the model has not been adequately tested outside the institution within which it has been developed. Therefore it is impossible to make judgements about its applicability in other settings.

2.4.2.7 Development of MPM II

In 1993 the authors published an updated version of the MPM model (MPM II) (4). The model used data gathered on 6 US adult general ICUs as well as data from the European and North American Study of Severity Systems used in the development of the SAPS II model. Data were collected from 137 ICUs from 12 different countries as well as the 6 US ICUs. The data from the 6 US units were divided by time, 65% were taken from the earlier part of the study to form a development set. The remaining time period making up a validation set. The patients from the joint study were randomly divided in a 65/35 split with 65% forming the development set and the remaining patients going to the validation set. The two development sets and the two validation sets were combined to form one development set and one validation set. Data were collected on 19,124 patients who met the study inclusion categories, with 12,610 patients in the development set and 6,514 patients in the validation set.

Table 2.4 Variables in MPM_0 model

Coma or deep stupor
Heart rate ≥ 150 beats/min
Systolic blood pressure ≤ 90 mm Hg
Chronic renal insufficiency
Cirrhosis
Metastatic neoplasm
Acute renal failure
Cardiac dysrhythmia
Cerebrovascular incident
Gastrointestinal bleeding
Intracranial mass effect
Age
Cardiopulmonary resuscitation prior to admission
Mechanical ventilation
Non elective surgery

Table 2.5 Variables in MPM_{24} model

Variables ascertained at admission
Age
Cirrhosis
Intracranial mass effect
Metastatic neoplasm
Medical or unscheduled surgical admission
24 hour assessments
Coma or deep stupor at 24 hour
Creatinine > 176.8 μmol/L (2.0 mg/dl)
Confirmed infection
Mechanical ventilation
Partial pressure of oxygen (PaO_2)< 7.98 kPa (60 mm Hg)
Prothrombin time >3 sec above standard
Urine output < 150 ml in 8 hour
Vasoactive drugs ≥ 1 hour intravenously

Two models were developed, one for patients on admission (MPM_0) and one for patients who remained in ICU at 24 hours (MPM_{24}). The MPM_0 model contained 15 variables (Table 2.4) all requiring yes/no answers. The MPM_{24} model was developed using 5 of the variables from the MPM_0 model which were significantly associated with mortality

for patients who remained in the ICU after 24 hours and 8 new variables significant over the 24 hour period. The authors used univariate analysis to select variables that were significantly associated with mortality before entering all significant variables into a multivariate logistic regression.

The authors reported good discrimination and calibration, with an area under the ROC curve of 0.824 and a P-value for the Hosmer-Lemeshow GOF statistic of 0.327 in the validation set for the MPM_0 model. The MPM_{24} model was reported to have an area under the ROC curve of 0.836 and a P-value for the Hosmer-Lemeshow GOF statistic of 0.231. The original MPM models were also applied to the data showing poor calibration with significant Hosmer-Lemeshow statistic ($P<0.001$) for both the MPM_0 and the MPM_{24} model. The authors argued that this demonstrated the need for regular updating of severity of illness models. The authors also demonstrated significant differences between patients remaining in the ICU at 24 hours and those patients discharged alive over the 24 hour period. This confirmed the argument for constructing two models for different time periods.

Lemeshow et al developed models for 48 and 72 hours from the same database (122). Using the same variables from the 24 hour model, the authors re-weighted the variables to reflect changing probabilities over time. Both models showed better performance over time for the patients remaining in at 48 and 72 hours, than that demonstrated by the MPM_{24} model on the same cohorts.

Application of MPM II: Like the APACHE II, APACHE III, and SAPS II models, the MPM II model has been used and evaluated in a number of studies (123-128). Results from analysis of the European and American multi-centre study showed that the older MPM models had poorer performance than the newer MPM II models (21). However, this is not surprising as these represented a large proportion of the patients from which the MPM II models were developed. Data from the EURICUS-1 study showed that the MPM_0 model had poorer performance than that described in the original study, with the MPM_0 model significantly overestimating mortality (114,129). Although the model had reasonable discrimination, with an area under the curve of 0.785, the calibration was poor, with a significant Hosmer-Lemeshow test ($P<0.001$). The model also showed poor performance when analysed in important subgroups, with significant differences in the mortality observed and that estimated by the model. (115).

2.5 Developments since the beginning of study

Funding for this study began in June 1994 with data collection beginning in January 1995. At the beginning of the study little had been done to consider the performance of these models outside the populations in which they had been developed. Since the study began there has been a considerable amount of work carried out and published in this field some of which has already been described in this chapter. This literature has not only concentrated on looking at the performance of models in different populations, but has also included customisation of the models described earlier, where performance has been shown to be poor. There has also been considerable debate as to whether these models can be used for detecting differences in the quality of care in different units. The following sections will attempt to describe the developments in the area of severity of illness scoring models since the beginning of this study.

2.5.1 Comparisons between models

There have been a considerable number of studies which have looked at the performance of the most up to date models we have described (APACHE II, APACHE III, SAPS II and MPM II) (68,83,100,101,111-113,130-136). Most of these have been small, or have described the performance in certain types of patients. There have been two large international studies that have compared the performance of these models. The first of these used the data collected as part of the European and American multi-centre study (21). The development of both the SAPS II model and the MPM II model were based on data from this study. The study used all the data available when comparing the performance of the older systems (SAPS, MPM_0, MPM_{24} and APACHE II) allowing the comparison on 12,802 patients. The newer models (APACHE III, MPM II, and SAPS II) were compared on the validation sample from both the MPM II and SAPS II studies, allowing the inclusion of 4,101 patients in the analysis. By using only the validation sample the authors were attempting to avoid the bias of using the development set, although the study only included the APACHE III score as the rest of the equation was not available. The study reported that the newer models had better performance then the older models. It concluded that

> ".....the new systems represent real improvement in severity model performance."

The authors go on to say that no one model is superior to another and that all the models in the study can be used with reliability. However, they stress the importance for testing models to check for adequate fit before being applied in different countries. However, the

MPM II and SAPS II models were developed on data taken from the same ICUs. As the development set was a randomly selected group it would be expected that the profile of the validation cohort would be very similar to that of the development cohort. It would be expected that the MPM II and SAPS II models would have good performance, especially in calibration. The other "new" model, the APACHE III, could not be tested for calibration as the score was the only part of the model available. As none of the older models were developed on this population it is not surprising that these models showed poorer calibration. A critical appraisal and systematic review of these models published in the same year (68) concluded

> "Direct comparisons of existing severity indices with respect to their calibration, discrimination, and reliability in different population sets are necessary in order for consumers to make informed choices between available models."

Teres and Lemeshow (137) in an editorial in Critical Care Medicine argue that there are only five studies (83,102,110,114,138) that have published enough detailed information on calibration and discrimination to allow any assessment on the performance of models in alternative settings to those in which they were originally developed. These, they argue, are consistent in showing that the new generation of models have good discrimination but poor calibration. One of these studies is the only other large study to have compared the performance of models in a large data set (114). This study is part of the large EURICUS-1 study collecting information from 89 ICUs within 12 European Countries. The data on this study have been published in a number of papers (114,129,139,140). Reporting results on data collected from 16,060 patients, they include comparisons of performance between the SAPS II and MPM II (MPM_0) models. The study shows better discrimination in the SAPS II model, which has an area under the ROC curve of 0.822 compared with the area under the ROC curve of 0.785 in the MPM_0 model. The calibration in both models has been poor with SAPS II having a Chi Squared value of 208.4 and the MPM_0 model having a Chi Squared value of 368.2. Both these values show significant differences between the observed mortality and the mortality predicted by the models ($P<0.0001$). The authors have also introduced the idea of uniformity of fit (115), where they argue that models must also have good performance in important subgroups. If models are to reliably identify quality of care issues then the adjustment for case mix must be adequate. This cannot be the case if certain subgroups with certain characteristics are poorly estimated. Other authors have pointed to this issue (73,86,110). The study showed in large important subgroups that there were significant differences between actual mortality and the mortality estimated by the models.

Data for the EURICUS-1 study were collected on ICUs from 12 different European Countries. The ICUs are not necessarily representative of the ICU culture in their countries. There is a wide range of severity of illness in the different ICUs used in the study and this is reflected in the range of estimated mortality from the different ICUs (SAPS II, 2.9%-34.8%; MPM_0, 3.9%-39.3%). This study has reported an average hospital mortality rate of 20.0% which is considerably lower than the 27.7% and 26.3% reported in other studies from the UK (20,83). There are two elements of these models that still remain unanswered:

- the performance of all the models (APACHE III, APACHE II, SAPS II, and MPM II) in the same independent large database.
- the performance of all the models in a large database in the same ICU culture or country.

So, although there has been considerable work in this area, these two areas of model performance still remain to be evaluated.

2.5.2 Customisation of models

A possible solution to the apparent lack of fit of these models when applied to different settings, may be to use a process of customising the existing models so that they better reflect the mortality experience in different countries. Lemeshow et al introduced the idea of customising models to adjust for groups of patients who show poor fit (4). Le Gall et al reported analysis from a study customising the MPM II and SAPS II models to provide better fit with sepsis patients (141). The authors reported improved calibration with an improved Hosmer-Lemeshow test for the new models.

Using data from six North American ICUs on 4,224 patients, Zhu et al investigated two strategies for customising the MPM II model (142). One strategy was to use the logit in a regression model to produce a re-weighted model, the other strategy involved re-weighting all the variables in the MPM II model. The study looked at the effect of customisation on changing accuracy of mortality of the MPM II model. This was done by simulating changes in mortality and analysing the effect this had on both the area under the ROC curve and the Hosmer-Lemeshow statistic. The authors then looked at the effect of the different customisation strategies and the effects these had on improving the accuracy of the mortality. The study concluded that both strategies could bring improved calibration to the model. However, the re-weighting of all the variables rather than just the logit was a better approach with a larger sample size.

There have been two large studies looking at the effect of customisation on the accuracy of different severity models. The first used data from the EURICUS-1 study (n=16,060)

and looked at the effect of the two strategies outlined previously on changing the accuracy of the MPM_0 model (129,139). The authors reported little change in discrimination in the validation set for either types of customisation. There was considerable improvement in the calibration using both strategies illustrated by improved Chi Squared values. Results from comparisons in the validation sample showed an improved Chi Squared value in the customisation of the logit only model of 133.76 to 21.68. In the variable model, the improvement in the Chi Squared value was from 133.76 to 20.94, with the lower Chi Squared value representing a better calibration. Both strategies resulted in models that had no significant differences between the observed mortality and the estimated mortality. However, despite the improvement in the model's calibration after customisation, the authors found insufficient improvement in the analysis of the uniformity of fit to overcome all the problems of applying the MPM_0 model to their data. Some subgroups of patients still had significant differences between the observed mortality and the mortality estimated by the MPM_0 model.

The other large study to have assessed the effect of customisation on the performance of a severity of illness model was the customisation of the APACHE III model in 10,929 patients from 86 Spanish ICUs (143). The authors reported good performance in the new model with an area under the ROC curve of 0.82 and non significant Hosmer-Lemeshow test (X^2=12.27). No results of the model before calibration were reported, except to say that the original American model had a slight underestimation of hospital mortality. No analysis of the effect on uniformity of fit was reported either. It is hard to assess the effect of customisation in this study as results for the model before customisation were not published.

2.5.3 To score or not to score?

There has been a debate as to whether it is possible to use severity of illness models or not. Although, this study is not designed to answer this question, there are some relevant aspects of the debate. The debate is often dichotomised, and there is a presumption that there is a yes or no answer. However, the issues are more complex than this and many authors express caution rather than a belief that severity models cannot be used.

It has been suggested that the severity models described above can be used to predict individual mortality. The reported accuracy of all these models would suggest that, as yet, this is not the case. Lemeshow et al (144) have said

"..models for estimating severity of illness in intensive care unit (ICU) patients, while demonstrating good agreement for describing patients in the aggregate, are shown to differ considerably for individual patients."

They go on to conclude

"This suggests that identifying patients unlikely to benefit from ICU care by using models must be approached with considerable caution."

Kollef et al claimed that most physicians did not rely on scoring models to make decisions about individual predictions (145). However, the models could be used to add evidence to the decision making process. This point was made by Knaus et al (6) who said that scoring models

".....can ensure that the experiences of the past are taken into consideration in an unbiased manner."

However, there would appear to be a general consensus that these models are not accurate enough to predict individual mortality (146)

In an editorial in the Lancet, Boyd and Grounds were among the first to question the use of scoring models for comparing ICU performance(147). With reference to APACHE II, the editorial highlights the effect treatment may have on the apparent performance of an ICU. The authors point to the fact that patients receiving poorer treatment but surviving will have a worse score than the equivalent patient who has received good treatment. The first patient's severity score will be higher because of poorer treatment. The authors conclude:

"The very accuracy of these scoring systems for assessing the severity of illness precludes their use for comparison and Audit"

On the other hand, a small study in Canada used APACHE II to flag patients with a low severity of illness who subsequently died (148). They compared this method against a 10% random sample in an effort to identify potential quality problems. Using a physician's evaluation as a gold standard, the study concluded that the performance of the APACHE II model in identifying problems of quality was much better than using a 10% random sample. Other studies have used severity models to highlight issues of quality. Wagner reported a New England study (149) that used a severity index to highlight problems of quality in cardiac surgery, allowing for a change in practice which

subsequently led to an apparent improvement in quality. A study by Zimmerman et al looked at data from nine units using the APACHE III model (94). The performance of the units was ranked using mortality ratios. Units were visited by a team of clinicians and organisational researchers who assessed the units on a number of aspects of the ward's performance. Areas thought to have a possible influence on outcome, like management structure, were assessed by the research team and then the ICUs were ranked according to how well the research team felt they had performed. The authors found that there was no correlation between the ranks produced by the team and those produced by the APACHE III model. The authors concluded that this showed clinicians were unable to assess quality of care. However, they could equally have concluded the same about the APACHE III model. The study's results were inconclusive and may have merely highlighted that there was little difference between the units.

Teres et al argued in 1993 that Intensive Care severity models should be used with caution (150). The authors suggested that, despite being based on larger databases and despite increasing accuracy, these models may not adequately describe important conditions (acute respiratory distress syndrome and multi-organ dysfunction).

Becker and Zimmerman (151) argued that

> "Any prognostic model that established a predicted hospital mortality rate for each ICU based on a representative data base and a patient-by-patient measurement of risk allows ICUs to compare their observed versus predicted outcomes".

They also argued that the mortality ratio (observed/estimated) provides an outcome based measure of the effectiveness of care. Concerns over the influence of prior therapy and treatment on the accuracy of probabilities, they argued, have not been supported by data or comparative studies. They also said, in the same paper, that increasing evidence would suggest that the mortality ratio is little influenced by prior therapy but gave no references to support this claim. They went on to state that too much time and effort is spent trying to "pass final judgement" on whether systems are "good" or "bad". The authors said that there is substantial evidence from existing studies that demonstrate the usefulness of these models.

Sherck and Shatney in the same journal argued that it was not possible to measure ICU performance or quality of care using these models (146). The authors pointed to four reasons why the reliability of these models in comparing ICUs should be questioned.

1. The methodological problems of earlier systems have not been corrected entirely by the new systems. Missing variables are treated as normal, which remains questionable especially in the case of the GCS. Choosing a single diagnosis can be difficult especially in more complex Intensive Care patients.
2. Some subgroups are poorly predicted.
3. Potential variations in patient mix precludes accurate comparison. Models have been shown that they do not transport easily to other cultures.
4. The models only evaluate mortality which is not in itself an accurate measure of quality. Other measures like quality of life may be more valid in ICU performance.

They concluded that researchers have been driven by pressures from regulators and the desire to predict mortality objectively and reliably. They also concluded that despite the advances in this area there is still no system that provides useful prognostic data.

Fidler, using data from the EURICUS-1 study, looked at the performance of 79 different ICUs (140). Using two different types of regression (multiple regression and fixed effects logistic regression) they looked at a number of factors and their effect on the standardised mortality rate. The author found five factors that affected ICU performance. These were:
1. Number of beds (optimal number is 9)
2. Organisational commitment
3. Results oriented culture
4. Elementary organisational framework
5. Country
They also concluded that units who performed well on high risk patients did not necessarily perform well on low risk patients and vice versa.

Those expressing caution in using these models for measuring ICU performance and more generally for case mix adjustment, have pointed to a number of weaknesses in the methodology of the models. Rowan argues that the reliability of case mix adjustment is determined by a number of factors (152).

- Input variability
    - Location of the study
    - Ascertainment bias
- Procedure variability
    - Exclusion criteria
    - Rules and definitions governing data collection
    - Time frame for data collection
    - Handling of data prior to analysis

- User variability

The author went on to say that before adjustment for ICU performance takes place these factors must be taken into account.

2.5.4 League tables

Work in other areas of medicine has looked at the validity of ranking hospitals or units into "League tables". It is not the purpose of this thesis to examine the validity of League tables but there is some relevance in the debate to measuring ICU performance. The Clinical Resource and Audit Group (CRAG) in conjunction with the Information and Statistics Division (ISD) of the NHS' Common Services Agency have, since 1993, produced four reports into clinical outcomes (153-156). These are based on data collected on individual patients for every hospital admission in Scotland. These include 30 indicators from different areas of practice. There has been fierce debate as to the usefulness of these 30 indicators with their publication resulting in a number of studies into the legitimacy of League tables (157,158). However, in a recent paper Kendrick et al argued that despite the issues surrounding the validity of these indicators they are the best knowledge we have on outcome (159). If their use is collaborative and comparative then clinicians are more likely to use the best evidence around to judge best practice.

Goldstein and Spiegelhalter, while acknowledging the need for establishing appropriate measures of institutional outcomes, emphasise the need to understand the limitations of the statistical methods used in comparing performance of institutions (160). They say that

> "...the continuing official publication and ranking of unadjusted scores lends any comparisons based on them an authority that they do not have."

However, they also say that

> "....comparative information about institutions can be useful if handled sensitively with due regard for all their problems, and that this must inform public dissemination."

Leyland et al developed measures of performance of maternal and neonatal care by controlling for case mix (161), concluding that these measures were valid and allowed comparison between hospitals or comparison with the data for all Scottish maternity hospitals. Marshall and Spiegelhalter using data from 52 in vitro fertilisation clinics assessed the extent to which clinics could be "reasonably ranked" by live births rates (158). They concluded that the ranks were an extremely unreliable statistic in indicating

performance or change in performance. Parry et al also tried to assess the reliability of crude mortality as a measure of performance in neonatal ICUs and concluded that, as an indicator, it was not very reliable as a measure of performance or best practice (157). They also stated that any use of annual league tables based on crude mortality could be, at worst, detrimental or at best just irrelevant.

Although this thesis is not about assessing the value of "League tables" it is possible that scoring systems analysed in this study may be used in this way. Also, the effect of inaccuracies in the models and their possible impact on ranking will be discussed in this thesis.

It is clear that before using these models to assess ICU performance, they need to be fully tested in the population in which it is intended to use them. Before applying these models to any Audit of Scottish ICUs, it is important to understand their possible weaknesses. Also, given the most recent studies in this subject, if the performance is poor then it is possible this may be improved by some sort of adjustment for the Scottish population.

Chapter 3-Methods and Patients

Contents:

3.1 Participating units

In 1994 , 24 Scottish general ICUs were identified and were invited to take part in the study. Eighteen general ICUs of whom 8 were in teaching hospitals and 10 were from non teaching hospitals collected data for the study. Four units characterised as High Dependency Units (HDU) / ICUs also took part and were within non teaching hospitals. These units are in small district general hospitals and admit predominantly patients who do not require organ support. However, they do have the capacity to admit limited numbers of patients who require organ support. Consultants in these units were asked to identify from the outset when a patient was being admitted for a purpose other than recovery monitoring. Ideally, all admissions would have been admitted to the study, however, it would not have been possible for such units to undertake the work involved in the scoring of all admissions. Thus, 22 ICUs in 22 hospitals collected data for a two year period from the 1st January 1995 to 31st December 1997. This represented all but three general adult ICUs known at that time. Two of these units agreed to take part in the study but wanted to maintain their own software. However, due to the difficulty in changing their existing software these units were unable to provide the study with any data. A further unit was identified and has since been invited to join the Audit, but was not recruited in time for this study. All patients were followed up to hospital discharge to ensure that those HDU/ICU patients not entering the study had a mortality consistent with that expected for an HDU population .

3.2 Audit personnel

The Audit was co-ordinated from an office in the Victoria Infirmary anaesthetic department and had two full time staff (Table 3.1). The day to day running of and management of the Audit was undertaken by the author. A Research Nurse was employed to undertake a validation of the quality of the data collected for the study. The rest of the Audit's work was undertaken by volunteering medical and nursing staff. A consultant anaesthetist based in the Victoria Infirmary was responsible for the overall management of the project. Each unit designated a consultant with local responsibility for management of the Audit. A list of these can be seen in Appendix 1. The right to use the APACHE III model and software was purchased from APACHE Medical Systems Inc (AMS).

Table 3.1 Audit Personnel

Full time Audit Staff	Responsibilities
The Author	Day to day management of
(PhD studentship)	Audit
Research Nurse	Validation of quality of data
Volunteering personnel	
Consultant	Overseeing of Study
Local Consultant	Local management

3.3 Installation of computers and software

A computer, printer, modem, and software were installed on 21 of the participating ICUs. The remaining ICU chose to update their existing PC-based software to collect the dataset required for the study. A training package for the installation of the computers, use of the software, and implementation of the rules for data collection was developed. This training package was then piloted, changed and re-piloted before a final package was decided. This process and the issues around training have been reported elsewhere (162). In the pilot study training had taken place over a three day period. This was considered impractical as the time it would have taken to install all the computers would have seriously curtailed the data collection period. It was therefore decided to install computers and train staff in one day.

3.4 Computer training.

The training protocol was as follows:

Preliminary visit to unit: All units were given a preliminary visit, to arrange a date for the installation and training in the use of the computers and software, and to allow the discussion of a number of issues regarding the study, i.e. who would collect which data, who would be responsible for the management of the Audit, the need for a separate modem line and where the computer was situated (preferably on the unit). Although the study was being administered by the Scottish Intensive Care Society, a predominantly medical organisation, a member of both medical and nursing teams was requested to be present at this meeting.

Computer set up: Computers and software were set up for each ICU before being delivered to the ward. This meant customising software for each unit to include a bed plan of their unit, lists of admitting consultants and wards, and the setting up of password protection.

Installation of the computer package: On arrival at the ICU the computer, modems, and printers were installed on the unit.

Basic training: Training to use the computer and the software for basic data collection was given. This included computer based training on the reason for and purpose of the Study. This training was given in groups of two (preferable) and in no more than three. The training was interactive with the users being encouraged to try inputting data to the software.

Training of Audit Managers: It was recommended that there would be two Audit Managers, one from the medical team and one from the nursing team. ICUs were encouraged to choose senior staff for these roles. The same basic training was followed by more intensive training on the management of the Audit software and the rules for collection of the data. It was the responsibility of the Audit Managers to then teach the relevant staff on the units. Audit Managers were encouraged to see how the system could be used to meet the information needs of the individual ICUs and by so doing it was hoped to instil a sense of ownership in the data. The ICU that maintained its existing software was also given training in the data collection rules.

Follow up services: ICUs were encouraged to start collection of data from the day after the installation of the computers. The Audit Staff ensured that there was someone available by telephone for the two days after training days, to allow any questions to be answered. We also provided a phone service throughout the study to answer questions about any aspect of the data. This included any software problems that might arise. All computers on each of the ICUs had software that allowed a remote connection to be made and the ICU's computer could be viewed by us, allowing further explanation and training to be carried out from the central Audit office. As part of a data validation programme (see Chapter 4) the Research Nurse returned to each ICU once a quarter to allow staff an opportunity to talk about any problems they may have been experiencing. Manuals detailing the study protocol and data collection rules for the APACHE III model were left on all the wards (Appendix 2). Where issues of ambiguity arose, confirmation of the correct procedures or rules was sought from AMS.

Only data collected from 1995 onwards were used which meant that most ICUs had been using the software for a month or more before any of their data were used.

3.5 Ward Watcher Software

Data were collected on 21 ICUs using the Ward Watcher Software, which was provided under licence by AMS, with one ICU using PC-based software developed by Dundee University, Computing Science department. The Ward Watcher program is a comprehensive ICU Audit and clinical data system. The software was commercially available and was licensed by AMS for the collection and calculation of the APACHE III model. The software was considerably altered to collect information on all the models in the study, except the UK APACHE II model which was calculated retrospectively. On the one unit retaining their own software, data were downloaded from the system once every quarter. The data were imported into a Ward Watcher database at the central Audit office where the probabilities for all of the models were calculated.

There were certain advantages of all but one of the units using the Ward Watcher software. These were as follows:

- ICUs recorded the data in the same format and under the same data entry criteria which gave a consistency not available if all units collected on different software.
- Supporting the data collection was less problematic as ICUs were using the same computer set up.
- All rules for data collection were included within the software at the particular data entry point. When in a data entry area, users could click on a question mark to show the relevant rules.
- Range checking asked users to confirm values when they were at the outer limits of possible values.
- The rules for choosing the correct value in the different models are complex, so to reduce mistakes users entered the lowest and highest values. The value which was used to calculate the probabilities was then chosen by the software.
- Although primarily for the collection of data for the study, the software was a powerful clinical system allowing the users to collect a considerable amount of other data which was of specific interest to individual ICUs.
- The system also included easy-to-use search and report facilities allowing users to interrogate their own data.
- Providing units with a computer, a colour printer and clinical Audit software would, it was hoped, help to foster a feeling of ownership in the data and increase the possibilities of units collecting good quality data.
- The software was "user friendly" allowing easy to follow data entry (Ward Watcher screens can be seen in Appendix 3). This it was hoped would reduce the time it would take to learn how to use the software and reduce the amount of data entry errors.

3.6 Model calculation

Data were collected to generate scores and predictions for APACHE II, APACHE III, SAPS II, MPM_0 and MPM_{24}. Data collected for APACHE II and APACHE III used the original published protocols (5,6). Protocols and procedures were published in a data collection manual (Appendix 2) and in help screens on the Ward Watcher software (Critical Audit Ltd). SAPS II and MPM II were collected in accordance with the originally published papers (3,4).

APACHE II diagnoses were generated from APACHE III diagnoses using mapping provided by AMS (Appendix 4). APACHE II mortality predictions were derived using both the original coefficients and those generated by the UK APACHE II study (13). This required the author to map original APACHE II diagnostic groups to UK APACHE II diagnostic groups (Appendix 4). Predictions for all models were calculated using the Ward Watcher software.

3.7 Subjects

Data were collected on consecutive patient admissions from 1st January 1995 to 31st December 1996 in all ICUs and, as previously described, for ICU/HDUs (3.1). However, patients with burn injuries, patients under 16 years of age, patients who died in the first hour after ICU admission, patients admitted to die and who died in the first four hours, and patients in full cardio-pulmonary arrest who died in the first four hours were excluded from the study. The criteria for these exclusions were determined by similar exclusions used in the development of the original models (3-6).

3.8 Outcome measure

In keeping with the original studies the outcome used was hospital mortality, defined by the patient's status when discharged from hospital.

3.9 Data Analysis.

3.9.1 Discrimination

Discrimination is defined as the ability of a model to identify survivors and non-survivors (163). One measure of discrimination is the area under the Receiver Operating Characteristic (ROC) curve (164). The ROC curve is constructed by calculating sensitivity and specificity for a discrete number of cut-off points. Sensitivity is then plotted against 1-specificity and the resulting points, when joined, form the curve. The higher the number of cut off points the more accurately it is possible to draw the curve. The better the discrimination then the closer the area under the curve gets to 1. An area under the curve of 1.00 would indicate perfect discrimination with an area of 0.50 representing random chance (Fig 3.1). The most intuitive description of what the area under the curve represents would be, if you randomly chose a survivor and non survivor then the area under the curve represents the probability of the non survivor having a higher probability of mortality than the survivor. (4,13). ROC curves are only a measure of discrimination and give no sense as to how accurate the probabilities generated by the different models are (165,166). ROC curves in this thesis have been calculated using a 100 cut off points. The DeLong method for comparisons of the areas under the different ROC curves was used. The DeLong method is a non-parametric method for comparing the areas under the curves of two or more correlated samples (167). All ROC curves and comparisons of the areas under the curves were calculated using AccuRoc Non-parametric Receiver Operating Characteristic Analysis Version 3.0 (Dr S.Vida, McGill University).

Another means of measuring discrimination is by use of a classification table (168). This is generated by counting the numbers of survivors and non-survivors above and below a certain cut off point of estimated mortality. The sensitivity, specificity and correct classification rate can be compared in different models. Sensitivity is the proportion of observed deaths predicted to die, the specificity is the proportion of observed survivors predicted to survive and the overall classification rate is the proportion of patients correctly classified as survivors or non survivors (Table 3.2). Although this is a means of testing discrimination, and the value of it can be extended using different cut off points, the weakness in this approach is that it is possible to have good discrimination but for this not to be reflected in the classification table depending on which cut off point is used. It is also difficult to see what information is gained from classification tables that is not already described by an ROC curve. Hosmer and Lemeshow give a good description of the classification table and its weakness (168).

Figure 3.1 ROC curve for APACHE II model

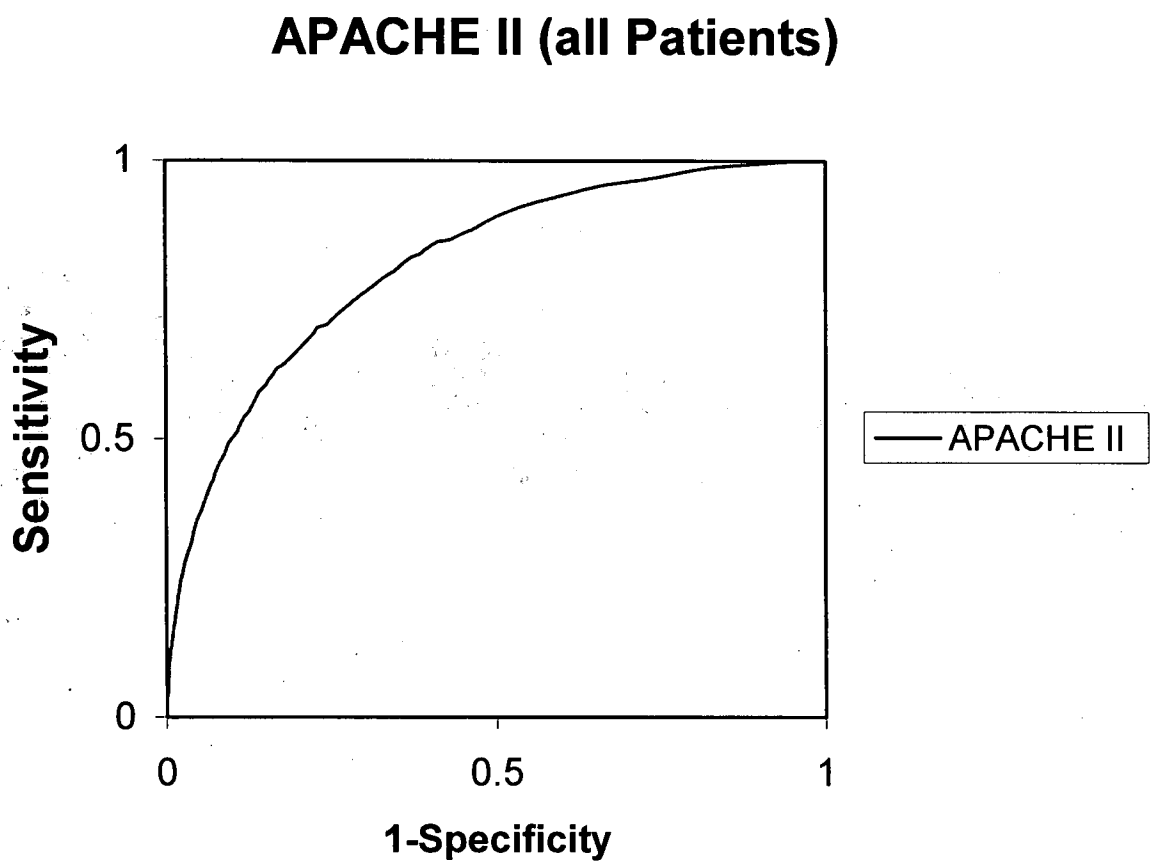


Table 3.2 APACHE II Classification table

Decision Criterion 50%	Predicted Alive	Predicted Dead	Total
Observed alive	6288	740	7028
Observed dead	1440	1380	2820
Total	7728	2120	9848

Overall correct classification: **77.86**

3.9.2 Calibration

The other area of assessing a model's performance, as previously mentioned, is calibration. Calibration is used here as an evaluation of the degree of correspondence between estimated mortalities and observed mortalities. There are a number of different ways of assessing the calibration of a model but the following are the most commonly used in the assessment of Intensive Care scoring models.

Hosmer-Lemeshow Goodness Of Fit (GOF) test: The Hosmer-Lemeshow GOF test creates 'bins' of increasing risk. There is no set number of bins but the number is usually 10. With large numbers it is possible to increase the threshold for obtaining a significant result by increasing the number of bins. Both the observed number of survivors and non-survivors and the predicted number of survivors and non-survivors are calculated for each mortality risk group. The Chi Squared statistic which compares the observed with the expected frequencies is calculated for each of the bins both for survivors and non-survivors using the following equation:

$$\sum((O-E)^2/E)$$

Where O = Observed number of patients
E = Predicted number of patients

A P-value can then be obtained from a one-sided Chi Squared distribution table (169). There are two approaches to the Hosmer-Lemeshow test. One approach (H) is to group the patients by a cut-off point of risk, i.e. all patients with a probability less than 10% and then all patients with a probability greater than 10% and a probability less than 20% and so on. The alternative approach (C) is to create equal size bins with increasing risk. So, for a sample size of 100, the 10 patients with the lowest risk would be in the first bin, the 10 patients with the next lowest risk would be in the second bin and so on. Hosmer and

Lemeshow have validated both these approaches but have reported their preferred use of the C statistic. (169) (personal communication). In this study, the C statistic has been used, as grouping the patients into equal bins of equal numbers would appear to be the most legitimate method. The probabilities from these models produce a distribution with relatively low numbers of patients who have high probabiities of death and smaller numbers in the higher bins. The H statistic gives equal weight to each of these bins. Any differences in the probabilities at the higher end of severity will mean that a relatively small number of patients have a disproportionate effect on the test.

Calibration curves: Calibration curves give a graphical representation of calibration. They are produced by grouping patients into bins of predicted mortality of equal width, plotting the observed mortality for each decile and comparing this plot to the line of equality. A frequency distribution plot should be included which allows areas of discrepancy, between observed and predicted mortality, to be related to the distribution of illness severity within a study population. Confidence Intervals (CI) for the observed mortality should be included to allow statistical assessment of the level of discrepancy. This has been done rarely in the intensive care literature but without CIs it is hard to make any real assessment of differences between observed mortality and that estimated by the model. This method has been criticised by Lemeshow because of the unequal distribution of patients in the different bins overestimating the importance of patients with high probabilities in the model (137) (personal communication). While this may be true, calibration curves offer a graphical representation of these data and the frequency distribution and the CIs allow the viewer to assess the imbalance in distribution. It is also the most intuitive way to view these data as you see the full range of severity in a regular scale (0-100). CIs for observed mortality are calculated using the following equations (170):

where p = the percentage of observed mortality
SE= standard error of p.
n= sample size

$$SE= \sqrt{(p(100-p)/n)}$$

$$p-(1.96*SE) \text{ to } p+(1.96*SE)$$

3.9.3 Mortality ratios

Mortality ratios (observed mortality/expected mortality) have been used to assess differences between units and different subgroups in severity models. However, there are a number of statistical issues around the use of mortality ratios and it has been recommended that caution should be used in drawing conclusions from apparent differences (171). To assess differences using mortality ratios, comparisons have only been made between ICUs/Subgroups and the whole cohort rather than between separate ICUs/Subgroups. Also, statistical significance was accepted only when CIs for the ICU/Subgroup lay outside the CIs for the whole population. Although there is no stated numbers above which mortality ratios can be legitimately used, comparisons have only been made when the numbers were greater than 300.

Another problem in using mortality ratios is calculating CIs. As most studies do not publish the variance from the logistic regression analysis it is not possible to use the variance of the model to calculate CIs. However, Hosmer and Lemeshow have suggested a method for calculating mortality ratios (172). They have shown that this method is comparable to other methods which use information taken from the output of the logistic regression model. This method has been used in practice by Rapoport et al (125) and the equation is as follows:

$$CI=(O \pm 1.96*\sqrt{Var})/E$$

Where O = observed number of hospital deaths

E = expected number of deaths

$Var = \sum(p_i*(1-p_i))$ where p_i is the estimated probability for patient i

3.9.4 Subgroups

The proposed use of these models is to adjust for case mix. It is important that different subgroups of patients representing different case mix should show comparable performance. When comparing the performance of models in subgroups it would not be appropriate to use ROC curves or the Hosmer-Lemeshow GOF tests. ROC curves will quite legitimately demonstrate differing discrimination as subgroups may represent different levels of severity of illness. Hosmer-Lemeshow GOF tests are sensitive to the numbers involved in the test, with a significant test more likely if the numbers are large (142). Therefore, it would be inappropriate to compare performance in subgroups as the numbers in different groups would be different. It seems more appropriate to compare subgroups using other methods. This study has chosen mortality ratios and their CIs, the

CIs for observed mortality, and Chi Squared tests to assess performance in subgroups. The Chi Squared tests have been calculated using the following equation (173)

$$X^2 = \sum((O-E)^2/E)$$

Where O = Observed number of deaths
E = Predicted number of deaths

3.9.5 Statistical Packages

The above analysis was carried out using a number of statistical packages. All ROC analysis were carried out using AccuRoc Non-parametric Receiver Operating Characteristic Analysis Version 3.0 (Dr S.Vida, McGill University).

Hosmer-Lemeshow GOF tests, mortality ratios, all CIs and Chi Squared values were calculated on Microsoft Excel version 4.

Bland Altman plots were graphed using Minitab version 10. All other graphs were generated using Microsoft Excel version 4.

CHAPTER 4-Data Variability and Quality

Aims: To measure the quality of the data collected in the study, assessing it is fit for use in these models

To discuss possible factors that may affect both the quality and the variability within the data.

Contents:

4.1 Introduction

The accuracy of the probabilities generated by the scoring models depends on the quality of data collected. The study attempted to ensure that data were both accurate and complete. These measures have already been described in chapter 3. In an effort to minimise on data errors, the study:

- collected high and low values
- had range checking in the software
- would not allow patients to be discharged from the system until the minimum dataset was complete
- spent a day on each ward installing the computer and explaining data collection
- a comprehensive data collection manual was supplied to all units
- answered questions by telephone on data collection on a nine-to-five basis and provided a 24 hour answering machine for questions outside those hours
- collected hospital outcome every three months, units being reminded about missing data until information was correct.
- re-scored 10% of all records in the first year and 5% of all records in subsequent years.
- provided feedback to units about where errors were occurring

However, it is clear there are a wide number of issues regarding scoring models and data quality. Some of these issues the study has attempted to address, others it was not possible for this study to tackle.

4.2 Validation Study

Aims: To monitor the effect of errors on the quality of data.

Sample: A 10% systematic sample of all records from each unit was re-scored by the Research Nurse.

4.2.1 Methods

Patients were selected by taking the first patient in every three month period and then every tenth patient thereafter. In addition to the 10% sample a further 5% was requested in the event that the original case notes were unavailable. These were selected by choosing the 18th patient admitted in the quarter and every 18th patient thereafter until the required number of extra notes had been identified.

A database was designed and constructed using the Ward Watcher data screens (Appendix 3). This allowed the Research Nurse to enter the variable values in similar conditions to the original scorer. Values from the original scorer were withheld from the Research Nurse to allow re-scoring without bias. For variables contributing to the APACHE III, II and SAPS II physiology scores, the effect of the value entered by the Research Nurse on the score was calculated by the software and then compared to that of the original scorer. When changes in any of the three physiology scores (APACHE III, APACHE II and SAPS II) were detected an error screen was then produced. The screen informed the nurse of the differences and the original scorer's value whereupon she would re-check the value and re-enter what she believed to be the correct value. Differences in values entered into the MPM II fields and all other fields for other models were compared directly which prompted the nurse to check the value when differences were found. Forcing the Research Nurse to re-check when differences were found, allowed her to correct any mistakes she may have made, and made her as close to a "Gold Standard" as was possible.

Data were exported from the validation data base into a 'dummy' Ward Watcher and probabilities (%) generated for both the original and the re-scored data.

4.2.1.1 Data analysis

Student t-tests were used to test for any significant bias in the differences between the probabilities (%) generated by the original scorer and the Research Nurse (174). The t-tests allowed for any significant over or underscoring to be highlighted by testing if the mean differences are significantly different from zero. The differences in the probabilities (%) were also plotted against the mean of the two scorers' probabilities (%), with the lines representing ± two standard deviations (SD) also being plotted. This plot allows the assessment of the relationship between the differences and the mean. Ideally, for high and low probabilities (%), there should be no pattern of under or overscoring. We would expect that 95% of the observations will fall within ±2SD (175).

4.2.2 Results

Over the first nine months of the study 3,390 patients were admitted to the study and of these patients, 339 were re-scored by the Research Nurse. Of those patients re-scored, 54% were male and 46% were female.

All models except MPM_0 showed no significant difference from an expected mean of zero (APACHE III, P=0.12; APACHE II, P=0.69; MPM_{24}, P=0.23; SAPS, P=0.0.91). MPM_0 had a mean difference of 2.16 which showed a significant overscoring (P=0.014) (Table 4.1).

Table 4.1 Comparison of original scorer and Research Nurse for each model

	Mean difference (Original Scorer-Reseach Nurse)	Significance of t-tests P =	+2SD	-2SD
APACHE III	-1.29	0.12	28.87	-31.45
APACHE II	-0.31	0.69	27.69	-28.31
SAPS II	-0.09	0.91	27.31	-27.49
MPM 24	0.70	0.23	22.24	-20.86
MPM 0	2.16	0.01	34.22	-29.32

SD, Standard Deviation; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Model.

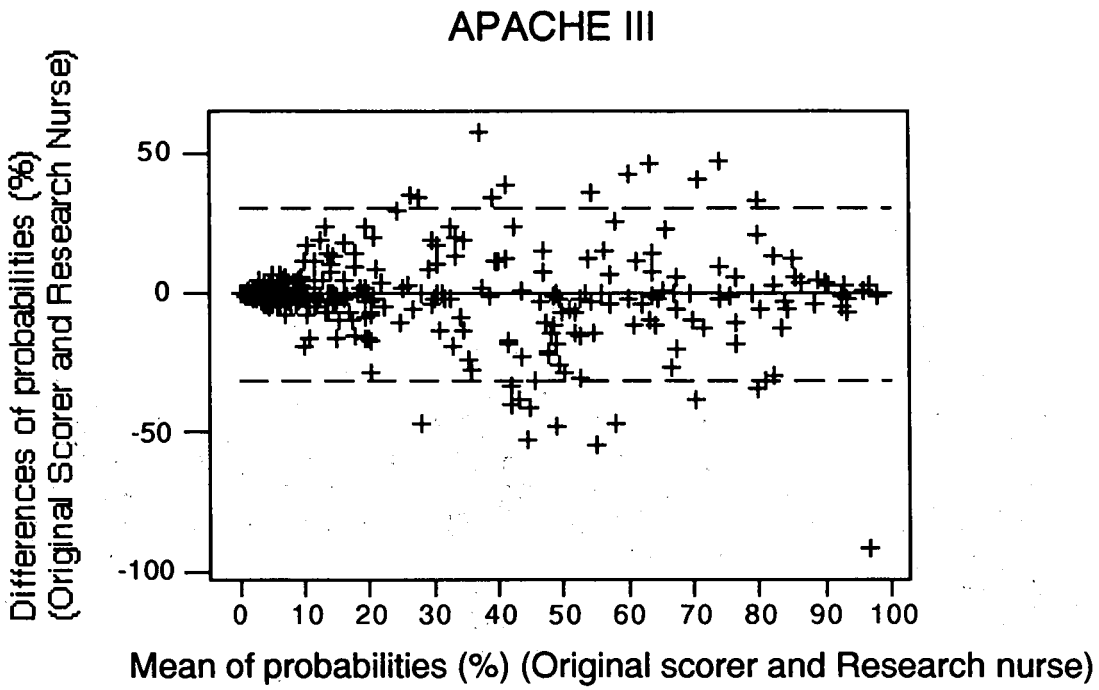
Plots of the differences versus the mean of the two probabilities (%) show no signs of over or underscoring (Figure 4.1 - Figure 4.3). There would appear to be no relationship between the level of probability and any over or underscoring. Using ±2SD, 95% of the differences are expected to fall between 34.22 and -29.32 at worst (MPM_0) and between 22.24 and 20.86 at best (MPM_{24}) (Table 4.1).

4.2.3 Discussion

The results from the re-scoring of 10% of all records in the first nine months showed that there was no significant over or underscoring, except with the MPM_0 model, when the original probabilities (%) were compared to those generated by the Research Nurse. Although the Research Nurse returned to the ICUs as soon as possible, it was typically 3-4 months after the patient's admission. This may have meant that the information in the notes was different to that available the original scorer.

It was believed that those involved in the data collection would not attempt to inflate the predictions by deliberately inputting incorrect data ("gaming"). The results seem to support this belief with no perceptible overscoring except in the MPM0 model. The overscoring in the MPM_0 model appears to be as a result of the incorrect choice of the coma variable. This is believed to be the result of a looser interpretation by the scorers of the term "coma" than was originally defined in MPM II paper (4).

Figure 4.1 Plots of differences (Original Scorer - Research Nurse) against the mean of the two probabilities (%) for the APACHE III and APACHE II models. Each cross represents one patient (n=339). Broken lines, ±2SD.



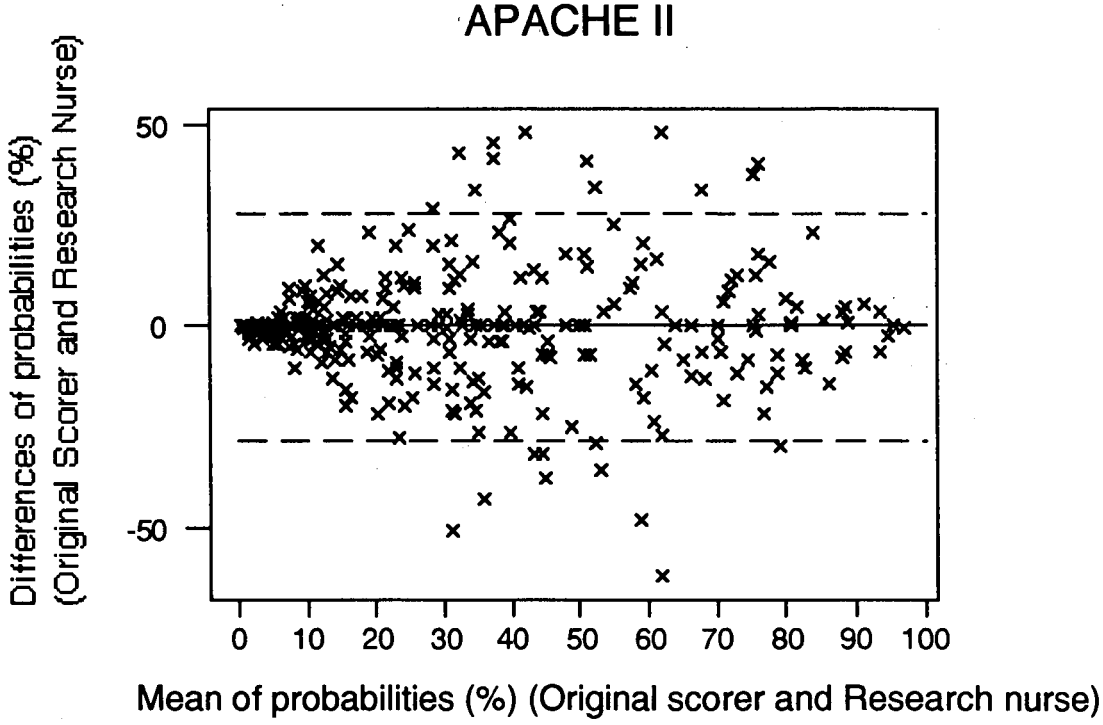
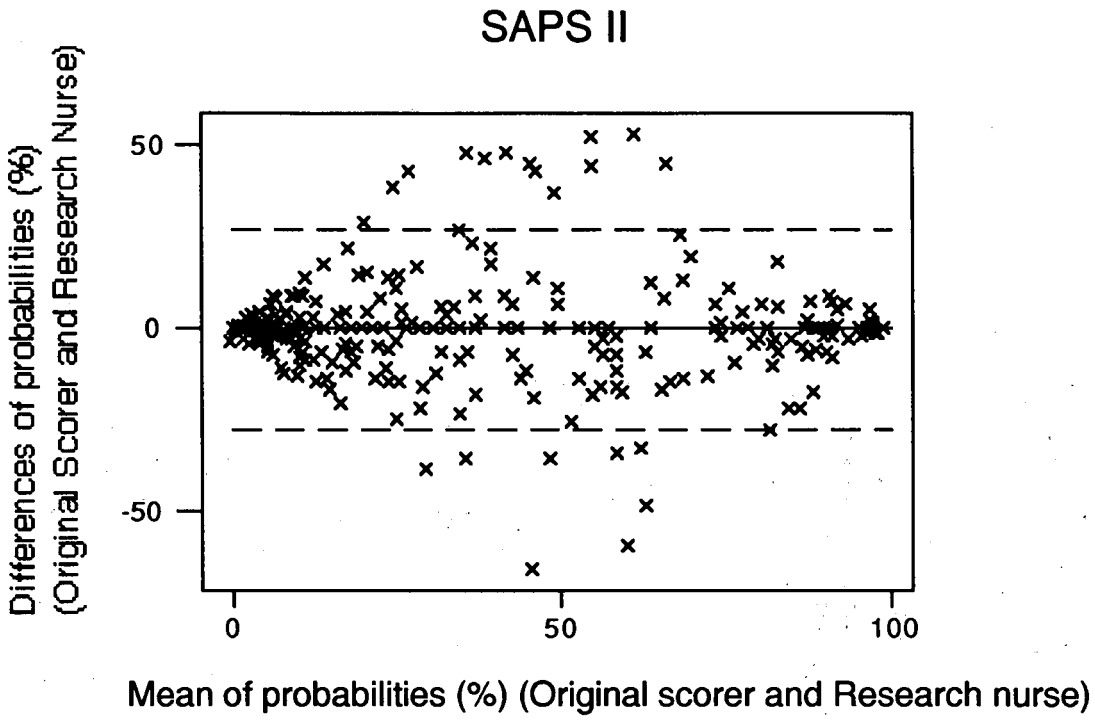


Figure 4.2 Plots for the SAPS II and MPM_{24} models of differences (Original Scorer - Research Nurse) against the mean of the two probabilities (%). Each cross represents one patient (n=339). Broken lines, ±2SD.



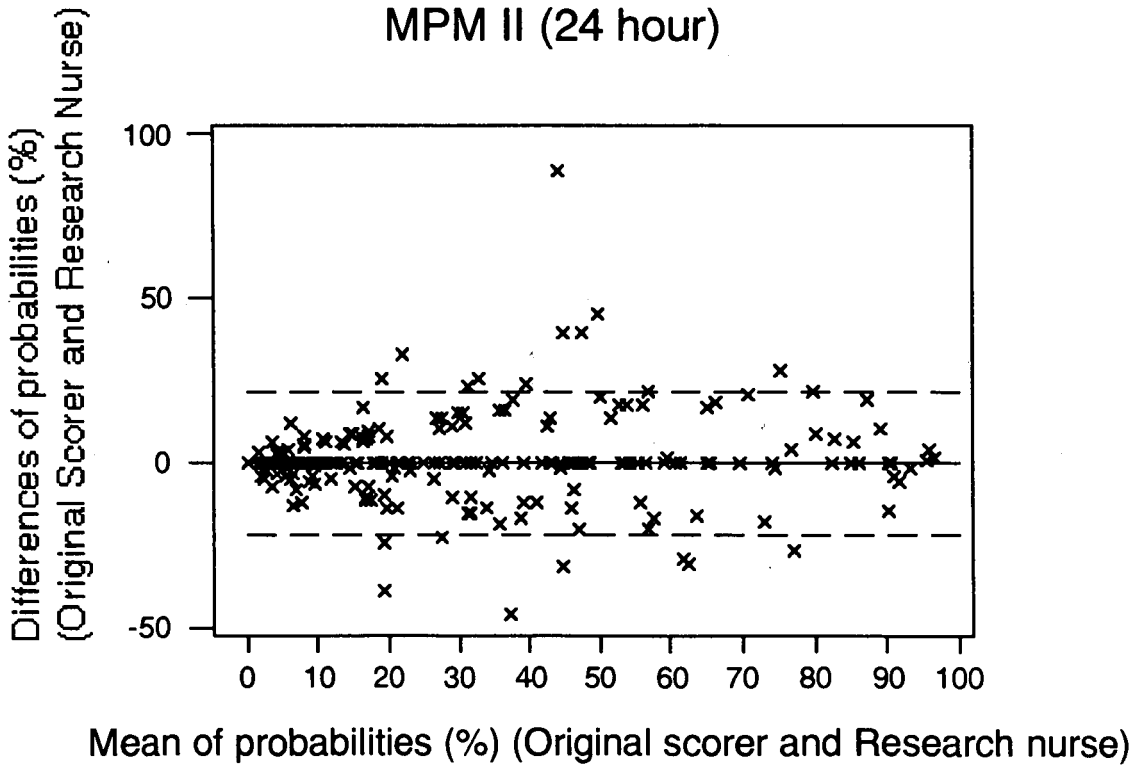
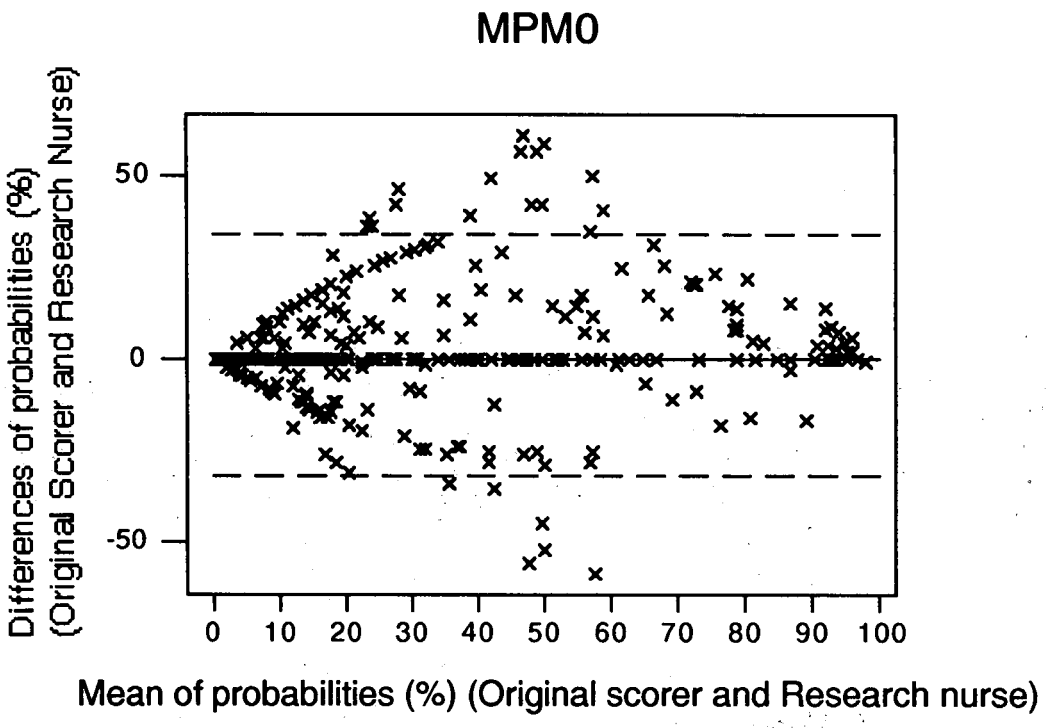


Figure 4.3 Plot of differences (Original Scorer - Research Nurse) against the mean of the two probabilities (%) for the MPM_0 model. Each cross represents one patient (n=339). Broken lines, ±2SD.



With the exception of MPM_0 the size of the differences between the original scorers and the Research Nurse are largest in those models with the most variables i.e. APACHE III (SD ± 15.08), and smallest for those with fewest variables i.e. MPM_{24}. It is clear that the more variables a model has, then the more opportunity for errors there are, and as MPM_{24} has only 13 boolean variables with all but age requiring only yes or no answers it is not surprising that the standard deviation is smallest in this model (SD± 10.78).

How small the differences between the Research Nurse and the original scorers should be, is not a question of statistical significance but one of judgement (173). There would appear to be quite considerable variation in the differences with 95% of the differences falling within +34.22 to -29.32 for the MPM_0 model. Although this model is the worst, all the models have considerable variation. Differences of 20-30% on the estimated mortality could make a considerable difference to calibration if concentrated in a certain group of patients. To assess the effect that errors made to the overall performance of the models, it would be important to look at the differences they made to the measures of the performance, like ROC curves and Hosmer-Lemeshow tests. The numbers in this sample were too small to allow these tests to be carried out accurately.

Whilst the study found no significant over or underscoring in the overall population, it was possible that some ICUs made more errors than others. It was not possible to test whether this was the case as the numbers of re-scored records in the individual ICUs were too small (mean = 15.4). However, the evidence from the overall sample would suggest there is nothing to support the hypothesis that errors were making the estimated mortality of individual ICU's significantly higher or lower. There are no data at the moment that would allow this study to determine where these errors predominantly occurred, which ICUs made most errors, and in which variables errors were most common.

Although an attempt was made to ensure that data were collected in a uniform and accurate way, there were areas where it was uncertain if there were any significant effects in the data. As part of the re-scoring process ICUs were informed of the errors they had made and the possible effect these errors may have had. It is unclear whether this information had any effect on the practice of recording data. It was not possible to determine how data were collected as this would have been seen as infringing on the autonomy of the clinicians. Where ICU treatment was begun, how often data were recorded, what patients were admitted to ICU or selected for entry into the study by

the 4 HDUs, were only a few of the decisions that clinicians determined individually. However, the knowledge that errors made in scoring cause no significant over or underscoring is reassuring and given a lack of data to answer other questions in this area it is presumed that the data were accurate enough to be used and were "fit for purpose".

Although for this study we have accepted that the data we have is "fit for purpose", there are still a number of areas where we have no information and further work is needed. There is no record of the number of case notes not available from different units. This could be a bias in the validation study as it would have been possible for units to withhold case notes on patients where it was known there was a problem. At present the study has no information on missing case notes, however, it is thought unlikely that any case notes were deliberately withheld. It also has to be acknowledged that it would have been a simple process to have monitored the unavailability of case notes and any future work should do this.

As has already been stated, there were not enough patients involved in the validation study to look at the difference in the error rates in the ICUs. Data from the validation study shows us that there is potential for considerable differences in probabilities caused by errors. If poor data collection or recording of the data for the models is concentrated on certain units, then errors might affect the apparent performance of a unit. Although it was hoped that by informing each unit of the effect of their errors this would improve their data quality, there is no evidence that this was the case. Any future work in this area should involve the collection of enough patient data to allow the assessment of errors on the different units and assessment of the effect of reporting errors to the units.

These models, especially those that are physiologically based (APACHE II and III, and SAPS II), have a considerable number of variables. It is probable that some variables are more susceptible to errors than others. The reason for the significant overscoring seen in the MPM_0 model has, anecdotally at least, been blamed on the inaccuracies of the Coma variable. As variables in the models have unequal weight, the effect of errors in the collection of some variables will be larger than others. Any future validation should be able to analyse if specific variables were more prone to errors and also their effect on the probabilities generated by the models.

Data collection rules: The study attempted to ensure that the same rules for data collection were applied by all the units and scorers. After the computers were installed on each of the ICUs, a day was also spent explaining the rules for data

collection and some of the issues surrounding them. The rules and definitions for the data collection of APACHE III were provided in a manual (Appendix 2). Rules and definitions for the other models in the study were also available through help screens in the Ward Watcher software. Any ambiguities in the rules or definitions could be clarified by phoning the central Audit office. However, it was impossible to force staff to use manuals or help screens and it was presumed that time and workload pressures on the ward would have some effect on the data quality.

Data collectors: The staff who collect the data may be an important factor in the quality of data. At the beginning of the study there was no apparent evidence to suggest who would best collect the data required for the severity models. However, Holt et al have since reported that inter observer variability in large patient groups has minimal effect on predicted mortality (176). They also reported that significant variance may occur at an individual patient level. They reported that, at this individual level, residents scored patients more accurately than nurses. In this study, at an anecdotal level, the Research Nurse who re-scored the patients, reported that it was nearly always medics who recorded the severity of illness data. She also reported that the quality of data was best when recorded by fewer people. It was not possible for the Audit to decide who would collect the data on each of the ICUs. Different ICUs faced different pressures with regard to the time and the resources they had available for collecting the information needed for the Audit. It was therefore left to the individual ICUs to make the decision as to who was best able to collect the severity of illness information. However, it would have been possible to record who was collecting the data on each of the ICUs. This would have allowed some analysis of who most accurately collects information and if certain groups of staff are better at recording this information than others.

Amount of data and frequency of sampling: Another potential effect on both the quality and variation of data was the amount of data routinely recorded on each of the units. Although the study had a minimum data set (Appendix 3) some data were not necessarily collected for each patient. Some units may routinely have collected certain tests, while other ICUs may have only ordered these tests when they felt they were indicated. There was an increased possibility of highlighting extreme values if clinical tests were carried out routinely rather than when clinicians felt they were indicated. ICUs in this study had differing policy depending both on the clinician's views within the unit and budgetary constraints. The study was not in a position to recommend when these tests should be done and it was therefore left to each ward to determine their own policy.

Also, some units may record observations more frequently than others. With the increase in computer technology in ICUs some monitoring systems allow the storage of continuous data for review at a later point. If data are available for the whole of a 24 hour period rather than sampling every 15 minutes there is increased potential for recording abnormal values (177).

Any future validation study should profile ICUs, looking at where ICU treatment starts, how data were collected, who collects it, and the effects of these processes on the accuracy of the data. If data from different units is to be consistently collected then it is important to understand the effect, if any, of these factors.

Chapter 5- The models' overall performances.

Aims: To compare the overall performance of the models in a Scottish setting.

Contents:

5.1 Results

5.1.1 Intensive Care Units

The ICUs in the study varied in size and patient turnover with a range of patient admissions, during the study, of 159 to 956. This represents a range of 1.53 to 9.20% of the patients admitted to the study. The ICUs in the study had a mortality range of 16.97%-42.31% and an APACHE II score (a proxy for severity of illness) range of 14.34-21.77. The percentage of operative patients on the different ICUs was in the range of 24.68-73.26 (Table 5.1).

Table 5.1 Profile of Intensive Care Units collecting data for the study.

	Patients[a]	% of overall admissions	% Mortality	Mean APACHE II Score	% of operative patients
A	389	3.74	17	14	43
B	609	5.86	19	16	52
C	394	3.79	21	16	50
D	456	4.39	22	17	57
E	414	3.98	22	16	67
F*	309	2.97	23	16	64
G	585	5.63	27	19	41
H	390	3.75	28	17	36
I	444	4.27	29	18	59
J	304	2.93	29	18	33
K	814	7.83	29	19	39
L	404	3.89	31	19	59
M	513	4.94	32	20	25
N	672	6.47	33	18	68
O*	268	2.58	33	19	73
P	472	4.54	33	19	32
Q	470	4.52	33	18	63
R	767	7.38	35	20	41
S	396	3.81	35	21	58
T*	159	1.53	36	20	47
U	956	9.20	37	20	39
V*	208	2.00	42	22	49
Total	10,393	100	29.4	18	48

APACHE, Acute Physiology and Chronic Health Evaluation;

*ICU/HDUs; a. Number of patients in study after exclusions.

Those units identified as ICUs/HDUs were analysed to assess the impact of excluding HDU type patients from the study. The mortality of HDU patients varied from 3.44%-9.87%. The mortality in these units for those patients who have been included in the study had a range of 23.30%-42.31%. This compares with a mean 29.09% in the

other ICUs in the study. In three of the HDUs/ICUs the estimated mortality (APACHE III and APACHE II) was higher than that in the ICUs (Table 5.2), with the remaining HDU/ICU having the 6th lowest APACHE II score in the study (Table 5.1).

Table 5.2 Profile of high dependency units mortality and severity of illness.

	HDUs/ICUs				ICUs
Hosp ID	F	T	V	O	
		In study			
Subjects	309	159	208	268	9449
Deaths	72	57	88	89	2749
Mortality (%)	23.30	35.85	42.31	33.21	29.09
APACHE III Score	53.05	66.03	70.30	66.22	60.25
APACHE III estimated mortality (%)	17.46	28.23	32.76	27.16	23.72
APACHE II Score	16.33	19.82	21.77	18.71	18.24
APACHE II estimated mortality (%)	24.54	36.22	41.39	31.41	29.98
		Exclusions			
Subjects	60	324	834	436	-
Deaths	5	32	71	15	-
Mortality (%)	8.33	9.87	8.51	3.44	-

HDU, High Dependancy Unit; ICU, Intensive Care Unit; APACHE, Acute Physiology and Chronic Health Evaluation.

5.1.2 Patients and exclusions

Data were collected on 13,291 consecutive admissions to the 22 Scottish ICUs in the years 1995/96. Of the 13,291 patients, 2,898 were either HDU patients, and excluded under the study's exclusion criteria or excluded for other reasons (Table 5.3) leaving 10,393 patients for analysis.

Table 5.3 Study exclusions

Reason for Exclusion	Number of patients
Readmissions	498
Missing Outcomes	114
Prospective study exclusion categories	632
HDU patients	1,654
Total	2,898

HDU, High Dependancy Unit.

The specific entry criteria for each system were applied to these 10,393 patients. This led to further exclusions although the numbers excluded varied because of the different criteria. When comparing the models it is important to use a common data set (for example, the different entry criteria concerning minimum period of admission will lead

to different samples being selected for each scoring system). Diagnostic categories were not available for 31 records, therefore, a probability for the APACHE models could not be calculated. Three records had the APACHE III diagnosis of "Rule out MI" which has no coefficient and therefore no probability of mortality was available for these records. Fifty-nine records had incomplete SAPS data and therefore probabilities could not be calculated on these patients for the SAPS model. The numbers excluded and the extra exclusions from each model can be seen in Table 5.4.

Table 5.4 Model exclusions

Model	Patients	Other reasons for exclusion	Exclusions and data missing
APACHE III	10,326	missing diagnosis & rule out MI	67
APACHE II	9,848	Stay < 8 hours & missing diagnosis	545
SAPS II	10,334	Missing Data	59
MPM0	10,393	------	-----
MPM24	7,343	Stay < 24 hours	3,045
UK APACHE II	9,848	Stay < 8 hours & missing diagnosis	545

APACHE, Acute Physiology and Chronic Health Evaluation; MI, mycardial infarction, SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Model.

Table 5.5 Subject characteristics for (a) survivors, (b) non-survivors, and (c) overall.

	Survivors n=7,338	Non-survivors n=3,055	All n=10,393	Observed Mortality %
Age(±SD)	56.8±18.7	64.07±14.8	58.9±17.9	---
Male (%)	4094 (55.8)	1,661(54.4)	5,755 (55.4)	28.9
Female(%)	3244 (44.2)	1,394 (45.6)	4,638 (44.6)	30.0
Non-operative(%)	3205 (43.7)	2,136 (69.9)	5,341(51.4)	40.0
Operative(%)	4133 (56.3)	919 (30.1)	5,052(48.6)	18.2
Elective(%)	1999 (27.2)	215(7.0)	2,214 (21.3)	9.7
Emergency(%)	2134 (29.1)	704 (23.0)	2,838 (27.3)	24.8

Data given are mean (standard deviation) for age, n (percent for other factors).
In addition observed mortality (%) is shown.

Although patients under 18 years of age were excluded from the original SAPS II and MPM II studies, all patients between 16-18 were scored and included in the study. Hospital mortality in our study population was 29.4%, with an ICU mortality of 20.5%. Fifty one percent of patients were non-operative and 48.6% operative. Patients had an average length of ICU stay of 4.64 days, with non-operative patients having an average length of ICU stay of 5.67 days compared to 3.56 in operative patients.

The demographics of the population used for analysis are shown in Table 5.5. A further breakdown of this population can be found in Table 5.6.1 and Table 5.6.2 which show the 20 most common operative and non operative APACHE III diagnoses.

Table 5.6.1 List of twenty most frequently chosen APACHE III operative diagnoses

Diagnosis	Number of patients	% Operative Patients
GI Neoplasm(not perforation/rupture)	703	13.95
GI perforation/rupture	522	10.33
GI obstruction(any cause)	390	7.72
Other Miscellaneous	357	7.07
Aortic aneurysm: pre-leak/dissection	265	5.25
Aortic aneurysm: Rupture	229	4.53
Other GI Surgery	214	4.24
Aortic-femoral, fem-fem bypass graft	162	3.21
Bleeding-ulcer	134	2.65
Peritonitis	115	2.28
Cholangitis/cholecystitis	109	2.15
Trauma - extremities	108	2.14
GI vascular insufficiency/embolism/infarction	97	1.92
Renal neoplasm	82	1.62
GI inflammatory disease	80	1.58
Other renal surgery	74	1.46
Neoplasm-mouth/sinuses	72	1.43
Other cardiovascular surgery	69	1.37
Localised GI abscess/cyst	63	1.24
Peripheral ischaemia	63	1.24
All	5,052	100

GI, Gastrointestinal

Table 5.6.2 List of twenty most frequently chosen APACHE III non operative diagnoses

Diagnosis	Number of patients	% Non-operative patients
GI neoplasm (not perforation)	564	10.56
Aortic aneurysm	507	9.49
GI perforation/rupture	369	6.90
GI obstruction (any cause)	297	5.56
Other Miscellaneous	241	4.51
Other GI disorder	163	3.05
Other respiratory disorder	159	2.98
Peripheral ischaemia	153	2.86
Peritonitis	135	2.53
Other cadiovascular disorder	128	2.40
Bleeding - ulcer	106	1.98
GI vascular insufficiency/embolism/infarction	85	1.59
Septic shock - gastrointestinal tract	76	1.42
GI inflammatory disease	72	1.34
Cholangitis/cholecystitis	71	1.32
Renal neoplasm	67	1.25
Congestive heart failure	60	1.12
Other renal disorder	59	1.10
Trauma - multiple site without head/brain	59	1.10
Localised airway obstruction/oedema (mechanical)	55	1.03
All	5,341	100

GI, Gastrointestinal

5.1.3 Model performance

The results for each model are summarised in Tables 5.7 and 5.8. APACHE III demonstrated the best discrimination for the area under the ROC curves (0.845), whereas MPM_0 had the poorest discrimination with an area under the ROC curve of 0.785. Statistically valid comparisons can only be made for models that include the same patients. Thus, only comparisons of APACHE II and UK APACHE II can be made. The area under the ROC curve for UK APACHE II was not significantly greater than that for APACHE II (APACHE II = 0.805 and UK APACHE II = 0.809).

All the models had Hosmer-Lemeshow goodness of fit tests, which were statistically significant ($P<0.001$). This indicates statistical significance between observed and predicted mortality in all models. This may not be surprising, given that small differences can achieve significance with large sample sizes. Although there is no formal way of comparing Chi Squared statistics, it would appear that there was a considerable difference between the APACHE II Chi Squared value ($x^2=36.39$, $p<0.001$) and the APACHE III Chi Squared value ($x^2=331.65$, $p<0.001$).

Table 5.7 ROC values and Hosmer-Lemeshow goodness of fit tests using each model's exclusion criteria and for common subset

	Patients using each model's exclusion criteria			Patients in common subset		
Model	N=	ROC Area under the curve	Hosmer Lemeshow GOF Chi Squared	N=	ROC Area under the curve	Hosmer Lemeshow GOF Chi Squared
APACHE III	10,326	0.845	331.7*	7,338	0.795	365.7*
APACHE II	9,848	0.805	36.4*	7,338	0.763	67.4*
SAPS II	10,334	0.843	57.8*	7,338	0.784	142.0*
MPM_0	10,393	0.785	307.5*	7,338	0.741	451.9*
MPM_{24}	7,343	0.799	101.9*	7,338	0.791	100.8*
UK APACHE II	9,848	0.809	307.5*	7,338	0.756	236.6*

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Model; ROC, Receiver Operating Characteristic; GOF, Goodness of fit; * $P<0.001$.

Table 5.8 Hosmer-Lemeshow goodness-of-fit statistics for all models

	Deciles of Predicted Mortality (%)	Observed survivors	Expected survivors	Observed deaths	Expected deaths	GOF statistic & df
APACHE II	0<4	958	959.63	27	25.37	
	4<8	919	925.57	66	59.43	
	8<12	887	889.95	98	95.05	
	12<17	847	846.67	138	138.33	
	17<23	796	792.41	189	192.59	
	23<31	710	722.93	275	262.07	
	31<40	658	637.11	327	347.89	
	40<52	599	532.13	386	452.87	
	52<69	451	396.72	534	588.28	df=8
	69<100	203	186.50	780	796.50	36.39*
APACHE III	0<2	1010	1023.32	23	9.68	
	2<3	986	1009.26	47	23.74	
	3<5	952	992.35	81	40.65	
	5<8	913	966.27	120	66.73	
	8<13	856	927.62	177	105.38	
	13<20	791	869.41	242	163.59	
	20<30	678	779.37	355	253.63	
	30<45	575	649.89	458	383.11	
	45<68	384	454.68	649	578.32	df=8
	68<100	141	173.18	888	855.82	331.65*
UKAPACHE II	0<8	952	933.29	33	51.71	
	8<13	938	875.75	47	109.25	
	13<18	880	829.17	105	155.83	
	18<24	839	777.24	146	207.76	
	24<31	813	714.64	172	270.36	
	31<39	720	645.39	265	339.61	
	39-48	670	565.16	315	419.84	
	48<59	574	466.89	411	518.11	
	59<3	424	343.37	561	641.63	df=8
	73<100	218	163.55	765	819.45	307.51*
SAPS II	0<3	1011	1018.09	22	14.91	
	3<5	992	994.20	41	38.80	
	5<9	958	962.54	75	70.46	
	9<14	897	919.84	136	113.16	
	14<20	841	862.12	192	170.88	
	20<29	792	784.72	241	248.28	
	29<41	681	676.18	352	356.82	
	41<58	559	534.23	474	498.77	
	58<80	415	333.42	618	699.58	df=8
	80-<00	145	108.92	892	928.08	57.75*
MPM 0	0<3	991	1013.51	48	25.49	
	3<6	950	991.01	89	47.99	
	6<9	923	960.37	116	78.63	
	9<14	856	920.19	183	118.81	
	14<20	819	863.52	220	175.48	
	20<26	761	803.75	278	235.25	
	26<37	695	718.79	344	320.21	
	37<53	636	570.85	403	468.15	
	53<75	457	377.37	582	661.63	df=8
	75<100	250	131.27	792	910.73	307.47*
MPM 24	0<4	715	714.19	19	19.81	
	4<7	675	694.97	59	39.03	
	7<11	662	670.26	72	63.74	
	11<16	590	636.51	144	97.49	
	16<22	568	596.65	166	137.35	
	22<30	500	545.46	234	188.54	
	30<40	495	479.72	239	254.28	
	40<53	397	395.13	337	338.87	
	53<71	310	278.27	424	455.73	df=8
	71<100	179	119.80	558	617.20	101.87*

GOF, Goodness of fit; df, degrees of freedom; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Model; * P<0.001

The impact of these differences in calibration can be seen in the calibration curves and goodness of fit tables (Figures 5.1-5.3 and Table 5.8). APACHE II was seen to provide agreement between observed and predicted mortality across the full range of severity of illness, with the confidence intervals for observed mortality predominantly including the line of equality. Conversely, the observed mortality was consistently higher than that predicted by the APACHE III model with the line of equality lying outside the 95% confidence intervals. The confidence intervals for observed mortality for the UK APACHE II model were predominantly below the line of equality indicating that the mortality predicted by the model exceeds the observed mortality (Figure 5.2). These differences in calibration were reflected in the different frequency distribution plots of mortality prediction (Figures 5.1-5.3). Thus, while more than 40% of patients have a predicted mortality of 0 to 10% with APACHE III, UK APACHE II gives a figure of just over 20% for the same population. A similar pattern of observations was evident in the goodness of fit table (Table 5.8).

All the models, with the obvious exception of MPM_{24}, showed poorer discrimination after excluding patients discharged in the first 24 hours (Table 5.7). Thus, the rank order of performance changes with MPM_{24}'s performance relatively improved. Overall APACHE III, SAPS II and MPM_{24} have significantly superior discrimination compared with APACHE II, UK APACHE II and MPM_0 (Table 5.9). UK APACHE II and APACHE II were significantly superior to MPM_0.

Table 5.9 Comparison of ROC values in common dataset (showing P values)

N= 7,338	APACHE III P=	APACHE II P=	SAPS II P=	MPM_{24} P=	MPM_0 P=	UK APACHE II P=
ROC Values	0.795	0.763	0.784	0.791	0.741	0.756
APACHE III		0.0000	0.9921	0.3369	0.0000	0.0000
APACHE II			0.0000	0.0000	0.0006	0.5003
SAPS II				0.2298	0.0000	0.0000
MPM_{24}					0.0000	0.0000
MPM_0						0.0002
UK APACHE II						

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology; MPM, Mortality Probability Model.
Significance tested using the DeLong method.

Figure 5.1 Calibration curves for APACHE III and APACHE II models.

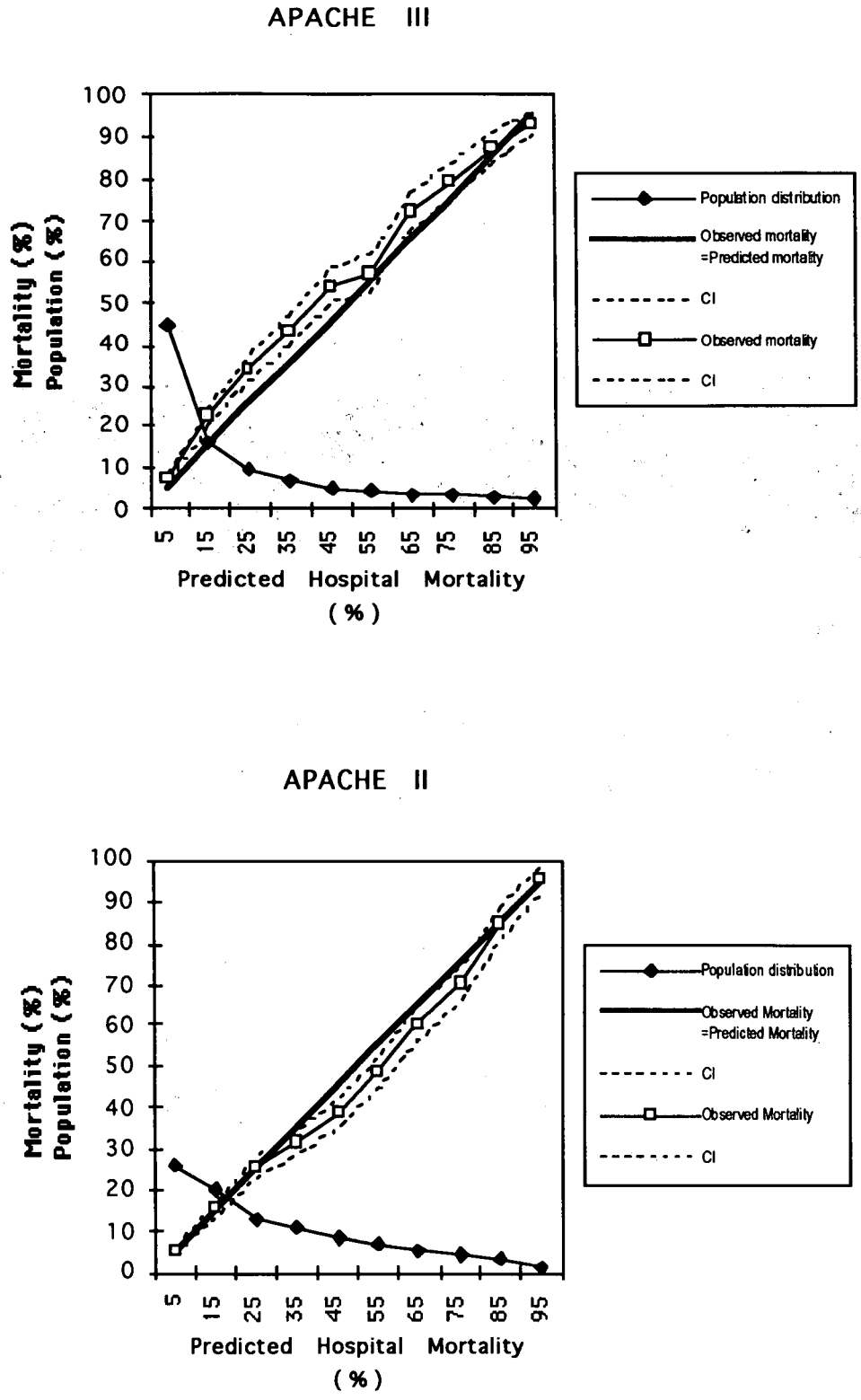


Figure 5.2 Calibration curves for the SAPS II and MPM_0 models.

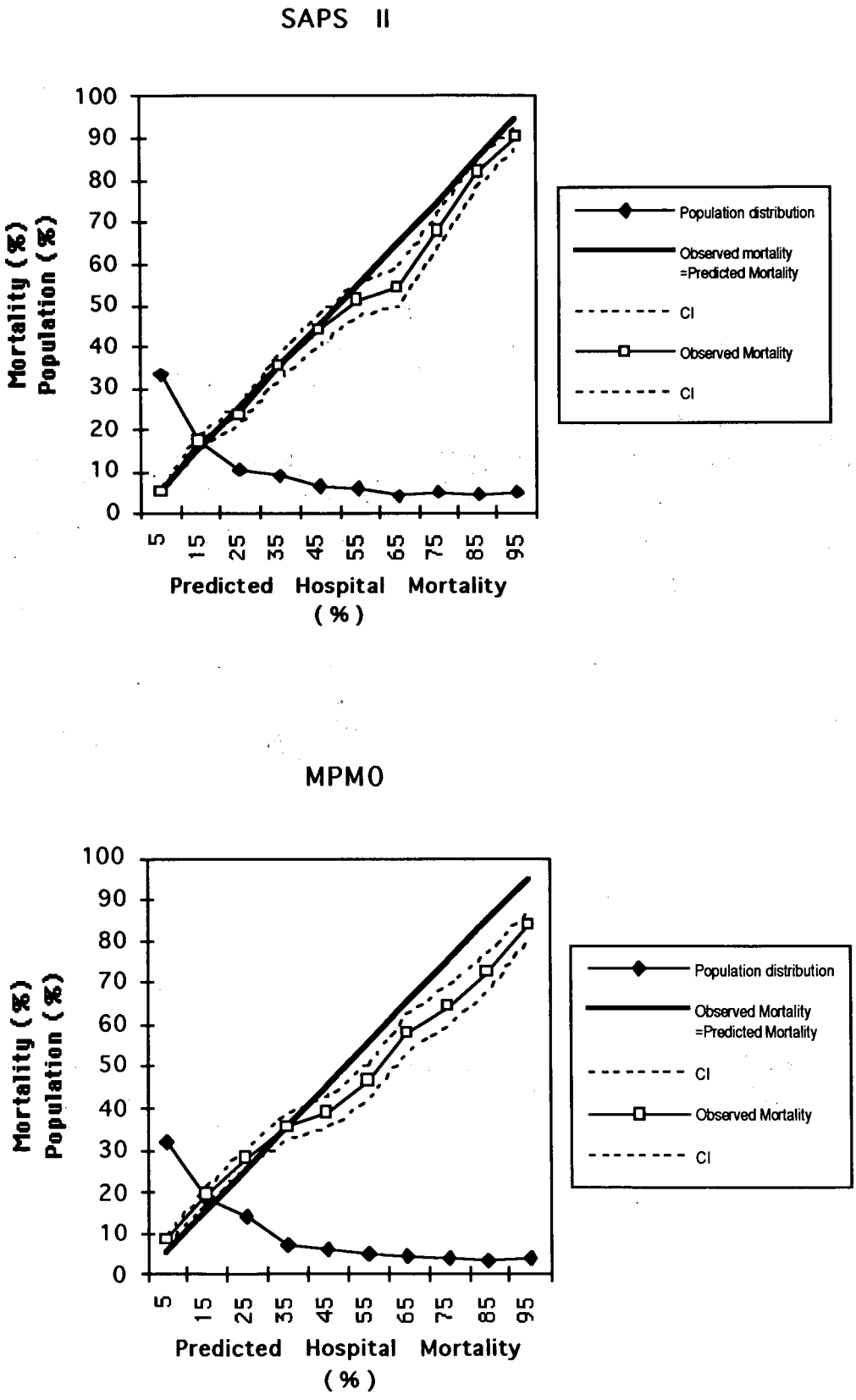
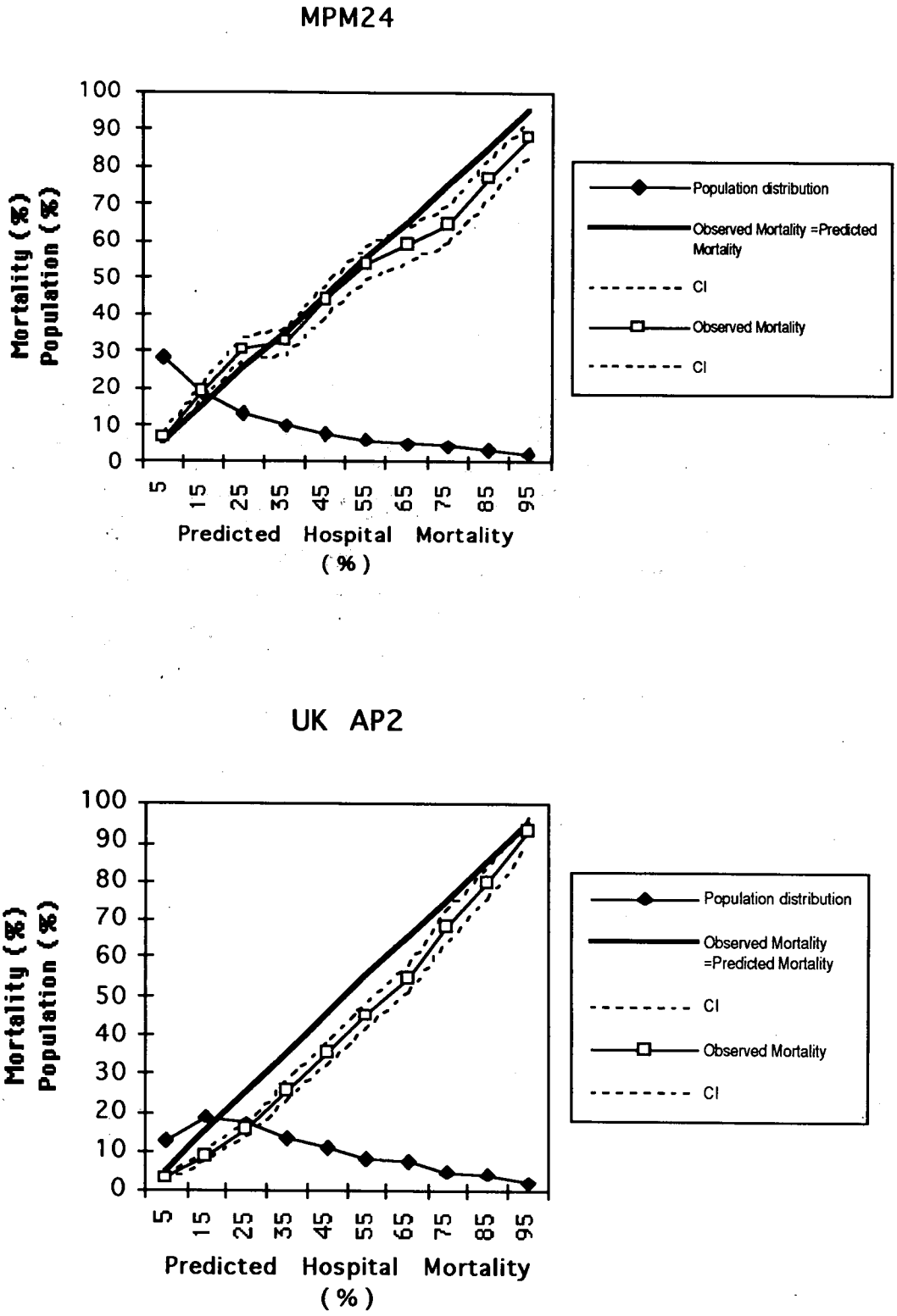


Figure 5.3 Calibration curves for MPM_{24} and UKAPACHE II models.



The explanation for the deterioration in the ROC values of all models, with the exception of MPM_{24},was evident from the ROC curves of the excluded population (APACHE III, 0.931; APACHE II, 0.908; UK APACHE II, 0.9132; SAPS II, 0.925; MPM_0, 0.875). This demonstrated high predictive ability of all models when restricted to patients discharged within 24 hours.

Model calibrations, which exclude patients discharged in the first 24 hours, were little changed from calibrations derived from the total study population with no consistent effect (Table 5.7 and 5.8). All produce significant differences with the Hosmer-Lemeshow goodness of fit test. UK APACHE II has improved calibration with a reduced Chi Squared value, while SAPS II, APACHE II, APACHE III, and MPM_0 have increased Chi Squared values.

5.2 Discussion

Within each prognostic scoring system the development and validation cohort were generated by randomly splitting the overall cohort. There have been several studies assessing the performance of scoring systems outside the ICU populations in which they were developed (53,73,83,84,102,110,115). A study of APACHE II in the UK showed similar discrimination (ROC Area) and calibration to that of the original model (73). However new diagnostic coefficients derived from the UK database showed improved discrimination and calibration when tested on that database (13).

Unlike other scoring systems, the weights needed to generate probabilities from the APACHE III model are not within the public domain. There has, therefore, been limited assessment of its performance outside the United States. A small study in a single UK ICU (83) showed APACHE III to have better discrimination than APACHE II (0.847 and 0.806 respectively) with both systems underestimating the mortality rate. A large study in the UK assessed the ability of APACHE III to adjust for case mix in 17 ICUs (84). APACHE III considerably underestimated mortality in this population, however discrimination, as judged by the ROC area, was similar to that in the original validation population.

A European study, taking data from the original multi-centre database used to validate SAPS II and MPM II, compared the performances of these new models with the original versions of these scoring systems (SAPS and MPM) and with that of APACHE II and APACHE III (score but not prediction as the latter was not available) (21). All three scoring systems demonstrated improved performance of the newer versions with no individual model being clearly superior.

When assessing the relative performance of severity of illness scoring systems it is important to appreciate that the most relevant assessment of their performance will depend on the proposed application. If models are to be employed to assess quality of care, by derivation of standardised mortality ratios, then calibration is the more significant measure of performance with discrimination being secondary to this. Wagner said,

> "*When you talk about performance evaluation you are fundamentally talking about calibration issues of the model. You are not as concerned about discrimination or ROC areas. Higher ROC areas are always better...*" (9th European Congress of Intensive Care Medicine and personal communication).

Indeed it has been argued by Castella et al that a model must have appropriate calibration before it can be applied to a population outside its original development population and only then can the model's discrimination be analysed (21).

Model calibration, as indicated by the Hosmer-Lemeshow goodness of fit test, demonstrated that for all systems there was a significant difference between the observed and predicted mortality. As the number of patients increases, the magnitude of the difference between observed and predicted mortality, required to generate a significant difference, decreases. This sensitivity to larger patient numbers was demonstrated by Zhu et al (142). Given the size of our study population, it may not be surprising that the goodness of fit tests were significant. However, the original studies for both the MPM II and SAPS II models demonstrated no statistical significance for the Hosmer-Lemeshow test (MPM_0 P=0.327; MPM_{24} P=0.231; SAPS II P=0.104) despite large numbers. The poor calibration may be in part due to higher average mortality in our cohort of patients (29.4 %) compared to that of the original studies (MPM II, 20.8%; APACHE III, 17.3%; APACHE II, 19.7%; SAPS II, 21.9%). This difference reflects differing admission criteria which presumably derived from a lower number of available ICU beds per head of the Scottish population.

Although there are no formal means of directly comparing the Chi Squared values derived from the goodness of fit tests, it would appear that there were large differences between some of the models. Thus, the calibration characteristics were best for APACHE II (36.4) and SAPS II (57.8) and poorest for APACHE III (331.7), MPM_0 (307.5) and, most surprisingly, for the UK derived version of APACHE II (307.5). Given that Scotland has a similar health care system to the UK as a whole and that four of the units contributed data both to this study and the UK APACHE II study, from which this model was derived, the poor performance of the UK APACHE II is difficult to explain.

While goodness of fit statistics allow statistical evaluation of relative performance it is possible that meaningful information on a given model's limitations can be visualised by examination of calibration curves. In particular, they demonstrate the extent to which the model predicts accurately across the full range of severity of illness. Consequently, whenever a given model is to be used there should be evaluation of this aspect of its performance within the population, prior to derivation of standardised mortality ratios. The low Chi Squared value for APACHE II appears to be associated with a close relationship between observed and predicted mortality across the full range of illness severity. The plots of the frequency distribution of predicted mortality demonstrate why the APACHE III prediction of mortality is lower than the observed rates, with a far higher proportion of patients predicted to have a mortality of 10% or less, compared with APACHE II. The reverse is seen with the UK APACHE II model.

While discrimination and calibration are both important characteristics of performance, calibration is the more important where it is to be used to predict mortality rate for populations. However, it remains important that it should demonstrate adequate discriminant characteristics. Statistical evaluation of the models required that patients who were discharged, alive or dead, within the first 24 hours were excluded from analysis, as suggested by Lemeshow et al (4). The performance of models which normally include such patients, all deteriorated when these patients were excluded. This results from the ability of the models to discriminate maximally for the excluded group. Other studies have also demonstrated this deterioration in model performance over time (92,109,178). When this approach was taken a statistical superiority in discrimination was demonstrated for APACHE III, SAPS II and MPM_{24} when differences were tested using the Delong method.

Using the models own exclusion criteria, when the results of both methods of evaluation were taken together it appears that the performance of SAPS II, was perhaps best overall. The SAPS II model had the second best rank order performance in terms of calibration and was one of the three best models in terms of discrimination. However, APACHE II, due to its superior calibration would, perhaps, be the most appropriate model from which to derive standardised mortality ratios in Scotland. The APACHE III model, which includes additional weightings for aspects such as source of admission, may be more health care system specific. The only other comparable study of its performance in the UK (83) demonstrated a similar pattern of relatively poor calibration associated with an underestimate of mortality, but with very good discrimination. The performance of the MPM_0 model was relatively poor in terms of discrimination and calibration. The performance of MPM_{24} was good in terms of discrimination and had comparable

calibration. Although the UK APACHE II model had comparable discrimination to that of the original APACHE II model, it had poorer calibration. This is surprising and not easily explained given that this study shares the same health care system.

A possible explanation for the underestimation of mortality by the APACHE III model and the better calibration demonstrated by the APACHE II system could reflect an increase in the standard of care required in ICUs by the APACHE III model. This might be the case, reflecting APACHE III's development from more recent data collection when compared with APACHE II. However, the data used to develop the SAPS II and MPM II models were drawn from a similar period and show no such underestimate of mortality. Furthermore, Beck et al (83) showed in a UK ICU that both APACHE II and APACHE III underestimated mortality in its population.

Another possible explanation, for at least some of the underestimation of mortality in the APACHE III model and the poor fit in general of the other models, may be the effect of deprivation in the Scottish population. There is a lot of recent evidence to suggest that deprivation affects the outcome of a number of diseases and surgical procedures (179-181). If the Scottish population is a more deprived than the populations represented in the original models databases then it is possible that this may have a considerable effect on the performance of the models. However, the ICUs involved in collecting data for the development of these models may have had considerable deprivation within their catchment area.

One possible criticism of the study is the inclusion of patients from the HDU/ICUs. This is because the inclusion of these patients is decided by clinical judgement of the Consultant on the different units. It is possible that this will lead to bias when included with ICUs who include all patients in the study. However, clinicians on ICUs make the same judgement about the inclusion of their patients in the study by deeming them an "Intensive Care type patient" when they admit them to an ICU. Also, it was important for the Audit not to exclude "Intensive Care type" patients from HDU/ICUs as they represented 9.08% of Intensive care patients in the Study. The mortality rate in three of the units is higher than the mortality rate in the ICUs, with five ICUs having a lower mortality rate than the remaining HDU/ICU. The mortality estimated by the APACHE II model and the mean APACHE II score for three of the HDU/ICUs is higher than the average estimated APACHE II mortality and score for the ICUs (Table 5.4). The remaining HDU/ICU has an APACHE II score that is higher than 6 of the ICUs (Table 5.5) . These results would suggest that patients from HDU/ICUs were included appropriately. Analysis of those patients not included in the study by the HDU/ICUs shows a mortality rate with a range of 3.4 to 9.87 which is considerably lower than the

mortality rate in those patients included in the study. It is unlikely that the HDU/ICUs were excluding patients who have a poor prognosis as this would be reflected in the mortality rate of the excluded patients (Table 5.4).

While it appears that APACHE II and SAPS II perform best in comparison to other models, it is not clear how accurate that performance is. Significant Hosmer-Lemeshow GOF tests in the models, especially those which would appear to have better calibration (APACHE II and SAPS II), may be a result of the large numbers in the study. These differences may be of no clinical significance, with the numbers in the study making the sensitivity test reveal differences that have no impact on the clinical application of the study. Nevertheless it is important, given the lack of any other evidence, not to dismiss the significant GOF tests. As all the statistics used in measuring calibration are necessarily a result of averaging (as the outcome variable is a dichotomous one), it is possible, given the large numbers, that smaller groups of patients over and underestimate the mortality but appear to have good calibration when combined.

There are considerable differences in the performances of all the models used in this study compared with the original published data for both discrimination and calibration. It would therefore seem appropriate that before these models are used for comparing different ICUs further analysis is carried out on their accuracy in different groups of patients.

Chapter 6-Uniformity of fit

Aim: To assess the uniformity of fit of the severity of illness models.

Contents:

6.1 Introduction

As well as having good calibration and discrimination, a model must have adequate case mix adjustment for large subgroups of patients (Uniformity of Fit) (73,115,129). If a model does not adequately adjust for case mix in a subgroup, an ICU admitting a large number of patients from this subgroup will appear to have poor performance. For example, if the APACHE II model underestimates the risk of mortality in medical patients, a unit that admits mainly this type of patient will appear to perform poorly.

The profile of each of the 22 units will be different in some way. Some of the units are based in large teaching hospitals others in smaller more general hospitals. Some of the units admit mainly operative patients with others receiving mainly medical patients. It is because of this imbalance in the type of patients being admitted that case mix adjusted outcomes are so important.

6.2 Methods

To test uniformity of fit patients were grouped into mutually exclusive subgroups by both the APACHE II system and source of admission. The APACHE III admitting diagnosis is categorised by nine systems (cardiovascular, respiratory, neurological, gastrointestinal, renal, metabolic/endocrine, haematological, trauma, and general). Source of admission was recorded as part of the APACHE III data set and grouped by Accident and Emergency (A&E), recovery/theatre, ward in this hospital, other ICU in this hospital, ICU in another hospital, other area in another hospital, and home/clinic. Given the poor performance and the incompatibility of the diagnosis, the UK APACHE II model has been dropped from any further analysis.

6.3 Data analysis

In the previous analysis, goodness of fit tests and ROC curves were used to assess the performance of the different models. However, these tests are not appropriate in the analysis of subgroups. As the numbers of patients increase so the margins for finding significant results in the Hosmer and Lemeshow GOF test will narrow. As the numbers in each group (system and source) were different, and were smaller than the overall database, it was not appropriate to use GOF tests to compare calibration in the different sub groups (142,163).

Comparing ROC values in different subgroups as a measure of performance would also not be appropriate as discrimination in different subgroups will naturally produce

different types of patients. Patients within some groups may represent the middle range of severity of illness making discrimination more difficult i.e. patients remaining in the ICU longer than 24 hours (Table 5.7). It is more important that a model's calibration in groups is adequate.

CIs (95%) were calculated for the observed mortality in each subgroup to assess the difference between the mean observed mortality and the mean estimated by each of the models (see Chapter 3.9.2). The mean estimated mortality was considered to be significant if it lay outside the confidence intervals for the observed mortality.

P-values from the Chi Squared test were calculated to assess the significant differences between the observed survivors and non survivors and those estimated by the models (see Chapter 3.9.4).

Mortality ratios were calculated with their confidence intervals (see Chapter 3.9.3). Ratios whose confidence intervals lay outside the confidence intervals for the overall population were treated as significantly different to the overall population mortality ratio for a particular model. This allows the prediction of subgroups to be related to the overall estimated mortality generated by a model.

6.4 Results

6.4.1 APACHE II

Within the source of admission category the estimated mortality lay outside the observed mortality CIs for the A&E, recovery/theatre category, ward and other area in another hospital (Table 6.1). Of the APACHE system categories the estimated mortality lay outside the CIs for the cardiovascular, neurological, gastrointestinal and general groups (Table 6.2). With the model underestimating mortality in the cardiovascular and neurological groups and overestimating the mortality in the gastrointestinal and general groups. These results were confirmed by the Chi Squared test. The mortality ratio CIs (observed/expected) lay outside the mortality ratios CIs for the overall population in those patients admitted from A&E, recovery/theatre, ward and another area in another hospital.

Table 6.1 Patients by source of admission for the APACHE II model

APACHE Source	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	χ^2 P=
A&E	1,183	1.09* (1.03-1.16)	30.8*	33.6 (31.0-36.3)	0.080
Recovery/theatre	4,909	0.79* (0.74-0.83)	22.8*	17.9 (16.9-19.0)	0.001
Ward in this hospital	2,703	1.08* (1.04-1.12)	41.7*	45.0 (43.1-46.8)	0.008
Other ICU in this hospital	168	1.01 (0.85-1.18)	37.0	37.5 (30.2-44.8)	0.920
ICU in another hospital	315	1.08 (0.95-1.22)	34.3	37.1 (31.8-42.5)	0.380
Other area in another hospital	560	0.81* (0.71-0.92)	31.2*	25.4 (21.8-29.0)	0.014
Home/clinic	10	1.45 (0.60-2.30)	27.6	40.0 (9.6-70.4)	0.458
Overall patients	9,848	0.95 (0.93-0.98)	30.0*	28.6 (27.7-29.5)	0.011

APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; A&E, Accident and Emergency, ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Table 6.2 Patients by APACHE III diagnostic system for the APACHE II model

APACHE System	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	χ^2 P=
Cardiovascular	2,131	1.09* (1.05-1.14)	37.7*	41.2 (39.1-43.2)	0.009
Respiratory	1,846	1.04 (0.98-1.10)	30.6	31.7 (29.6-33.9)	0.377
Neurological	850	1.35* (1.25-1.45)	21.7*	29.3 (26.2-32.4)	0.001
Gastrointestinal	3,276	0.77* (0.73-0.81)	34.1*	26.3 (24.7-27.8)	0.001
Renal	299	0.81 (0.62-0.99)	22.8	18.4 (14.0-22.8)	0.114
Metabolic/ Endocrine	126	1.04 (0.80-1.29)	27.4	28.6 (20.7-36.5)	0.806
Haematological	48	1.08 (0.80-1.37)	42.3	45.8 (31.7-59.9)	0.708
Trauma	697	1.18* (0.99-1.37)	11.0	13.1 (10.6-15.6)	0.110
General	575	0.49* (0.32-0.66)	15.7*	7.7 (5.5-9.8)	0.001
Overall patients	9,848	0.93 (0.95-0.98)	30.0*	28.6 (27.7-29.5)	0.011

APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; A&E, Accident and Emergency, ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Patients admitted with cardiovascular, neurological, gastrointestinal, trauma and general diagnosis had mortality ratios that fell outside the confidence intervals for the overall groups and also had significant Chi Squared tests.

6.4.2 APACHE III

The estimated mortality of those patients who were admitted from A&E, recovery/theatre, ward and ICU in another hospital lay outside the CIs for observed mortality (Table 6.3). Those patients with a cardiovascular, respiratory, neurological, gastrointestinal, haematological or trauma diagnosis had an estimated mortality which fell outside the CIs for observed mortality (Table 6.4). All significant groups underestimated mortality. The analysis of the expected mortality compared with the observed mortality was confirmed by the significance in the Chi Squared test. Only those admitted from other ICU in this hospital and another area in another hospital or with a neurological haematological or general diagnosis had a mortality ratio whose CIs fell outside the CIs for the overall model.

6.4.3 SAPS II

Patients admitted from A&E, recovery/theatre, the ward (the three largest groups) and another area in another hospital had an estimated mortality which fell outside the CIs for observed mortality in these groups (Table 6.5). The model overestimated mortality in the A&E, other area in another hospital and recovery/theatre groups and underestimated mortality in patients admitted from the Ward. Those with neurological, trauma, renal and general diagnosis also had estimated mortality ratios that fell outside the CIs for observed mortality, with all significant groups overestimating mortality (Table 6.6). These results were again confirmed by the significance of the Chi Squared test in both the source of admission and the diagnoses, with the exception of a non-significant result in the other area in another hospital group. Patients admitted from A&E, recovery/theatre and the ward as well those with a diagnosis that fell in the respiratory, neurological, renal, general and trauma groups all had mortality ratios whose CIs fell outside those for the overall patients.

Table 6.3 Patients by source of admission for the APACHE III model

APACHE Source	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	X^2 P=
A&E	1,300	1.27 (1.21-1.34)	26.6*	33.8 (31.3-36.4)	0.001
Recovery/theatre	5,030	1.23 (1.17-1.28)	14.9*	18.2 (17.2-19.3)	0.001
Ward in this hospital	2,917	1.24 (1.20-1.28)	37.1*	46.0 (44.2-47.8)	0.001
Other ICU in this hospital	175	1.02* (0.86-1.17)	37.7	38.3 (31.1-45.5)	0.888
ICU in another hospital	321	1.29 (1.15-1.43)	29.1*	37.7 (32.4-43.0)	0.005
Other area in another hospital	572	1.02* (0.91-1.13)	25.3	25.9 (22.3-29.5)	0.791
Home/clinic	11	1.86 (0.87-2.85)	19.6	36.4 (7.9-64.8)	0.209
Overall patients	10,326	1.23 (1.20-1.25)	24.0*	29.4 (28.6-30.3)	0.001

APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; A&E, Accident and Emergency, ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Table 6.4 Patients by APACHE III diagnostic system for the APACHE III model

APACHE System	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	X^2 P=
Cardiovascular	2,281	1.25 (1.21-1.30)	34.8*	43.6 (41.6-45.7)	0.001
Respiratory	1,918	1.20 (1.14-1.26)	26.6*	32.0 (29.9-34.1)	0.001
Neurological	926	1.44* (1.34-1.53)	19.6*	28.2 (25.3-31.1)	0.001
Gastrointestinal	3,368	1.17 (1.12-1.22)	23.0*	27.0 (25.5-28.5)	0.001
Renal	312	1.14 (0.94-1.33)	16.4	18.6 (14.3-22.9)	0.330
Metabolic/ Endocrine	130	1.25 (1.02-1.48)	22.8	28.5 (20.7-36.2)	0.174
Haematological	53	1.88* (1.50-2.27)	22.0*	41.5 (28.2-54.8)	0.003
Trauma	738	1.34 (1.17-1.52)	10.0*	13.4 (11.0-15.9)	0.003
General	600	0.85* (0.63-1.08)	9.0	7.7 (5.5-9.8)	0.286
Overall patients	10,326	1.23 (1.20-1.25)	24.0	29.4 (28.6-30.3)	0.001

APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Table 6.5 Patients by source of admission for the SAPS II model

APACHE Source	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	χ^2 P=
A&E	1,301	0.83* (0.78-0.88)	40.6*	33.9 (31.3-36.5)	0.001
Recovery/theatre	5,031	0.88* (0.84-0.93)	20.6*	18.2 (17.2-19.3)	0.001
Ward in this hospital	2,922	1.11* (1.08-1.15)	41.3*	46.0 (44.2-47.8)	0.001
Other ICU in this hospital	175	0.95 (0.81-1.10)	40.2	38.3 (31.1-45.5)	0.689
ICU in another hospital	321	1.00 (0.89-1.12)	37.5	37.7 (32.4-43.0)	1.000
Other area in another hospital	573	0.86 (0.76-0.96)	30.2*	25.8 (22.2-29.4)	0.057
Home/clinic	11	1.16 (0.56-1.75)	31.4	36.4 (7.9-64.8)	0.764
Overall patients	10,334	0.97 (0.95-0.99)	30.4*	29.4 (28.6-30.3)	0.084

SAPS, Simplified Acute Physiology Score; APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; A&E, Accident and Emergency, ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Table 6.6 Patients by APACHE III diagnostic system for the SAPS II model

APACHE System	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	χ^2 P=
Cardiovascular	2,283	1.03 (0.99-1.07)	42.4	43.7 (41.6-45.7)	0.337
Respiratory	1,918	1.06* (1.01-1.12)	30.2	32.0 (29.9-34.1)	0.140
Neurological	927	0.79* (0.72-0.86)	35.7*	28.2 (25.3-31.1)	0.001
Gastrointestinal	3,368	0.99 (0.95-1.04)	27.2	27.0 (25.5-28.5)	0.806
Renal	312	0.77* (0.62-0.93)	24.1*	18.6 (14.3-22.9)	0.048
Metabolic/ Endocrine	130	0.80 (0.63-0.96)	35.7	28.5 (20.7-36.2)	0.167
Haematological	53	1.22 (0.92-1.53)	33.9	41.5 (28.2-54.8)	0.343
Trauma	738	0.74* (0.61-0.86)	18.2*	13.4 (11.0-15.9)	0.002
General	600	0.63* (0.45-0.81)	12.2*	7.7 (5.5-9.8)	0.002
Overall patients	10,334	0.97 (0.95-0.99)	30.4*	29.4 (28.6-30.3)	0.084

SAPS, Simplified Acute Physiology Score; APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; A&E, Accident and Emergency, ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

6.4.4 MPM_0

The mean estimated mortality from the MPM_0 model fell outside the CIs for the observed mortality for patients admitted from A&E, recovery/theatre, and the wards (Table 6.7). Patients with a cardiovascular, respiratory, neurological, haematological, trauma and general diagnosis also had mean estimated mortalities lying outside the CIs for the observed mortalities (Table 6.8). The model overestimated mortality in the trauma, neurological, general, A&E and recovery/theatre groups and underestimated mortality in the cardiovascular, respiratory and ward groups. These results from both the source of admission and the diagnoses are confirmed by the significance in the Chi Squared value. Patients admitted from A&E, recovery/theatre and the ward, and patients with a cardiovascular, respiratory, neurological, trauma, general and haematological diagnosis all had mortality ratio CIs that lay outside the CIs for the overall patients.

6.4.5 MPM_{24}

Patients admitted from A&E, recovery/theatre, ward in this hospital, and another area in another hospital and with a cardiovascular, respiratory, renal and trauma diagnosis had a significantly different expected mortality to that observed using CIs. All but those admitted from another area in another hospital and those with a cardiovascular diagnosis had a significant chi squared test. The model overestimated the mortality in the groups recovery/theatre, other area in another hospital, renal and trauma and underestimated mortality in patients admitted from the wards and with a respiratory and cardiovascular diagnoses. The mortality ratio CIs from the MPM_{24} model lay outside the CIs for the overall patients for patients admitted from A&E, recovery/theatre, the wards, other area in another hospital and patients with an admitting diagnosis from the respiratory, renal and trauma groups (Table 6.10).

Subgroups showing significance using the CIs for observed mortality for all models can be seen in Table 6.11.

Table 6.7 Patients by source of admission for the MPM_0 model

APACHE Source	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	χ^2 P=
A&E	1,313	0.84* (0.78-0.89)	40.6*	34.0 (31.4-36.5)	0.001
Recovery/theatre	5,052	0.90* (0.85-0.95)	20.2*	18.2 (17.1-19.3)	0.001
Ward in this hospital	2,940	1.20* (1.16-1.23)	38.3*	45.8 (44.0-47.6)	0.001
Other ICU in this hospital	178	0.99 (0.84-1.14)	38.6	38.2 (31.1-45.3)	0.920
ICU in another hospital	322	0.98 (0.86-1.09)	38.4	37.6 (32.3-42.9)	0.806
Other area in another hospital	575	0.90 (0.79-1.01)	28.9	25.9 (22.3-29.5)	0.190
Home/clinic	13	1.59 (0.80-2.38)	24.3	38.5 (12.0-64.9)	0.299
Overall patients	10,393	1.00 (0.98-1.03)	29.3	29.4 (28.5-30.3)	0.819

MPM, Mortality Probability Model; APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; A&E, Accident and Emergency, ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Table 6.8 Patients by APACHE III diagnostic system for the MPM_0 model

APACHE System	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	χ^2 P=
Cardiovascular	2,291	1.10* (1.06-1.14)	39.7*	43.6 (41.5-45.6)	0.003
Respiratory	1,923	1.13* (1.07-1.18)	28.4*	31.9 (29.8-34)	0.003
Neurological	933	0.72* (0.66-0.79)	38.9*	28.2 (25.3-31.1)	0.001
Gastrointestinal	3,377	1.03 (0.99-1.08)	26.0	26.9 (25.4-28.4)	0.303
Renal	315	0.83 (0.66-1.00)	22.6	18.7 (14.4-23)	0.153
Metabolic/ Endocrine	130	0.83 (0.66-1.00)	34.3	28.5 (20.7-36.2)	0.260
Haematological	53	1.57* (1.18-1.96)	26.5*	41.5 (28.2-54.8)	0.034
Trauma	739	0.72* (0.59-0.85)	18.6*	13.4 (10.9-15.9)	0.001
General	604	0.60* (0.42-0.78)	12.7*	7.6 (5.5-9.7)	0.001
Overall patients	10,393	1.00 (0.98-1.03)	29.3	29.4 (28.5-30.3)	0.819

MPM, Mortality Probability Model; APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Table 6.9 Patients by source of admission for the MPM_{24} model

APACHE Source	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	X^2 P=
A&E	781	1.16* (1.08-1.24)	32.3*	37.5 (34.1-41.3)	0.011
Recovery/theatre	3,445	0.86* (0.81-0.91)	24.1*	20.8 (19.5-22.0)	0.001
Ward in this hospital	2,223	1.17* (1.12-1.21)	37.5*	43.8 (41.8-46.0)	0.001
Other ICU in this hospital	139	0.96 (0.77-1.14)	36.1	34.5 (26.6-42.3)	0.752
ICU in another hospital	291	0.95 (0.83-1.07)	38.4	36.4 (30.9-42.5)	0.584
Other area in another hospital	455	0.84* (0.72-0.96)	28.5*	24.0 (20.0-27.3)	0.069
Home/clinic	9	1.74 (0.76-2.72)	25.6	44.4 (12.0-80.7)	0.262
Overall patients	7,343	1.02 (0.99-1.05)	30.1	30.7 (29.6-31.9)	0.396

MPM, Mortality Probability Model; APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; A&E, Accident and Emergency, ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Table 6.10 Patients by APACHE III diagnostic system for the MPM_{24} model

APACHE System	n	Mortality ratios (CI)	Expected mortality %	Observed mortality % (CI)	X^2 P=
Cardiovascular	1,599	1.06 (1.01-1.12)	39.1*	41.7 (39.2-44.1)	0.104
Respiratory	1,519	1.20* (1.13-1.27)	28.2*	33.8 (31.5-36.2)	0.001
Neurological	574	0.95 (0.85-1.04)	33.6	31.9 (28.1-35.7)	0.475
Gastrointestinal	2,482	0.96 (0.91-1.01)	29.3	28.0 (26.3-29.8)	0.240
Renal	213	0.72* (0.54-0.89)	27.5*	19.7 (14.4-25.1)	0.030
Metabolic/ Endocrine	83	0.93 (0.69-1.16)	32.5	30.1 (20.3-40.0)	0.699
Haematological	37	1.33 (0.98-1.69)	36.5	48.6 (32.5-64.8)	0.221
Trauma	507	0.77* (0.61-0.93)	18.3*	14.0 (11.0-17.0)	0.025
General	325	0.80 (0.57-1.03)	13.8	11.1 (7.7-14.5)	0.180
Overall patients	7,343	1.02 (0.99-1.05)	30.1	30.7 (29.6-31.9)	0.396

MPM, Mortality Probability Model; APACHE, Acute Physiology and Chronic Health Evaluation; CI, confidence interval; ICU, Intensive Care Unit; *Significance as indicated by confidence intervals

Figure 6.1 Mortality ratios for ICUs for the APACHE II, APACHE III and SAPS II models. Open circles, APACHE II. Open triangles, APACHE III. Open squares, SAPS II.

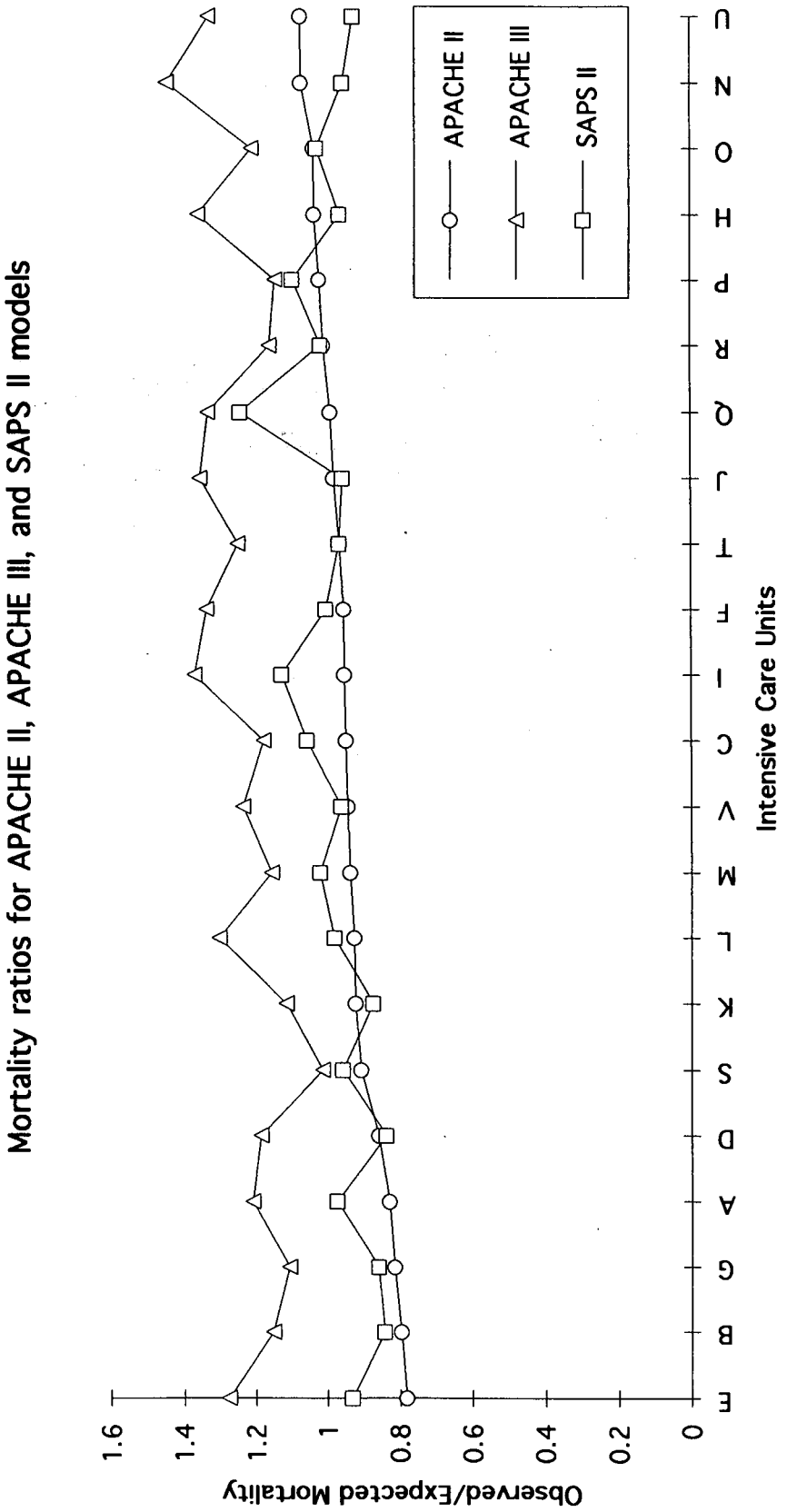


Figure 6.2 Mortality ratios for ICUs for the APACHE II, MPM_0, and MPM_{24} models. Open circles, APACHE II. Asterix, MPM_0. Open diamonds, MPM_{24}.

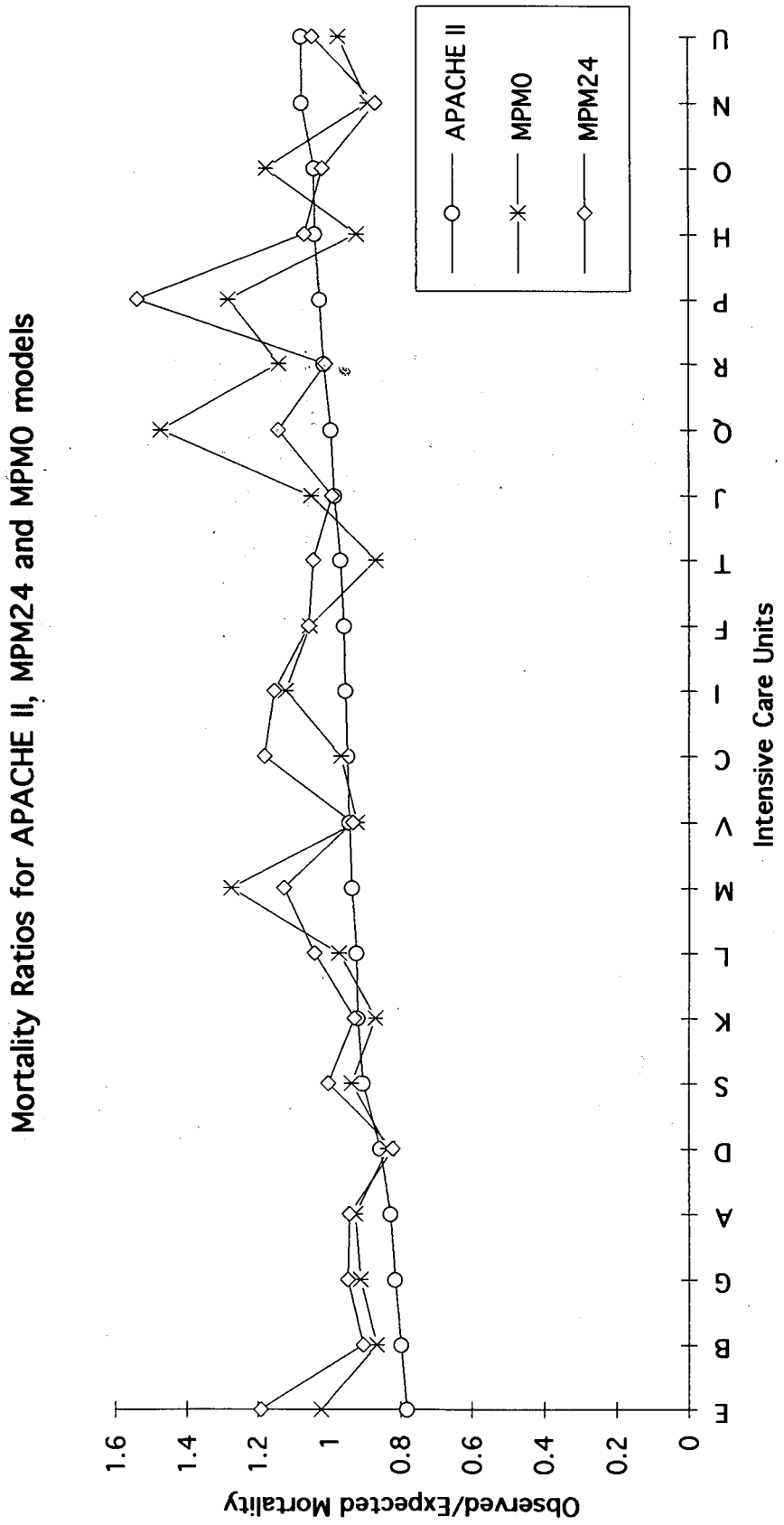


Table 6.11 Subgroups[a] where estimated mortality is significantly[b] different than that observed

Model	Subgroups with significant underestimation of mortality	Subgroups with significant overestimation of mortality
APACHE II	A&E, ward, cardiovascular, neurological	recovery/theatre, other area in another hospital, gastrointestinal, general
APACHE III	A&E, Recovery/theatre, ward in this hospital, ICU in another hospital, cardiovascular, respiratory, neurological, gastrointestinal, trauma	
SAPS II	ward in this hospital	A&E, recovery/theatre, Other area in another hospital, neurological, renal, trauma, general
MPM_0	ward in this hospital, cardiovascular, respiratory	A&E, recovery/theatre, trauma, general, neurological
MPM_{24}	A&E, ward in this hospital, cardiovascular, respiratory	recovery/theatre, other area in another hospital, trauma

APACHE, Acute Physiology and Chronic Health Evaluation; A&E, Accident and Emergency; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Model.
a. Only subgroups with more than 300 patients are shown
b. Significance as tested by observed mortality CIs

When mortality ratios for each of the units from each model are plotted together and ordered by APACHE II there appears to be no pattern of increasing mortality ratios for each of the models. (Figures 6.1-6.2).

6.5 Discussion

If these models are to be accurately used to assess an ICU's performance, then they must adequately adjust for case mix in large subgroups. If a model is sensitive to where a patient is admitted from or the patient's underlying disease then the profile of an ICU will profoundly affect the unit's mortality ratio and apparent performance (115,129).

ICUs in the study could arguably be seen as a homogeneous group, with much more in common with each other than the ICUs contributing data to these models' development. However, the ICUs (Table 5.1) in the study all have different profiles. As already stated, 36% of the ICUs are within teaching hospitals and 64% in non teaching hospitals. Some ICUs take mostly surgical patients others mostly medical patients. There are also high

concentrations of diagnoses within certain ICUs. It is important for any model to have good accuracy and calibration within subgroups otherwise case mix adjusted comparisons between units will only reflect failures within a model and not ICUs.

The study has used three statistics for assessing uniformity of fit.

1. The estimated mortality compared to the CIs of the proportions of the observed mortality.
2. The Chi Squared statistic.
3. The CIs for the mortality ratio in subgroups compared to the CIs for the overall population

Comparing the actual mortality to that estimated by the model using the Chi Squared statistic and the CIs for the observed mortality is a measure of calibration within a particular group. The mortality ratio in each group, on the other hand, is related to the overall mortality ratio for each model. If there is a systematic bias and the probabilities from a model are consistently low compared to the actual mortality experience then subgroups may not show significance when using mortality ratios. A model showing poor calibration in the different subgroups but with no significant mortality ratios in the subgroups may possibly be legitimately used to compare ICUs within the study.

When grouping patients by their source of admission and their diagnostic systems it is important to realise that some of the groups have small numbers and therefore any conclusions drawn from these groups may be unreliable. Therefore the focus was on subgroups of 300 patients or more. Wagner recommended that mortality ratios should not be compared until they were over 300 patients in a group (European Congress of Intensive Care Medicine).

All the models have subgroups which either show significant differences between the estimated and observed mortality or ratios which are significantly different from the overall ratio. It also has to be borne in mind the number of statistical tests carried out in this chapter, with each model having 48 tests. It would be expected that a certain number of tests would be significant because of random variation. However, the number of statistical significant tests found was greater than the number you would expect just from random chance. ICUs admitting patients from any of these subgroups may have their performance affected by the type of patient they are admitting and from where they are admitting them. Not all the subgroups underestimated mortality, in some groups the observed mortality was less than that predicted by the different models. It is possible that the overall performance of a model appears good but that good calibration in either deciles or decades of prediction is a result of averaging.

In the assessment of the overall performance of the different models APACHE II and SAPS II were the models with the best performance. Although both models had Hosmer-Lemeshow GOF tests that were significant ($P<0.0001$), it is possible that this significance was due to the large numbers in the study. However, both the APACHE II and SAPS II models had large subgroups of patients that were poorly predicted with significant differences in all three tests used. Also, those groups demonstrating significance contained a large number of patients.

APACHE III had performed relatively poorly compared to the other models in the study with a particularly large Chi Squared value ($x^2=331.65$). However, when the performance of the model is considered in the subgroups, APACHE III appears to have less variation than the other models. As the overall evaluation of the model shows, the estimates of mortality in the APACHE III model are lower than the actual mortality. It is therefore not surprising that the estimated mortality is significantly different to that actually observed in the largest subgroups using both CIs and the Chi Squared tests. However, when comparing the mortality ratios in the subgroups to the overall mortality ratios there is much less variation than in the other models. Patients admitted with a neurological and general diagnosis and patients admitted from an other area in another hospital are the only large subgroups (more than 300 patients) with mortality ratios significantly different than the overall mortality ratios of the groups. It is possible that comparisons of mortality ratios from the ICUs in this study would be more meaningful using the APACHE III model than the other models in this study. The consistency in the subgroups may be as a consequence of the presence of the diagnosis and source of admission as variables in the APACHE III model. The lack of fit in the neurological patients may, in part, be due to the rules for the collection of the GCS score. The rules state that if a patient is sedated in the first 24 hours then the GCS is presumed to be normal. As the GCS has the potential to contribute 19% to the APACHE III score, presuming patients who are sedated have a normal GCS could have a profound effect on the probabilities of mortality generated by the model.

There is considerable variation in the performance of these models in estimating the mortality in the different subgroups. An ICU's mortality ratio is dependent on the type and source of patients admitted to the unit. This seriously questions the ability of these models to accurately adjust for case mix. As all the models have been developed in other ICU cultures, it is not surprising that the resulting logistic regression models appear not to adjust adequately for case mix. There could be any number of differences in the ICU culture in Scotland, and the units participating in the original model development. When Intensive Care is commenced and admission criteria are just two possibilities where differences in ICU culture may have an effect but there are many more. The higher

average mortality in patients entering this study, compared to original study, has already been highlighted and may indicate that patients are, on average, sicker than those in the original studies. In their present form and based on this analysis there would appear to be no value in comparing ICU performance based on the output from these models.

From the apparent differences of the models in predicted mortality in the different subgroups it might be expected that the different ICUs in the study would perform differently in each model depending on the type and source of patients admitted. This would appear to be the case, with different ICUs having very different mortality ratios depending on the model (Figures 6.1-6.2). This apparent disparity between the models is confirmed by the correlation coefficients when each model is compared to the other. Given the relatively good ROC values in all the scores (0.79-0.85) this would suggest that each of the models is measuring a different aspect of severity of illness and accounts for a different percentage of variation in mortality

Chapter 7-The use of pre-sedation GCS value when calculating APACHE scores for sedated patients.

Aim: To assess the impact of using a pre-sedated GCS value when calculating APACHE scores, when the patient has been sedated for the first 24 hours precluding the use of the GCS in the APACHE models.

Contents:

7.1 Introduction

The GCS is an important component of both the APACHE II and APACHE III models and has the potential to contribute 17% of the theoretical maximum acute physiology score in APACHE II, and 19% in APACHE III. This is more than any other single variable. It has been shown that there is a significant relationship between GCS and outcome in general ICU patients and that the overall predictive capability of APACHE III is improved by incorporating the GCS (182).

In view of the importance of the GCS, it is unfortunate that there are a significant number of Intensive Care patients in whom the use of sedation prevents its accurate assessment. In APACHE II and APACHE III a normal GCS value (83) is assigned to all such patients. An alternative approach in such cases, as used in the SAPS II model, is to substitute the GCS recorded before the patient was sedated (3). There are theoretical arguments as to the better approach but there is no published evidence to support one methodology in preference to the other, and certainly no direct comparison within the same system. We hypothesised that the choice might have a significant effect on the performance of scoring systems, particularly if the number of sedated patients was large. In this study 50% of patients were sedated, preventing the use of the GCS score in the APACHE model (Table 7.1). The impact would be different in each of the ICUs with the percentage of sedated patients varying considerably in individual units in the study (10-70%).

Analysis from Chapter 5 showed that patients with a neurological diagnosis had an estimated mortality that was significantly lower than that observed. As already stated, the GCS is an important factor in the APACHE models. It is even more important for neurological patients that some assessment of their neurological status is available in the score. It may be that the use of a pre-sedated GCS might make an important difference to the estimation of mortality for these patients.

This chapter presents the effect of either assigning a normal GCS value or adopting the pre-sedation value on the performance of both APACHE II and APACHE III in the data collected for this study.

7.2 Methods

Data from the study, described in chapter 3-5, were used for the analysis in this chapter.

Where the use of sedative drugs precluded accurate assessment of the GCS during the first 24 hours of Intensive Care, the pre-sedation GCS was recorded, based on clinical observation, communication with referring staff or the clinical notes. The APACHE II and III score and prediction were calculated for these patients in two ways. Firstly this was done in the usual way, assuming that the GCS was 15 (normal GCS), producing database A. The calculations were then repeated using the pre-sedation value of GCS, producing an alternative database, B.

7.2.1 Data analysis

The performance of the two databases was then compared. Discrimination was again assessed using the area under the ROC curve (164) and the DeLong method was used for statistical assessment of the differences (167). The goodness of fit of the actual and predicted data was assessed using calibration curves and the Hosmer-Lemeshow statistic (163,169). In addition to an assessment of the overall effects, an analysis was made of the performance within the sub-group of patients with altered scores. The pattern within different primary APACHE diagnostic categories was also studied.

As there are differences in the sample sizes of different groups within this analysis, when assessing the differences in database A and B, comparisons have only been made where the sample sizes are the same. This is particularly important with the Hosmer -Lemeshow test as it is particularly sensitive to sample size (as previously discussed in Chapter 3).

7.3 Results

A total of 13,291 patients were admitted to the participating units during the study period. The 1,654 patients classed as HDU or CCU admissions were excluded, as were 498 re-admissions and 114 patients for whom outcome data was not available. The specific exclusion criteria for APACHE II and III were also applied. These criteria differ slightly, so scores and predictions were available for 9,848 patients using APACHE II and for 10,326 using APACHE III. The demographic details of the patients are shown in Table 7.1.

Table 7.1 Demographics of the study populations

	APACHE II group (n=9,848)	APACHE III group (n=10,326)
Age range (yrs) (mean)	16-99 (59)	16-99 (59)
Male (%)	5,457 (55%)	5,714 (55%)
Non-operative (%)	4,939 (50%)	5,030 (49%)
Sedated, GCS not recorded (%)	4,965 (50%)	5,124 (50%)
Sedated patients with pre-sedation GCS < 15 (%)	2,066 (21%)	2,281 (22%)

APACHE, Acute Physiology and Chronic Health Evaluation; GCS, Glasgow Coma Score.

Of the 9,848 patients with APACHE II predictions, 4,965 were recorded as sedated and 2,066 of these had a pre-sedation GCS less than 15. The APACHE II scores of 21% of all patients were therefore altered when the pre-sedation GCS (Method B) was used. The effects of this are summarised in Table 7.2. In the APACHE III database, 2,281 (22%) of 10,326 patients had their scores altered and the effects are summarised in Table 7.3. The increase in the mean APACHE scores is inevitable, since incorporating a pre-sedation GCS of less than 15 into an APACHE score must increase it. It follows that the predicted mortality will increase and that the mortality ratio, the ratio of observed to predicted mortality, must therefore be reduced.

Table 7.2 The effect on APACHE II score and performance

	APACHE II score (mean)	Mortality ratio	Hosmer-Lemeshow GOF	Area under ROC curve
All patients (n=9,848)				
Method A	18.3	0.953	36.39	0.805[a]
Method B	19.4	0.874	132.97	0.816[a]
Patients with altered scores (n=2,066)				
Method A	19	1.26	144	0.764[b]
Method B	25	0.89	65	0.771[b]

APACHE, Acute Physiology and Chronic Health Evaluation; GOF, Goodness of fit; ROC, Receiver Operaing Characteristic.
a P<0.0009 for difference between method A and method B
b P=0.13 not significant between method A and method B

Table 7.3 The effect on APACHE III score and performance

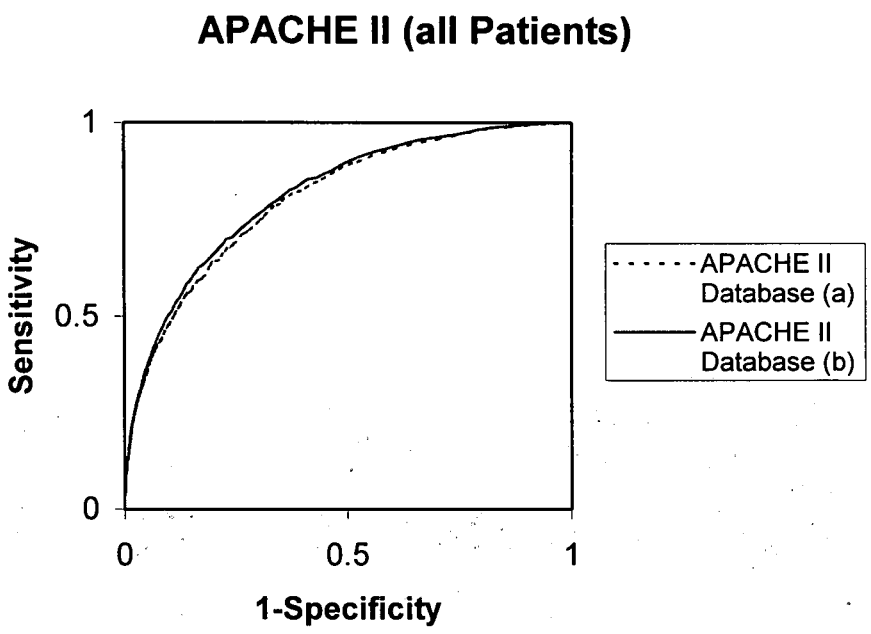
	APACHE III score (mean)	Mortality ratio	Hosmer-Lemeshow GOF	Area under ROC curve
All patients (n=10,326)				
Method A	60.8	1.23	331.65	0.845[a]
Method B	64.3	1.09	126.89	0.852[a]
Patients with altered scores (n=2,281)				
Method A	64	1.51	518	0.809[b]
Method B	81	1.00	47	0.805[b]

APACHE, Acute Physiology and Chronic Health Evaluation; GOF, Goodness of fit; ROC, Receiver Operaing Characteristic.
a $p<0.0001$ for difference between method A and method B
b $p=0.55$ not significant between method A and method B

There is a significant increase in the area under the ROC curve for APACHE II (Method B) compared to the original model (Method A), indicating improved discrimination (Table 7.2, Figure 7.1). The value of the Hosmer-Lemeshow statistic in database is increased in APACHE II (Method B) compared with the original model (Method A) suggesting that the fit of the data is poorer. The calibration curves (Figure 7.2) confirm this. The curve for APACHE II lies significantly below the line of equality (which represents the situation where actual mortality and estimated mortality are equal) when the estimated mortality is between 30% and 70%. This indicates that actual mortality is less than estimated. Since using the pre-sedation GCS increased estimated mortality, it must also increase this difference. Examination of the performance in the patients with altered scores however (Table 7.2) shows that for these patients there is an improvement in both the Hosmer-Lemeshow statistic and the calibration curve (Figure 7.3) although the increase in area under the ROC curve is not statistically significant.

The discrimination of APACHE III was also improved (Table 7.3, Figure 7.4), as was the Hosmer-Lemeshow statistic and the calibration curve. The calibration curve for APACHE III for the entire population (Figure 7.5) lies significantly above the line of equality for most of its length i.e. it underestimates mortality. The use of pre-sedation GCS moves the calibration curve closer to the diagonal, and for an estimated mortality of greater than 50% it is within the 95% confidence limits. If the calibration and goodness of fit are again analysed for the patients with altered scores (Figure 7.6, Table 7.3) an improvement in this group is confirmed.

Figure 7.1 ROC curves for APACHE II (all patients) for Database (a) and Database (b) and APACHE II (patients with altered scores) for Database (a) and Database (b)



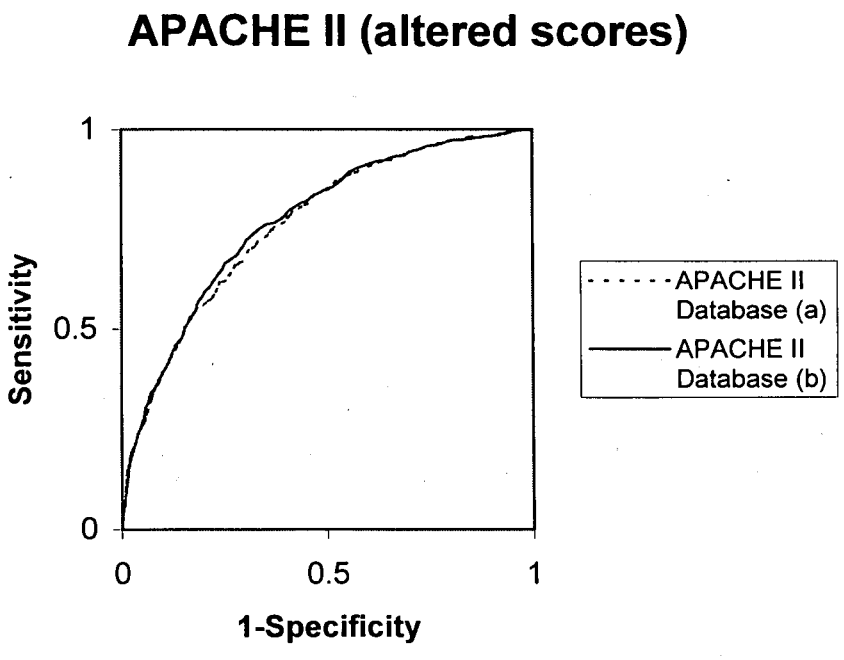
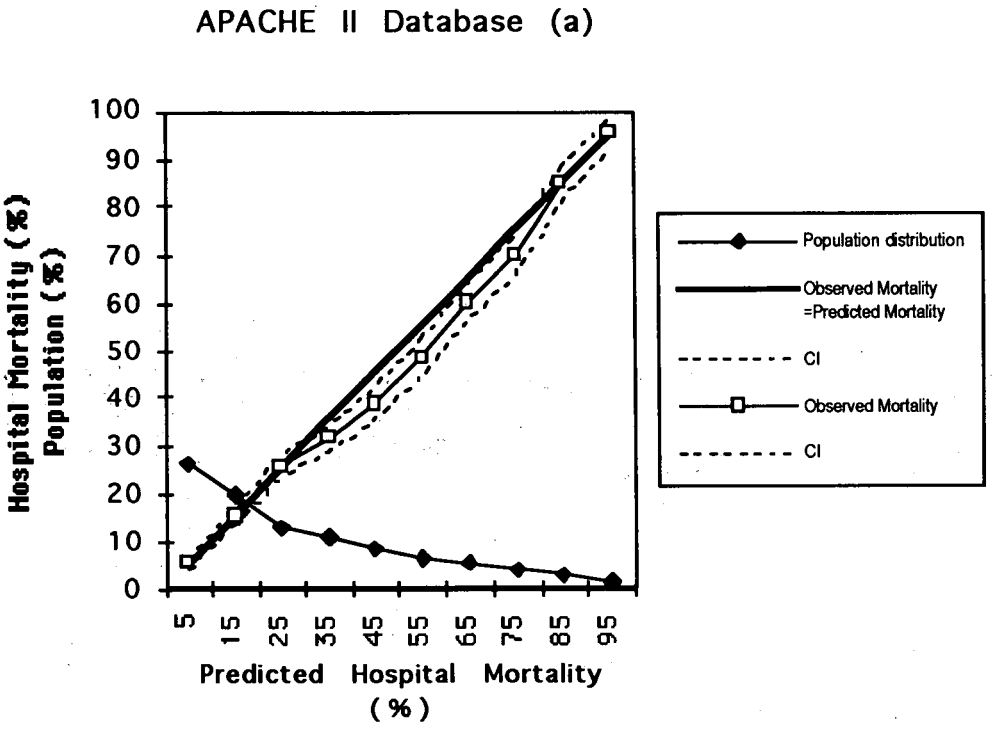


Figure 7.2 Calibration curves for APACHE II (a) assuming a normal GCS in sedated patients and APACHE II (b) using the pre-sedation GCS value for sedated patients.



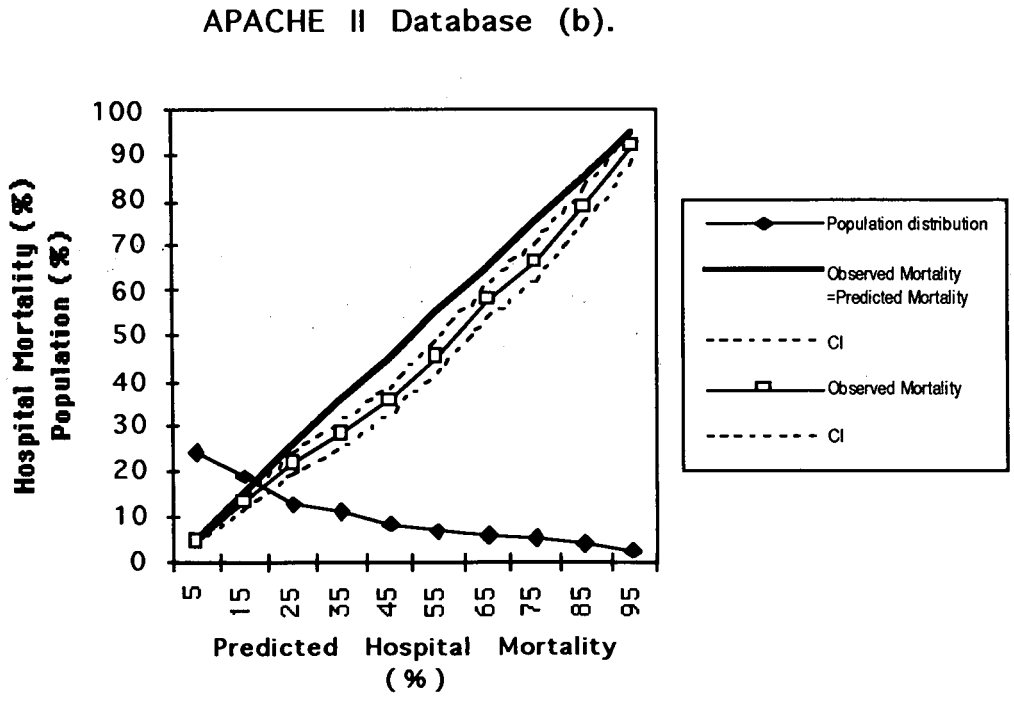
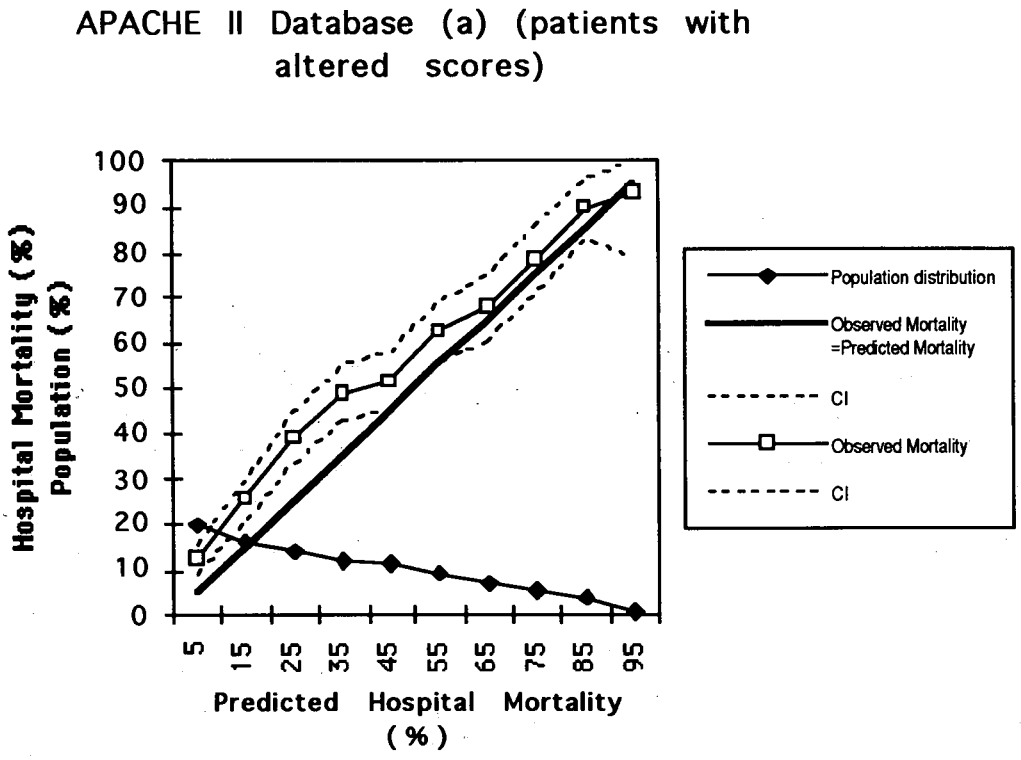


Figure 7.3 Calibration curves for APACHE II (a) (patients with altered scores) assuming a normal GCS in sedated patients and APACHE II (b) (patients with altered scores) using the pre-sedation GCS value for sedated patients.



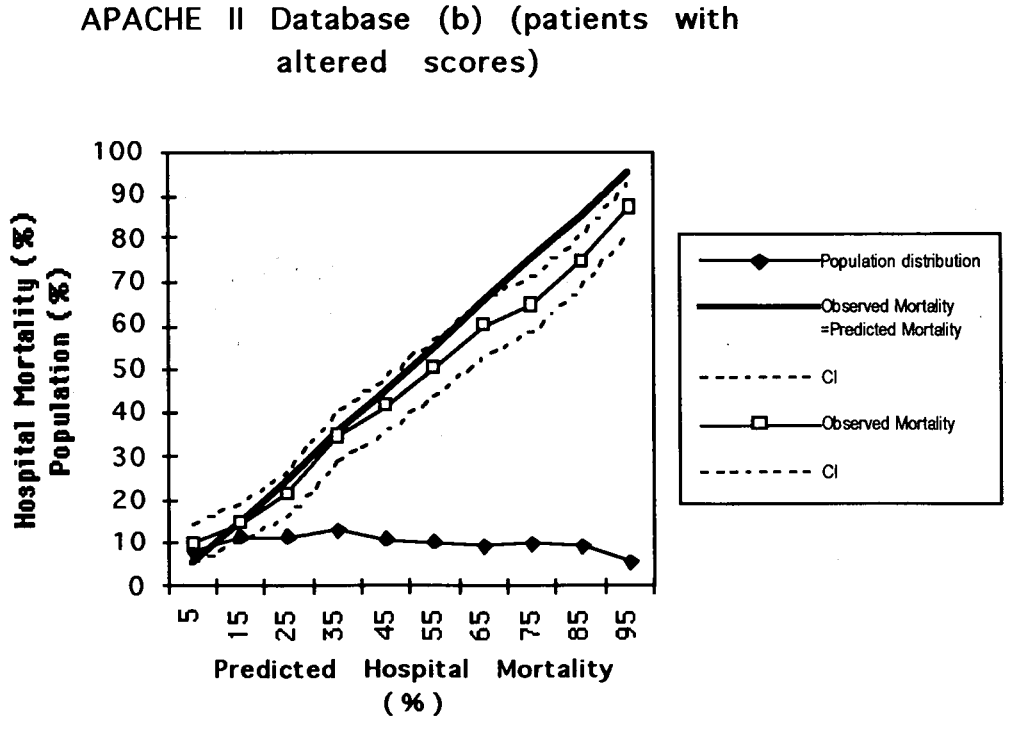
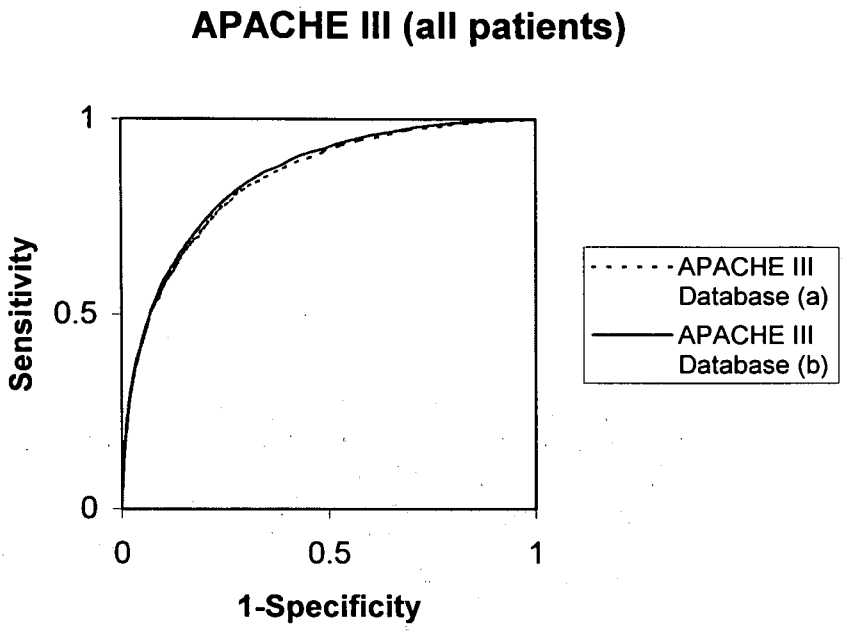


Figure 7.4 ROC curves for APACHE III (all patients) for Database (a) and Database (b) and APACHE III (patients with altered scores) for Database (a) and Database (b).



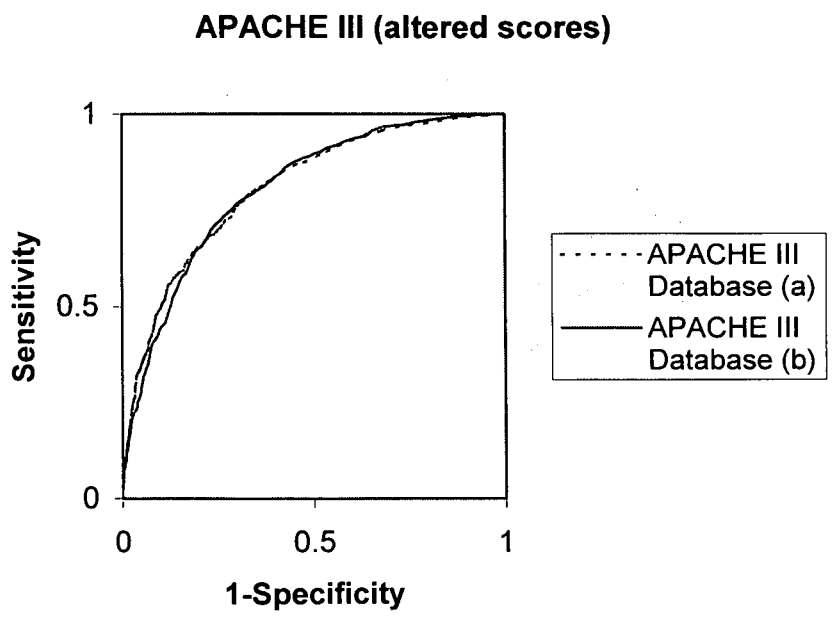
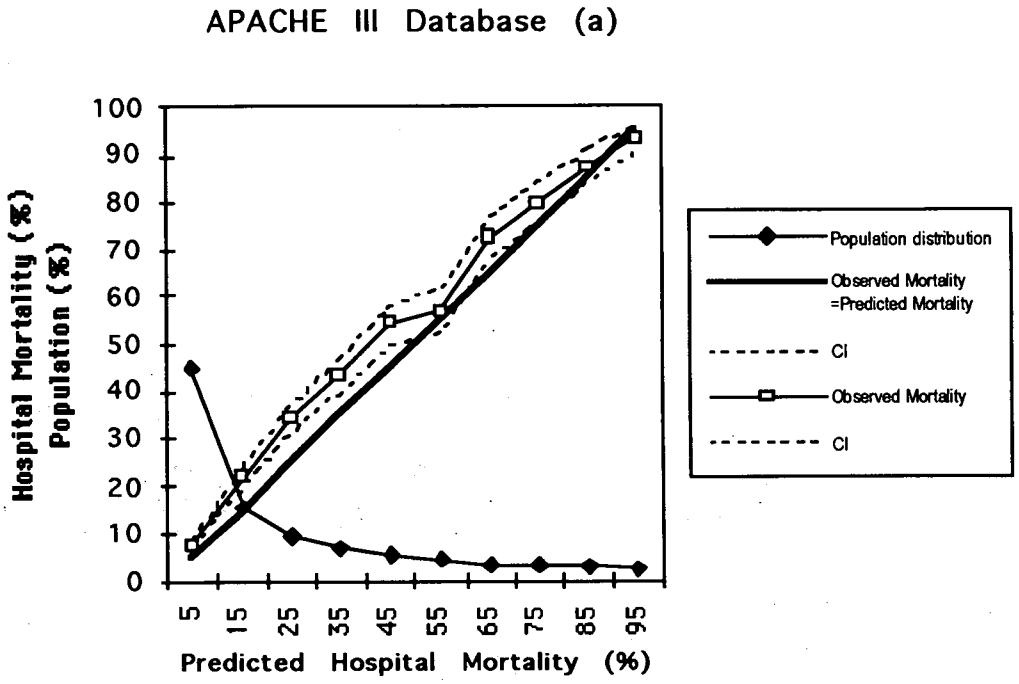


Figure 7.5 Calibration curves for APACHE III (a) assuming a normal GCS in sedated patients and APACHE III (b) using the pre-sedation GCS value for sedated patients.



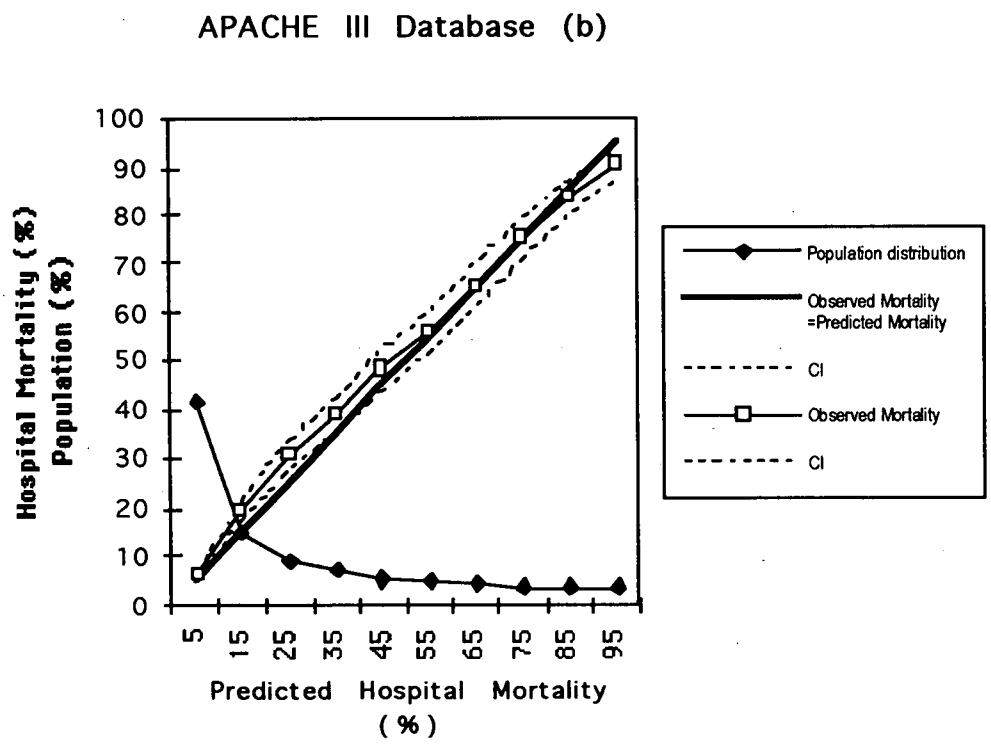
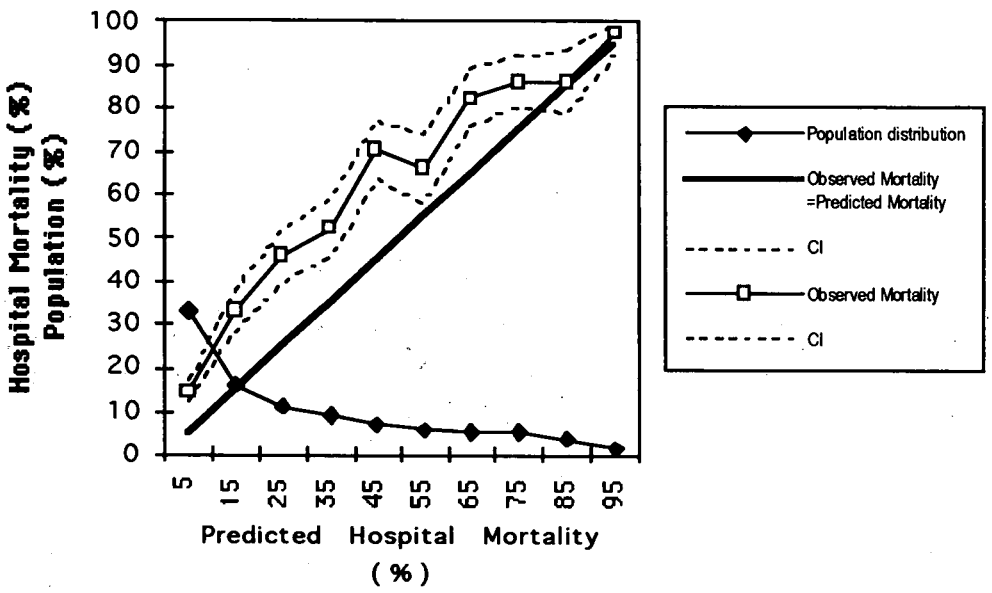


Figure 7.6 Calibration curves for APACHE III (a) (patients with altered scores) assuming a normal GCS in sedated patients and APACHE III (b) (patients with altered scores) using the pre-sedation GCS value for sedated patients.

APACHE III Database (a) (patients with altered scores)



APACHE III Database (b) (patients with altered scores)

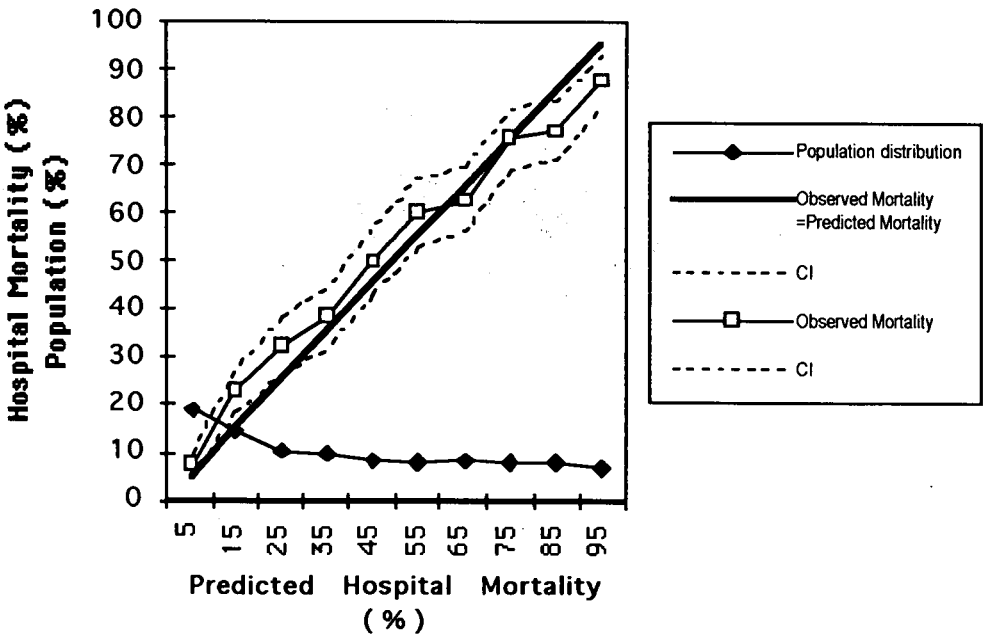


Table 7.4 Patients sedated and with altered scores in APACHE III database

APACHE diagnostic system category	No. of patients	No. sedated (%)	No. with score changed (%)
Cardiovascular	2,281	1,123 (49)	490 (21)
Respiratory	1,918	1,022 (53)	534 (28)
Neurological	926	466 (50)	395 (43)
Gastrointestinal	3368	1,735 (52)	522 (15)
Renal	312	109 (35)	31 (10)
Metabolic/endocrine	130	46 (35)	25 (19)
Haematological	53	22 (42)	13 (25)
Trauma	738	399 (54)	219 (30)
General	600	202 (34)	52 (9)
Overall	10,326	5,124 (50)	2,281 (22)

APACHE, Acute Physiology and Chronic Health Evaluation.

Table 7.5 The effect on actual and predicted mortality, APACHE II

APACHE diagnostic system category	No. of patients	Mortality ratio	
		Method A	Method B
Cardiovascular	2,131	1.09 (1.05-1.14)	1.01 (0.97-1.05)
Respiratory	1,846	1.04 (0.98-1.10)	0.93 (0.88-0.99)
Neurological	850	1.35 (1.25-1.45)	1.01 (0.92-1.09)*
Gastrointestinal	3,276	0.77 (0.73-0.81)	0.75 (0.71-0.79)
Renal	299	0.81 (0.62-0.99)	0.78 (0.60-0.96)
Metabolic/endocrine	126	1.04 (0.80-1.29)	0.94 (0.72-1.16)
Haematological	48	1.08 (0.80-1.37)	1.05 (0.75-1.33)
Trauma	697	1.18 (0.99-1.37)	0.89 (0.73-1.05)
General	575	0.49 (0.32-0.66)	0.46 (0.29-0.62)

APACHE, Acute Physiology and Chronic Health Evaluation.
*CIs for Method A and Method B that don't overlap

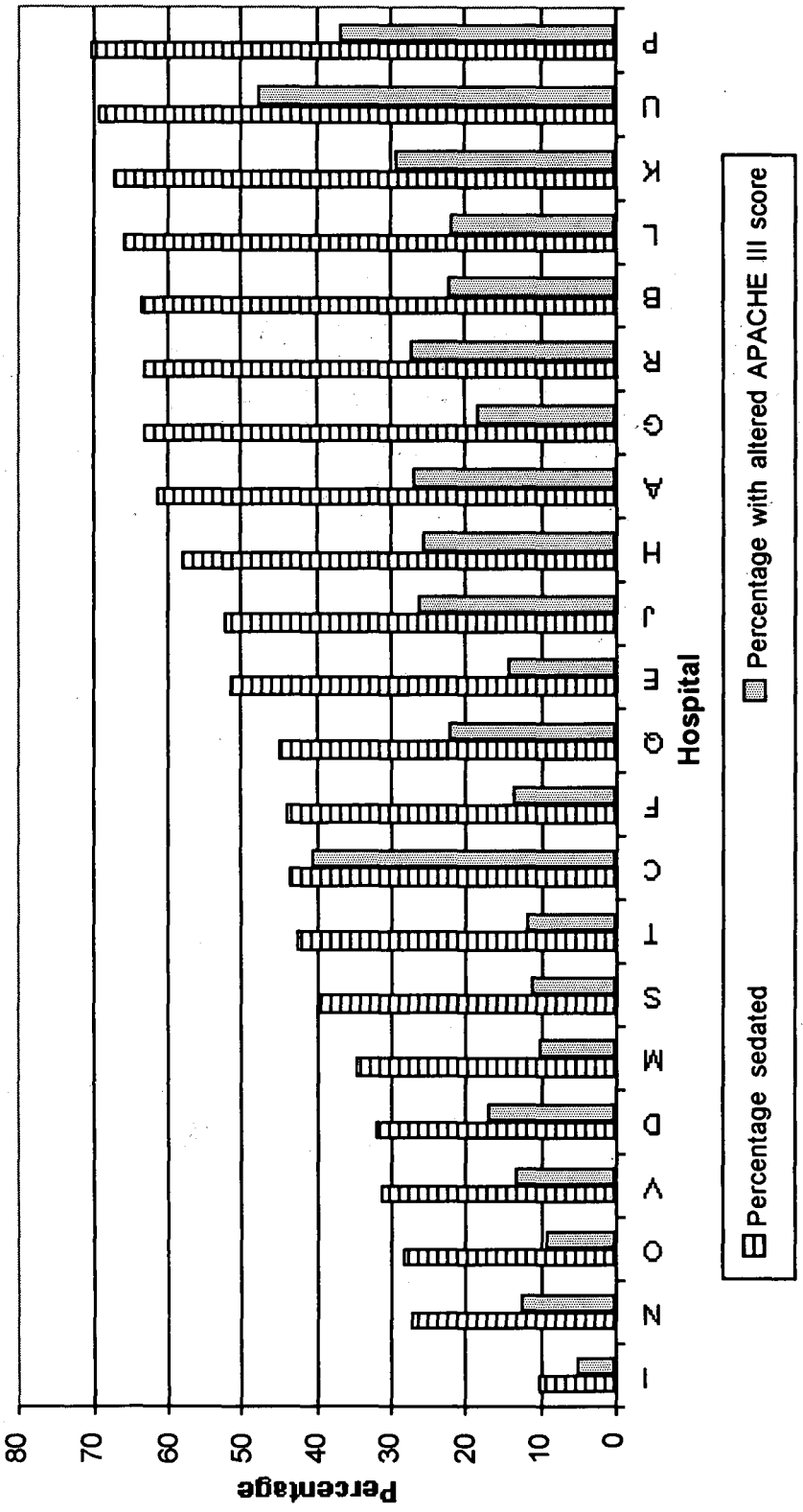
The effects of these changes were not confined to any particular diagnostic group. Although those patients whose primary APACHE diagnostic category was neurological or trauma were particularly likely to be affected, they accounted for only 27% of all altered scores. The number of patients sedated and with altered scores for each of the primary APACHE diagnostic categories are shown in Table 7.4 for APACHE III. The figures for APACHE II are almost identical. The effects on the ratio of actual to observed mortality for each category are shown in Tables 7.5 and 7.6. The proportion of patients who were sedated (10-69%) and who had their APACHE scores altered (5-47%) varied markedly amongst the participating units (Figure 7.7).

Table 7.6 The effect on actual and predicted mortality, APACHE III

APACHE diagnostic system category	No. of patients	Mortality ratio	
		Method A	Method B
Cardiovascular	2281	1.25 (1.21-1.30)	1.14 (1.11-1.18)*
Respiratory	1918	1.20 (1.14-1.26)	1.07 (1.02-1.13)*
Neurological	926	1.44 (1.34-1.53)	1.04 (0.96-1.12)*
Gastrointestinal	3368	1.17 (1.12-1.22)	1.13 (1.08-1.18)
Renal	312	1.14 (0.94-1.33)	1.11 (0.92-1.31)
Metabolic/endocrine	130	1.25 (1.02-1.48)	1.12 (0.90-1.34)
Haematological	53	1.88 (1.50-2.27)	1.84 (1.46-2.22)
Trauma	738	1.34 (1.17-1.52)	1.00 (0.85-1.14)*
General	600	0.85 (0.63-1.08)	0.80 (0.58-1.02)

APACHE, Acute Physiology and Chronic Health Evaluation.
*CIs for Method A and Method B that don't overlap

Figure 7.7 Percentage of patients sedated and with altered APACHE III scores for each ICU.



7.4 Discussion

The association between an impaired conscious level and mortality in Intensive Care is well recognised (183). Although the validity of using the GCS in general ICU patients has been questioned (77), there is a significant relationship between the GCS and outcome in such patients and the overall predictive capability of the APACHE III system is improved by incorporating the GCS (182). As already mentioned, the GCS is potentially the largest single contributor to the acute physiology score. There is no doubt that the GCS should be assessed directly whenever possible.

The purpose of the development of GCS was to facilitate consistent and reliable recording of level of consciousness, but difficulties do arise in practice. Ideally the GCS should be recorded in a patient who is fully resuscitated and is not sedated. Concern has been expressed about inconsistency of approach in patients with head injuries presenting to A&E (46) since modern practice emphasises early intubation and sedation. Similar problems clearly occur in the ICU where sedation and paralysis are commonly employed. In this study population, almost 50% of admissions were classed as being sedated for the first 24 hours.

There are four possible approaches to scoring an ICU patient who is sedated and in whom it is not possible to stop sedation. The first approach is to record the GCS observed. This is likely to result in an artificially low score, which is at least partly attributable to the sedative drugs rather than the underlying physiological disturbance. The second is to make a best guess at the underlying GCS with clear potential for systematic or random error. The third is to assume that the GCS is normal and the fourth is to assume that the GCS is unchanged from the pre-sedation value.

There is little support for either of the first two approaches and they have not been considered here. The third approach, used in APACHE II and III, is consistent with the general practice that missing values are recorded as normal. The reason for this is if there is no measurement of the PaO_2 then it is likely that the clinician was not concerned about it and it probably was normal. The same assumption cannot be made when the GCS is absent, indeed it may well be that the patient is kept sedated because of concern about their neurological condition. The main argument in favour of this approach is that although it runs counter to many clinicians' instincts and may underestimate the risk of death for an individual patient, it is, at least, consistent (49). When the number of sedated patients is large, however, there may be a considerable overall effect, which may not be equally distributed across different ICUs or diagnostic groups. The proportion of patients

sedated between units can vary considerably (0-26% in APACHE III (182)), 10% to 70% in the present study. This may hinder any comparisons between units.

The fourth approach, adopted in SAPS II, allows account to be taken of a previously documented abnormal level of consciousness but this has its own deficiencies. It assumes that the GCS has not changed since sedation was begun. The value ascribed to the pre-sedation GCS may be unreliable due to observer error or incomplete resuscitation. It is a single point measurement, whereas the principle of recording the worst in first 24 hours has gained general acceptance. Finally, it may have been recorded some time prior to ICU admission.

The choice of method is an important issue. In the present study 50% of patients were sedated, and 40% of these sedated patients had their APACHE scores changed. This had a significant effect on the performance of both APACHE II and III.

Ideally a severity scoring system would have both good discrimination and calibration. APACHE III was improved by the adoption of pre-sedation values of GCS rather than assuming a normal value. The calibration of APACHE III was also improved but the calibration of APACHE II was poorer. The difference here is related to the initial calibration of the two systems to the Scottish database. Since APACHE II overestimates mortality even when the GCS in sedated patients is taken as normal and using the pre-sedation GCS increases predicted mortality, this deterioration is inevitable. It does however require explanation since it suggests use of the pre-sedation GCS may be inappropriate. A clear improvement in the calibration seen when the patients with altered scores are analysed separately confirms the benefit of using pre-sedation GCS. It is probable that there are other limitations in the applicability of the existing APACHE II model in this data, normally concealed, that are exposed by the improvement seen in this group. The patients with a neurological or trauma primary diagnosis were most likely to have their scores affected, and the greatest changes in mortality ratios were also seen in these groups. It is noteworthy that it is in precisely these groups that the performance of both APACHE II and APACHE III has been least satisfactory in previous studies in the UK (20,83). It seems that the balance for these data favours the use of the pre-sedation value. It has the further advantage of reducing the spread of actual to observed mortality between diagnostic groups which will tend to reduce the effect of differing case mix on performance.

No previous published studies, as far as the author is aware, have examined this issue, and it is always difficult to generalise from a single study. The database is a large one, encompassing almost all general Intensive Care admissions in Scotland over this period

and the distribution of diagnoses was broadly similar to that of the UK APACHE II study (20,73). The population is probably representative of UK practice. The fact that the clinicians felt unable to assess the GCS in almost 50% of the patients was surprising, but it is difficult to know if this is an atypical figure. In the UK APACHE II study 52% of patients were sedated when the lowest GCS was estimated and 26% were sedated for all GCS measurements (K Rowan, personal communication). The great majority of the patients in the present study (70%) were ventilated and this may represent, unintentionally, a form of case selection compared to the original APACHE II and III databases. Despite the trend to lighter sedation in current practice it is not always possible to assess the GCS in such patients. If the pre-sedation values for these patients were nearly all 15 then the approach taken would have little practical impact, but in fact 40% of the sedated patients had a GCS less than 15 and using these values had a significant effect on scores, mortality prediction and performance.

Whilst it is important to re-emphasise the importance of recording the GCS accurately whenever possible, when sedation prevents this then the pre-sedation GCS, it would appear, should be used to calculate APACHE scores and mortality predictions. This approach improves the performance of the system and reduces one possible source of error in the mortality ratios for individual units.

It is clear that if any customisation or remodelling is to be done in this population then it should consider the use of a pre-sedated GCS where a normal GCS is not available.

Chapter 8- Customisation of Models.

Aims: To improve the performance of the existing models in the Scottish setting

To assess the ability of customisation to improve the performance of different severity of illness models

Contents:

8.1 Introduction

The models in this study have been tested and evaluated in several countries with varying results (53,73,83,84,102,110,184). The probabilities generated by all these models are based on logistic regression coefficients from their original data cohorts and therefore represent the weights for the variables in those data. These weights may not be representative of all ICU cultures and there is widespread agreement that these models must first be validated before being applied in different ICUs (21,22,83).

Results from previous chapters (Chapters 4-6) demonstrated significant differences between the predicted and observed mortality (Hosmer- Lemeshow goodness of fit test) in all models. The conclusions drawn from analysis of these data confirms that before any valid comparisons of outcome on different ICUs in Scotland could take place, the models needed to be customised to better reflect the mortality experience in Scottish ICUs. There are a number of studies where customisation has been shown to improve the accuracy of the MPM_0 part of the MPM II model (139,142). In a large study of over 11,000 patients after a process of customisation both the MPM_{24} and the SAPS II models had improved goodness of fit (141).

In an attempt to improve the accuracy and validity of four of the scoring models (APACHE II, SAPS II, MPM_0 and MPM_{24}) analysed in Chapters 4-6, the data from this study have been used to customise these models.

8.2 Materials and Methods

8.2.1 Added data

Data from a further six months of the study were added to the two years of data already collected. The same data protocols and conditions applied to these six months as in the first two years of data collection. It was hoped that by adding a further six months data this would help to generate more representative models with the extra observations. As some diagnostic categories and variables represent small numbers of patients, it was hoped that by adding a further six months data this would help to produce coefficients that were a true reflection of these patients.

8.2.2 Outcome measures

Patient status on hospital discharge was the outcome used in the original development of all models. However, patients being transferred to other ICUs would be recorded as

having survived, though these patients may not have survived after treatment in the second ICU. In the original analysis (Chapter 4-6) it was important to test the models using the original rules and outcome measures. However, the customisation process did not constrain the study to a particular outcome and the use of patient status at the end of hospital stay, being defined as one continuous stay in any hospital, was a possible alternative outcome measure.

The Information and Statistics Division's Scottish Morbidity Record (SMR) database routinely links all mortality data and all hospital admission records. All acute admissions to hospitals in Scotland require an SMR1 return to be generated on discharge. SMR1 records and the Registrar General's Office (RGO) mortality records are linked together for individual patients using probability matching. The accuracy of the linkage of these records to each other and to the GRO's death records is at least 99% (185).

Patients for the first year were linked to the Information and Statistics Division's linked database to extract status at the end of hospital stay. Of 669 patients transferred to other hospitals in 1995, 97 patients (14.5%) died before hospital discharge. This represents 1.96% of all patients entered in to the study in 1995, and 6.8% of the deaths in 1995. Record linkage was not available past 1995. An attempt was made to obtain status at the end of hospital stay but the data were difficult and time consuming to collect. The numbers suggest that the impact on the overall population was small but the size of the error may be larger for individual units. However, as end of hospital stay was not available for all patients, status on hospital discharge was chosen as the dependent variable in the customisation process. Status was defined as dead or alive on discharge from hospital.

8.2.3 Development and validation cohorts

All patients were allocated a random number between 1 and 100. The data were then divided approximately in half by allocating all patients with a random number of 50 or less (development cohort) into one group and the remaining patients forming the other (validation cohort). The development cohort was used to generate the new models and the validation cohort was used to test these new models. The data were grouped in a ratio of 50:50 to ensure that analysis was not weighted in favour of either the model building or validation exercise. This also allowed valid comparisons when using goodness of fit tests, which are sensitive to sample size, as it created cohorts of roughly equal size.

8.2.4. Data analysis

Models were customised using forward stepwise logistic regression. Variables in the regression that were not significant were not included in the final models. All regression was carried out using the SPSS software (SPSS Inc, Chicago, Illinois, US). Comparisons were then made between the performance of the original models and performance of the new customised models. Discrimination was assessed using the area under the ROC curve (164,186) with the DeLong method (167) being used to test for significant differences between areas under the curves. Calibration was assessed using the Hosmer-Lemeshow GOF C statistic (169) and calibration curves (see Chapter 3.9.2). As discussed in Chapter 5, it is important that the accuracy of the severity of illness models is maintained in large sub groups (115,129). Mortality ratios with their CIs were used to assess this uniformity of fit (See Chapter 3.9.3). CIs that did not overlap with the CIs for the whole population in the relevant models were treated as demonstrating significant differences. The Chi Squared test was used to test differences between estimated mortality and observed mortality in subgroups (see Chapter 3.9.4).

To convert the Logit (g(x)) to hospital mortality the following calculation was used:
$Pr(\text{Hospital mortality})=[e^{g(x)}/1+e^{g(x)}]$

8.2.4.1 APACHE II

Two new APACHE II models were created in the process of customisation. The first model used the APACHE II score as originally reported (New APACHE (A)). A second model was created using an adapted score which incorporated a pre-sedated GCS if the patient was sedated for the first 24 hours (New APACHE (B)). Analysis from Chapter 7 shows that using a pre-sedated GCS improves the APACHE II model. Three variables were entered into the regression to produce the new model, the APACHE II score, admitting diagnosis, and emergency operative status. The original APACHE II model was not calculated unless the patient stayed for longer than eight hours, this exclusion was not applied in the customisation study, and patients staying longer than an hour were included in the study if fulfilling all other criteria.

The study did not collect the APACHE II diagnosis as described in the original paper. The APACHE III list of diagnoses (230 separate diagnoses) was used to collect admitting diagnosis. Only diagnoses chosen for more than twenty patients were entered into the new models, remaining diagnoses with less than twenty patients were allocated to the general diagnostic groups (i.e. other respiratory).

The logistic regression equations for the two new APACHE models were as follows:
New APACHE II (A) = Logit=β_0+(APACHE II score*ß)+(Diagnostic ß)
+(Post-emergency surgery ß).
New APACHE II (B) = Logit=β_0+(APACHE II score(GCS amended)*ß)
+(Diagnostic ß).

8.2.4.2 SAPS II

The SAPS II score, post-emergency surgery, and, as in the original published paper, a log of the score were included in the regression to generate the New SAPS II model (new SAPS II (A)).

The logistic regression equation for the new SAPS II (A) model was as follows:
Logit=β_0+(SAPS IIscore*β_1)+(Log(SAPS IIscore+1) *β_2) +(Post-emergency surgery β_3).

8.2.4.3 MPM II

To generate a New MPM_0 model (New MPM_0(A)) all the variables, collected on admission to the ICU, from the original paper were entered into the regression. As in the published methodology all the variables collected on admission and those entered after 24 hours were entered into the regression for the New MPM_{24} model (New MPM_{24}(A)).

The logistic regression equations for the new MPM_0 (A) and the new MPM_{24} (A) models were as follows:
Logit $g(x)=\beta_0+\beta_1x_1+\beta_2x_2+.....\beta_kX_k$
where β_0 is the constant and β_ix_i is the estimated coefficient for the *i*th variable times the value of the *i*th variable, with i taking the values 1 to k and k being the number of variables in the model.

Previous studies customising the MPM model had also used a simpler method entering only the logit from the original model (139,142). To allow comparisons with these studies a second model was calculated for both MPM_0 (New MPM_0 (B)) and MPM_{24} (New MPM_{24} (B)) by entering the logit from the original score as the only independent variable .

The logistic regression equations for the new MPM_0 (B) and the new MPM_{24} (B) models was as follows:
Logit:=β_0+(ß*[Study]MPMLogit)

8.2.4.4 Selection check

A further random 50:50 split was carried out to check the robustness of the customisation process and to test the presumed stability of the models. The second random selection was used to calculate new coefficients for the APACHE II model incorporating a pre-sedated GCS. The same methods were used as in 8.2.4.1 with Hosmer-Lemeshow tests and ROC curves used to test the performance of the new APACHE II (check) model.

8.3 Results

8.3.1 Patients

A total of 16,701 patients were admitted to the ICUs participating in the study. Study exclusions, readmissions, HDU and patients with missing data, amounted to 3,693 leaving 13,008 patients for analysis (Table 8.1). Demographic details for the overall population, the development and validation cohorts can be seen in Table 8.2. Both the development cohort and the validation cohort had similar characteristics to the overall population. Importantly the severity of illness when measured by the SAPS II score and APACHE II score remained of a similar magnitude to the overall population.

8.3.2. APACHE II Models

Analysis for the APACHE II models used 12,936 patients, 6,422 in the development cohort and 6,514 in the validation cohort, with 72 patients excluded because of missing information.

Table 8.1 Study exclusions

Reason for Exclusion	Number of patients
Readmissions	637
Missing outcomes	150
Prospective study exclusion categories	817
HDU patients	2,089
Total	3,693

HDU, High Dependency Unit.

Table 8.2 Demographics

	All Patients n=13,008	Model Group n=6,458	Validation Group n=6,550
Age range (yrs) (mean)	16-99 (59)	16-99 (59)	16-99 (59)
Male (%)	7,163 (55)	3,548 (55)	3,615 (55)
Mean SAPS II score	39.58	39.51	39.66
Mean APACHE II score	18.38	18.32	18.44

SAPS, Simplified Acute Physiology Score; APACHE, Acute Physiology and Chronic Health Evaluation.

In the model New APACHE II (B) the variable "emergency surgery" was not significantly associated with the dependent variable (hospital mortality) and was therefore not included in the model.

The coefficients for the new APACHE II models are as follows (Table 8.3):

<u>New APACHE II (A)</u>	
New APACHE II (A) constant (β_0)	-3.8822
APACHE II (A) score	+0.1587
Emergency Operative	0.3984
<u>New APACHE II (B)</u>	
New APACHE II (B) constant (β_0)	-4.2952
APACHE II (B) Score	+0.1694

Table 8.3 Diagnostic coefficients for new APACHE II models

Diagnostic codes	op-nonop*	PtNo	New APACHE II (A)	New APACHE II (B)
Respiratory				
230 Respiratory arrest	N	88	0.0856	-0.2089
1 Neoplasm-mouth/sinuses	O	47	-0.8437	-0.5324
2 Neoplasm-larynx/trachea	O	33	-1.0666	-0.8253
3 Neoplasm-lung parenchyma	O	23	-0.2835	-0.065
8 Other respiratory surgery	O	50	-1.3997	-1.1845
9 ARDS (non cardiogenic pulmonary oedema)	N	52	0.619	0.6001
10 Pneumonia-viral	N	24	0.4298	0.4754
12 Pneumonia-bacterial	N	375	0	0
14 Pneumonia-aspiration/toxic	N	73	0.2492	0.1162
18 Pulmonary embolus	N	24	0.111	0.1343
20 Localised airway obstruction/oedema (mechanical)	N	36	-0.4347	-0.4774
21 Emphysema	N	67	-0.4549	-0.4649
22 Asthma	N	134	-1.5623	-1.4618
23 Smoke inhalation	N	28	-0.3391	-0.3757
25 Other respiratory disorder	N	187	-0.0459	0.0207

Cardiovascular				
27 Aorto-femoral, fem-fem bypass graft	O	97	-0.3581	-0.0992
29 Aortic aneurysm: pre-leak/dissection	O	143	-0.8961	-0.6899
30 Aortic aneurysm: dissection	O	20	-1.0327	-0.5151
31 Aortic aneurysm: rupture	O	141	-0.6262	-0.134
32 Peripheral ischaemia	O	38	-1.2287	-0.8971
51 Other cardiovascular surgery	O	119	-0.555	-0.2688
53 Aortic aneurysm	N	30	0.0462	0.0374
57 Rhythm disturbance	N	29	-0.2835	-0.5435
58 Acute myocardial infarction	N	23	0.6536	0.495
60 Congestive heart failure	N	126	-0.6489	-0.6403
61 Cardiogenic shock	N	117	1.0993	1.0758
64 Septic shock - lungs (pneumonia)	N	49	0.4502	0.2772
65 Septic shock - urinary tract infection	N	24	0.0372	0.0737
66 Septic shock - gastrointestinal tract	N	57	0.5327	0.4791
67 Septic shock - unknown origin	N	66	0.0427	-0.0252
69 Post cardiac arrest (± respiratory arrest)	N	260	1.1342	0.7538
73 Other cardiovascular disorder	N	84	-0.6772	-0.6816
Neurological				
74 Subarachnoid haemorrhage/intracranial aneurysm	O	29	0.8665	0.5716
75 Subdural/epidural haematoma	O	28	0.0729	-0.8359
82 Other neurosurgery	O	44	-0.3672	-0.5464
83 Subarachnoid haemorrhage/intracranial aneurysm	N	58	1.4534	1.0298
85 Intracerebral haemorrhage/haematoma	N	29	1.3838	0.8867
86 Cerebrovascular accident (CVA)/stroke	N	22	1.8432	1.6839
88 Seizures	N	74	-1.2772	-1.6031
92 Self-inflicted overdose	N	171	-2.2551	-2.2793
98 Non traumatic coma - cause unknown	N	22	-0.8737	-1.0917
99 Other neurological disorder	N	119	0.0299	-0.3233
Gastrointestinal				
110 Bleeding - ulcer	O	97	-0.17	0.3667
111 Bleeding - laceration/tear	O	25	-1.525	-1.3084
115 GI perforation/rupture	O	323	-0.5469	-0.1226
116 GI obstruction (any cause)	O	222	-0.7943	-0.3069
117 GI neoplasm (not perforation/obstruction)	O	439	-0.4572	-0.1434
118 Localised GI abscess/cyst	O	41	-0.7188	-0.2258
119 Peritonitis	O	70	-0.686	-0.1452
121 Cholangitis/cholecystitis	O	75	-1.5658	-1.1786
123 GI vascular insufficiency/embolism/infarction	O	69	0.5431	1.0942
124 GI inflammatory disease	O	53	-1.6879	-1.3188
125 Liver transplant	O	29	-1.8185	-1.7582
128 Other GI surgery	O	193	-0.3567	0.0171
129 Bleeding - ulcer	N	28	0.635	0.6855
131 Bleeding - varices	N	32	0.9495	0.6385
134 GI perforation/rupture	N	62	0.3575	0.4233
135 GI obstruction (any cause)	N	46	0.4924	0.5994
136 GI neoplasm (not perforation/obstruction)	N	39	-0.5951	-0.4306
138 Peritonitis	N	28	1.6085	1.6305
139 Pancreatitis	N	73	0.3084	0.3512
148 Hepatic failure - drug overdose	N	26	1.4763	0.7312

150 Other GI disorder	N	112	0.1852	0.1072
Renal				
157 Renal neoplasm	O	53	-0.5892	-0.3499
163 Other renal surgery	O	76	-0.9171	-0.6446
172 Other renal disorder	N	60	-0.3938	-0.4102
Metabolic/endocrine				
175 Other metabolic/endocrine surgery	O	26	-5.6774	-5.2667
187 Other metabolic endocrine disorder	N	51	-0.1379	-0.2333
Haematological				
190 Other haematological surgery	O	5	0.5649	0.8382
196 Other haematological disorder	N	31	0.5242	0.6359
Trauma				
199 Trauma - face	O	24	-0.6509	-0.1606
203 Trauma - extremities	O	72	-1.3135	-0.8928
204 Trauma - multiple sites plus head/brain	O	25	-0.5806	-0.3094
205 Trauma - multiple site without head/brain	O	39	-2.5173	-2.1282
209 Trauma - head/brain	N	47	-0.4306	-0.9344
212 Trauma - chest	N	45	-0.6287	-0.4448
216 Trauma - multiple sites plus head/brain	N	78	-0.0137	-0.5668
217 Trauma - multiple site without head/brain	N	45	-0.3644	-0.1725
General				
Obs and gynae				
222 Pre-eclampsia/eclampsia		21	-5.9244	-5.548
223 Hysterectomy		26	-5.6133	-5.1656
Elderly				
226 Other elderly disorder		43	-0.8461	-0.6574
Miscellaneous				
229 Other miscellaneous		393	-1.3054	-1.1199

APACHE, Acute Physiology and Chronic Health Evaluation;*N, Non-operative, O, Operative GI, Gastrointestinal.

In the development cohort, the New APACHE II (A) showed poorer calibration than the original APACHE II model using the Hosmer-Lemeshow GOF test. The New APACHE II (A) had a significant GOF test ($P< 0.001$). The New APACHE II (B) had improved calibration, with the Hosmer-Lemeshow GOF test not significant (Table 8.4). Both new models showed significantly higher areas under the ROC curve when compared to the original score (Table 8.4).

In the validation cohort, the New APACHE II (A) model had poorer calibration than both the original model and the New APACHE II (B) model and all had significant Hosmer-Lemeshow tests but the New APACHE II (A) model had a much larger Chi Squared value ($x^2=136.52$). This difference between the New APACHE (A) and the other APACHE II models was reflected in the calibration curves (Figure 8.1). Both new models showed significantly increased areas under the ROC curve when compared to the original model.

Table 8.4 GOF tests and Area under the ROC curves for APACHE II

		GOF	P<	ROC	P<	Mortality Ratios
APACHE II	Development	16.53	0.035	0.820	—	0.97 (0.94-1.00)
	Validation	34.12	0.001	0.807	—	0.94 (0.91-0.97)
NEW APACHE II (A)	Development	45.48	0.001	0.862	0.001*	0.92 (0.89-0.94)
	Validation	136.52	0.001	0.838	0.001*	0.88 (0.85-0.91)
NEW APACHE II (B)	Development	12.99	0.112	0.858	0.001*	1.00 (0.97-1.03)
	Validation	44.86	0.001	0.835	0.001*	0.97 (0.94-1.00)

GOF, Goodness of Fit; ROC, Reciever Operating Characteristic; APACHE, Acute Physiology and Chronic Health Evaluation; * P-value of area under the ROC curve when compared to the original model.

Table 8.5 Source of admission mortality ratios and P values for APACHE II

Source of Admission	n=	Original APACHE II	X^2 P=	New APACHE II (A)	X^2 P=	New APACHE II (B)	X^2 P=
A&E	810	1.06* (0.98-1.13)	0.345	0.84 (0.78-0.90)	0.003	0.93 (0.86-1.00)	0.224
Recovery/theatre	3,135	0.79* (0.73-0.85)	0.001	0.92 (0.86-0.98)	0.051	1.00 (0.94-1.07)	1.000
Ward in this hospital	1,877	1.05* (1.00-1.09)	0.167	0.91 (0.87-0.95)	0.007	0.99 (0.95-1.03)	0.740
Other ICU in this hospital	117	1.00 (0.80-1.20)	1.000	0.83 (0.67-0.98)	0.203	0.92 (0.75-1.09)	0.572
ICU in another hospital	207	1.04 (0.88-1.21)	0.708	0.81 (0.69-0.94)	0.077	0.92 (0.78-1.06)	0.458
Other area in another hospital	363	0.80 (0.66-0.93)	0.034	0.66* (0.55-0.78)	0.001	0.75* (0.62-0.87)	0.007
Home/clinic	5	1.70 (0.34-3.05)	0.450	1.12 (0.19-2.04)	0.888	1.15 (0.30-2.00)	0.841
Overall patients	6,514	0.94 (0.91-0.97)	0.009	0.88 (0.85-0.91)	0.001	0.97 (0.94-0.99)	0.121

APACHE, Acute Physiology and Chronic Health Evaluation; A&E, Accident and Emergency; ICU, Intensive Care Unit; *Significance as indicated by 95% CIs.

Variations in the model's ability to predict in large sub groups were seen in the mortality ratios for both source of admission and the APACHE System category in the validation cohort (Table 8.5 and 8.6). The CIs for the mortality ratios for patients admitted from A&E, Recovery/theatre, and Ward in this hospital all lay outside the CIs for the overall population in the original APACHE II model.

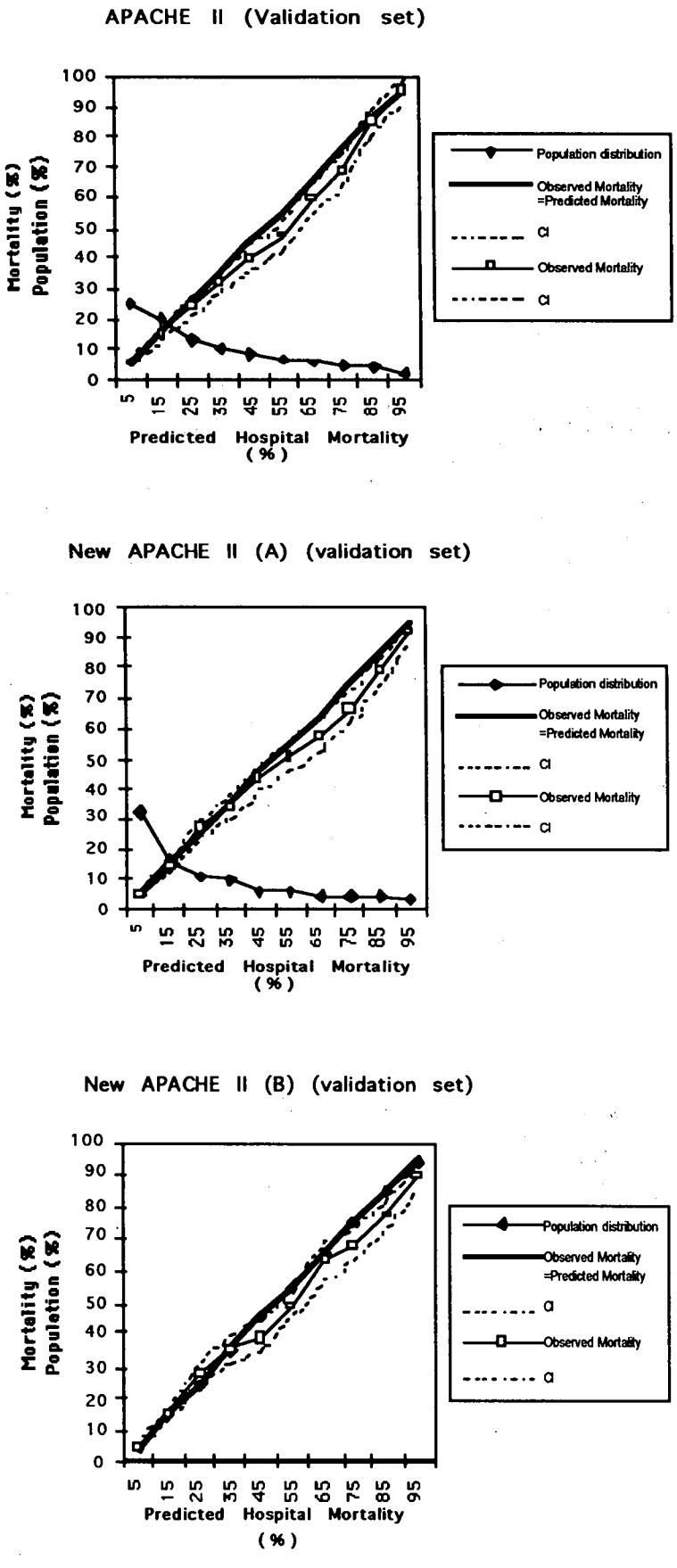
Table 8.6 APACHE System mortality ratios for APACHE II

APACHE System	N	Original APACHE II	χ^2 P=	New APACHE II (A)	χ^2 P=	New APACHE II (B)	χ^2 P=
Cardiovascular	1,406	1.07* (1.02-1.12)	0.086	0.91 (0.87-0.95)	0.016	0.97 (0.92-1.01)	0.406
Respiratory	1,241	1.01 (0.94-1.08)	0.791	0.87 (0.81-0.93)	0.005	0.95 (0.88- 1.02)	0.313
Neurological	573	1.27* (1.15-1.40)	0.002	0.75* (0.67-0.83)	0.001	0.95 (0.85-1.05)	0.493
Gastrointestinal	2,149	0.77* (0.72-0.83)	0.001	0.92 (0.86-0.97)	0.042	0.97 (0.92-1.03)	0.532
Renal	200	0.75 (0.54-0.97)	0.090	0.74 (0.54-0.94)	0.070	0.77 (0.56-0.98)	0.118
Metabolic/ Endocrine	73	1.02 (0.70-1.34)	0.920	0.75 (0.54-0.96)	0.198	0.82 (0.59-1.05)	0.365
Haematological	32	1.12 (0.77-1.47)	0.655	1.01 (0.72-1.31)	1.000	1.08 (0.76-1.40)	0.764
Trauma	461	1.13 (0.89-1.37)	0.351	0.80 (0.61-0.99)	0.092	1.03 (0.81-1.26)	0.806
General	379	0.58* (0.37-0.78)	0.001	1.20 (0.88-1.52)	0.275	1.34* (1.00-1.68)	0.080
Overall	6,514	0.94 (0.91-0.97)	0.009	0.88 (0.85-0.91)	0.000	0.97 (0.94-0.99)	0.121

APACHE, Acute Physiology and Chronic Health Evaluation;
*Significance as indicated by 95 % CIs

These were the areas admitting the bulk of patients to ICUs. The CIs for the mortality ratios for patients admitted from an "Other area in another hospital" lay outside the CIs for the overall population for both the New APACHE II models (New APACHE II (A & B)). However, the numbers in this group of patients were small. In the system categories, of cardiovascular, neurological, gastrointestinal and general, the CIs for the mortality ratios lay outside the CIs for the overall population in the original APACHE II model.

Figure 8.1. Calibration curve for APACHE II models in the validation cohort.



In the New APACHE II(A) model the neurological system category was the only category with mortality ratio CIs falling outside the overall CIs. In the New APACHE II(B) model the general system category was the only category with mortality ratio CIs falling outside the overall CIs, representing a small amount of patients.

The original APACHE II model showed significant differences in the Chi Squared value between the expected mortality and that observed for patients admitted from Recovery/ theatre, Other area in another hospital, and the patients with a neurological, gastrointestinal and general diagnosis (Table 8.5-8.6). The New APACHE II (A) showed significance using the Chi Squared test in patients admitted from A&E, Ward in this hospital, or Other area in another hospital and in patients with a cardiovascular, respiratory, neurological or gastrointestinal diagnosis (Table 8.5-8.5). The New APACHE (B) showed no significance in patients grouped by diagnosis only for patients admitted from Another area in another hospital, which had a relatively small amount of patients (Table 8.5).

8.3.3 SAPS II Models

Analysis for the SAPS II models used 12,944 patients, 6,426 in the development cohort and 6,518 in the validation cohort, with 64 patients excluded because of missing information.

The Chi Squared values for the Hosmer-Lemeshow GOF test in both cohorts were highly significant for the original SAPS II model but were not significant for the New SAPS II (A) model (Table 8.7). The New SAPS model showed improved calibration using the Hosmer-Lemeshow GOF test with the changes reflected in the calibration curves (Figure. 8.2). Discrimination was improved for the New SAPS II (A) model with slightly increased areas under the ROC curves for both the development and validation cohorts. Despite the slight change in the area under the ROC curves these changes were significant.

The coefficients for the new SAPS II (A) model were as follows:

New SAPS II constant (β_0)	-8.3937
SAPS II score (β_1)	+0.0518
Log of SAPS II score +1 (β_2)	+1.4439
Emergency Operative (β_3)	-0.4544

Table 8.7 GOF tests and Area under the ROC curves for SAPS II

		GOF	P	ROC	P	Mortality Ratios	CIs
SAPS II	Development	40.14	0.001	0.843		0.96	0.94-1.00
	Validation	56.04	0.001	0.837		0.95	0.92-0.98
New SAPS II (A)	Development	10.09	0.258	0.846	0.001*	1.00	0.97-1.03
	Validation	7.55	0.479	0.839	0.001*	0.98	0.95-1.01

GOF, Goodness of Fit, ROC, Receiver Operating Characteristic; CI, Confidence Interval; SAPS, Simplified Acute Physiology Score; * P-value of area under the ROC curve when compared to the original model.

The uniformity of fit did not seem to have been significantly improved with the New SAPS model (Table 8.8 and 8.9). The CIs for the mortality ratios, for patients grouped by source of admission, showed the same significance in the original model and the New SAPS II (A) model with admissions from A&E, Ward in this Hospital, Other area in another Hospital lying outside the overall CIs. Although not significant the original SAPS II mortality ratio for the group "Recovery/theatre" was low with a CI that does not include one. The New SAPS II (A) model had a mortality ratio of 1.00 in this group which suggests there had been an improvement. As this was the largest group of patients this may be important to the accuracy of the model.

When grouped by the APACHE diagnostic system, patients' CIs in the neurological, renal and trauma systems did not overlap with the CIs for all patients in both models with the general category falling outside the CIs for the New SAPS model.

Using the Chi Squared test patients admitted from A&E, Recovery/theatre, Ward in this hospital or Other area in another hospital, and patients admitted with a neurological, renal and trauma diagnosis showed significance for the original SAPS II model (Table 8.8 and 8.9). However, the renal category was small and represents only 200 patients. Using the same test for the New SAPS II (A) model, patients admitted from A&E, Ward in this hospital, or Other area in another hospital, and patients admitted with a neurological, renal, trauma and general diagnosis showed significant differences between observed and estimated mortality.

Table 8.8 Source of admission mortality ratios for SAPS II

Source of Admission	N	Original SAPS II	χ^2 P=	New SAPS II (A)	χ^2 P=
A&E	810	0.83* (0.76-0.89)	0.002	0.82* (0.76-0.89)	0.001
Recovery/theatre	3,135	0.88 (0.82-0.94)	0.002	1.00 (0.93-1.06)	1.000
Ward in this hospital	1,880	1.09* (1.04-1.13)	0.017	1.08* (1.03-1.12)	0.036
Other ICU in this hospital	117	0.98 (0.80-1.16)	0.888	0.96 (0.78-1.15)	0.806
ICU in another hospital	207	1.01 (0.86-1.16)	1.000	0.98 (0.83-1.13)	0.862
Other area in another hospital	364	0.78* (0.65-0.91)	0.020	0.75* (0.62-0.88)	0.008
Home/clinic	5	1.93 (0.79-3.06)	0.345	1.88 (0.66-3.10)	0.365
Overall patients	6,518	0.95 (0.92-0.98)	0.035	0.98 (0.95-1.01)	0.488

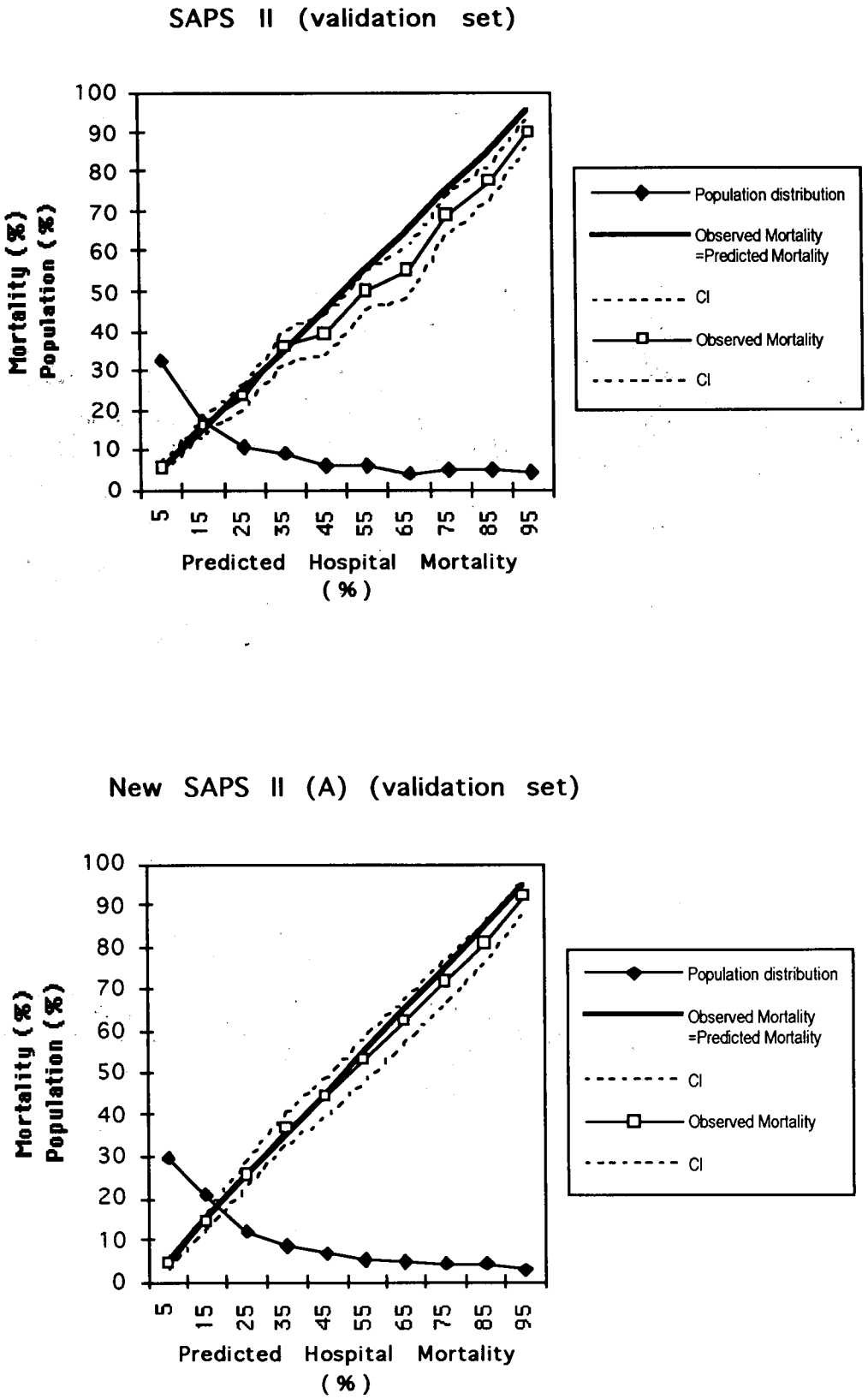
SAPS, Simplified Acute Physiology Score; A&E, Accident and Emergency; ICU, Intensive Care Unit; *Significance as indicated by 95% CIs

Table 8.9 APACHE System mortality ratios for SAPS II

APACHE System	n	Original SAPS II	χ^2 P=	New SAPS II (A)	χ^2 P=
Cardiovascular	1,406	1.00 (0.95-1.04)	0.920	1.03 (0.98-1.08)	0.431
Respiratory	1,241	1.05 (0.98-1.12)	0.359	1.02 (0.95-1.09)	0.718
Neurological	573	0.80* (0.71-0.89)	0.005	0.80* (0.71-0.90)	0.006
Gastrointestinal	2,149	0.98 (0.92-1.03)	0.603	1.07 (1.01-1.13)	0.093
Renal	200	0.68* (0.50-0.85)	0.018	0.68* (0.49-0.86)	0.020
Metabolic/Endocrine	73	0.79 (0.57-1.02)	0.303	0.79 (0.56-1.03)	0.299
Haematological	32	1.20 (0.87-1.54)	0.471	1.19 (0.85-1.54)	0.488
Trauma	461	0.69* (0.53-0.85)	0.006	0.71* (0.54-0.88)	0.010
General	379	0.73 (0.50-0.95)	0.059	0.70* (0.48-0.93)	0.035
Overall patients	6,518	0.95 (0.92-0.98)	0.035	0.98 (0.95-1.01)	0.488

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; *Significance as indicated by 95% CIs

Figure 8.2. Calibration curves for SAPS II models in the validation cohort.



8.3.4 MPM_0 Models

The MPM_0 model used all 13,008 patients for the analysis, with 6,458 patients in the development cohort and 6,550 patients in the validation cohort.

The coefficients for the new MPM_0 models are as follows (Table 8.10):

Table 8.10 Coefficients for the New MPM_0 (A) model.

Variable	ß
Constant	-4.9072
Physiology	
Coma or deep stupor	0.5531
Heart Rate >= 150	*
Systolic blood pressure	*
Chronic diagnosis	
Chronic renal insufficiency	0.5539
Cirrhosis	0.7652
Metastatic neoplasm	0.466
Acute diagnosis	
Acute renal failure	1.2697
Cardiac dysrhythmia	0.2983
Cerbrovascular incident	0.9335
Gastrointestinal bleeding	0.5208
Intracranial mass effect	0.4515
Other	
Age	0.034
Cardiopulmonary resuscitation prior to admission	1.2781
Mechanical ventilation	0.3341
Non-elective surgery	1.2483

MPM, Mortality Probability Model; SE, Standard Error; CI, Confidence Intervals;
*Variable not significant and not included model.

New MPM_0(B)	
Constant (β_0)	-0.3938
MPMLogit	0.6387

Table 8.11 GOF tests and Area under the ROC curves for MPM_0

		GOF	P _	ROC	P	Mortality Ratios	CIs
MPM_0	Development	209.56	0.001	0.787		1.00	0.97-1.02
	Validation	229.46	0.001	0.779		0.98	0.96-1.01
NEW MPM_0 (A)	Development	17.74	0.023	0.809	0.001*	1.00	0.96-1.03
	Validation	18.19	0.020	0.787	0.001*	0.98	0.94-1.01
New MPM_0(B)	Development	20.59	0.008	0.786	0.001*	1.00	0.97-1.03
	Validation	17.70	0.024	0.780	0.001*	0.98	0.96-1.01

GOF, Goodness of Fit; ROC, Reciever Operating Characteristics; MPM, Mortality Probability Model; CI, Confidence Interval

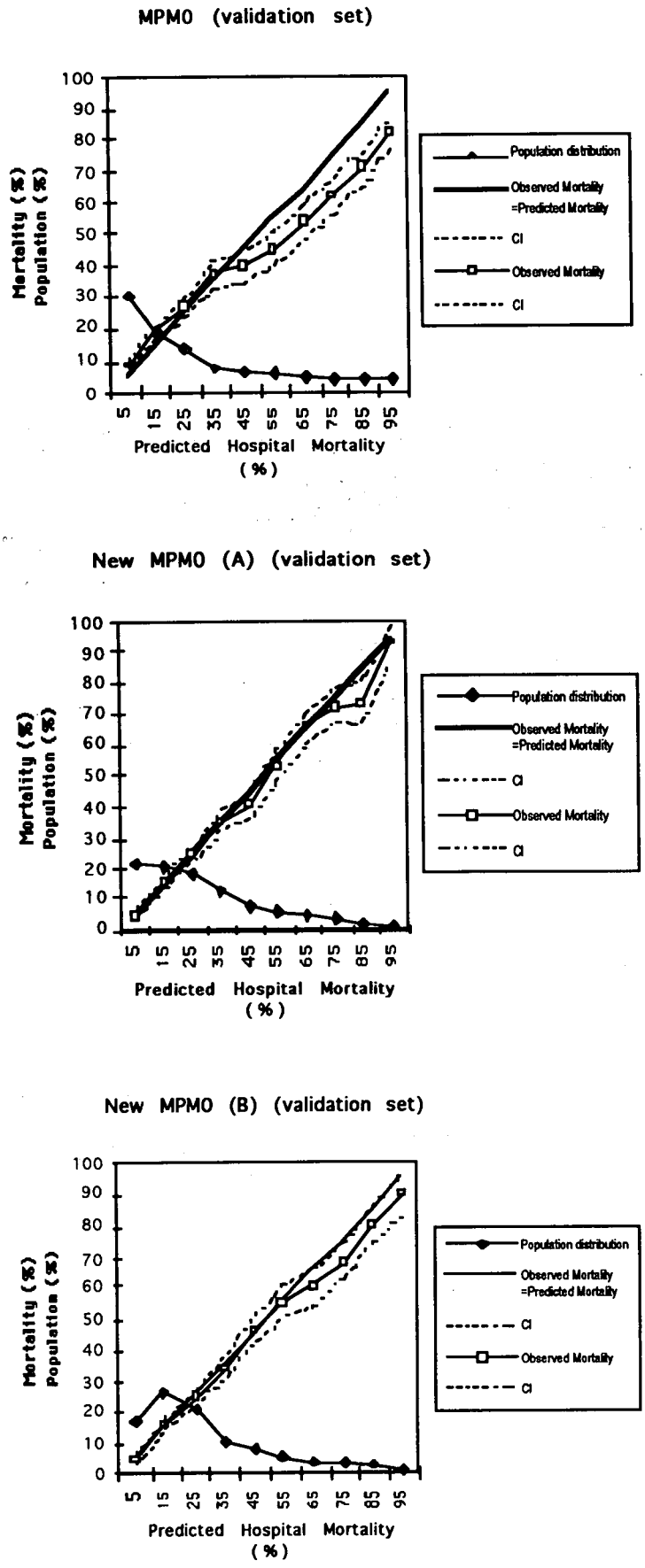
* P-value of area under the ROC curve when compared to the original model.

Table 8.12 Source of admission mortality ratios for MPM_0

Source of Admission	N=	Original MPM_0	X^2 P=	New MPM_0 (A)	X^2 P=	New MPM_0 (B)	X^2 P=
A&E	815	0.81* (0.77-0.85)	0.001	0.94 (0.87-1.02)	0.332	0.90* (0.84-0.95)	0.07
Recovery/theatre	3,147	0.89* (0.84-0.94)	0.004	0.84* (0.78-0.90)	0.001	0.85* (0.80-0.90)	0.001
Ward in this hospital	1,891	1.17* (1.14-1.19)	0.001	1.15* (1.10-1.20)	0.001	1.17* (1.13-1.22)	0.001
Other ICU in this hospital	118	1.08 (0.87-1.28)	0.610	0.98 (0.78-1.18)	0.888	1.04 (0.86-1.22)	0.78
ICU in another hospital	207	0.97 (0.88-1.06)	0.806	0.93 (0.77-1.08)	0.517	1.01 (0.86-1.15)	1.00
Other area in another hospital	365	0.83* (0.70-0.95)	0.076	0.77* (0.63-0.91)	0.018	0.78* (0.66-0.91)	0.02
Home/clinic	7	2.12 (0.74-3.51)	0.182	1.85 (0.71-3.00)	0.277	2.21* (1.22-3.20)	0.16
Overall patients	6,550	0.98 (0.96-1.01)	0.454	0.98 (0.94-1.01)	0.301	0.98 (0.96-1.01)	0.46

MPM, Mortality Probability Model; A&E, Accident and Emergency; ICU, Intensive Care Unit
*Significance as indicated by 95% CIs

Figure 8.3. Calibration curves for the MPM_0 models in the validation cohort.



The calibration of the New MPM_0 models (A & B) improved with the Chi Squared value for the GOF test in both the development and validation cohorts reduced compared to that of the original model (Table 8.11). Although the Chi Squared value was reduced for both new models the P-values remain significant for both. The calibration for all the MPM_0 models can be seen in Figure 8.3. Discrimination was increased in the New MPM_0 (A) model with a significant increase in the ROC values in both the developmental cohort and the validation cohort (Table 8.11). Discrimination was also improved in the validation cohort for the New MPM_0 (B) model, but, although there were significant differences, the differences were small.

Table 8.13 APACHE System mortality ratios for MPM_0

APACHE System	N=	Original MPM_0	X^2 P=	New MPM_0 (A)	X^2 P=	New MPM0 (B)	X^2 P=
Cardiovascular	1,408	1.04 (0.99-1.09)	0.348	1.05 (0.99-1.10)	0.258	1.07* (1.03-1.12)	0.08
Respiratory	1,247	1.12* (1.05-1.19)	0.022	1.10* (1.02-1.18)	0.056	1.08 (1.01-1.15)	0.13
Neurological	577	0.71* (0.63-0.80)	0.000	0.82* (0.72-0.92)	0.012	0.84* (0.76-0.92)	0.03
Gastrointestinal	2,157	1.03 (0.97-1.09)	0.532	0.96 (0.90-1.02)	0.365	1.00 (0.95-1.06)	0.92
Renal	202	0.77 (0.57-0.97)	0.106	0.70* (0.50-0.90)	0.030	0.70* (0.53-0.87)	0.03
Metabolic/ Endocrine	73	0.82 (0.59-1.06)	0.380	0.98 (0.68-1.27)	0.920	0.83 (0.63-1.03)	0.41
Haematological	32	1.47* (1.07-1.88)	0.132	1.49* (1.02-1.95)	0.121	1.49* (1.12-1.86)	0.12
Trauma	462	0.66* (0.49-0.82)	0.002	0.67* (0.50-0.85)	0.003	0.60* (0.47-0.73)	0.001
General	381	0.71* (0.49-0.94)	0.046	0.68* (0.45-0.92)	0.024	0.58* (0.43-0.73)	0.001
Overall patients	6,550	0.98 (0.95-1.01)	0.454	0.98 (0.94-1.01)	0.301	0.98 (0.96-1.01)	0.46

APACHE, Acute Physiology and Chronic Health Evaluation; MPM, Mortality Probability Model;
*Significance as indicated by 95% CIs

Despite the marked improvement in the calibration and the small but significant discrimination for the New MPM_0 models the uniformity of fit was very similar in all models for both source of admission and system category (Table 8.12 and 8.13). In the validation cohort, CIs for the groups of patients admitted from Recovery/theatre, Ward in this hospital, and Other area in another hospital all lay outside the overall CIs in all models. In the original model and the New MPM_0 (B) model the CIs for the group of patients admitted from A&E also lay outside the overall CIs. The same pattern was seen when the patients were broken down by the APACHE system category. Patients with an admitting diagnosis that fell in the respiratory, neurological, trauma, general or haematological system categories all had CIs that fall outside the overall CIs for the original MPM_0 model and the New MPM_0 (A) model. In the New MPM_0(A) model the CIs for the renal system category fall outside the overall CIs. Patients admitted with neurological, trauma, renal, general or haematological diagnoses all had CIs that fall outside the overall CIs for the New MPM_0 (B) model.

The Chi Squared test for the original MPM_0 model showed significant differences between estimated and observed mortality for patients admitted from A&E, Recovery/theatre, or Ward in this hospital and for patients admitted with a cardiovascular, neurological, gastrointestinal or general diagnosis. There were significant differences for patients admitted from Recovery/theatre, Ward in this hospital, Other area in another hospital and patients with a neurological, renal trauma and general diagnoses in both New MPM_0 models using the Chi Squared test.

8.3.5 MPM_{24} Models

As the MPM_{24} models exclude patients discharged in the first 24 hours, this left 9,212 patients for the analysis, with 4,549 patients in the development cohort and 4663 patients in the validation cohort.

New coefficients for the MPM_{24} models are as follows (Table 8.14):

Table 8.14 Coefficients for new MPM_{24}(A)

Variable	ß
Constant	-5.0243
Variables ascertained at admission	
Chronic renal insufficiency	0.4415
Cirrhosis	0.6915
Metastatic neoplasm	0.7541
Cerabralvascular incident	0.7114
Gastrointestinal bleeding	0.5204
Intracranial mass effect	*
Age	0.0315
Cardiopulmonary resuscitation prior to admission	1.0152
Non-elective surgery	0.7913
24 hour assessments	
Coma or deep stupor at 24 hour	1.4195
Creatinine >176.8mmol/L	0.325
Confirmed infection	0.3021
Mechanical ventilation	0.5525
Partial pressure of oxygen(PO_2) <7.98 Kilopascal (60 mm Hg)	0.5555
Prothrombin time > 3 sec above standard.	-0.0012
Urine output <150 ml in 8 h	1.0694
Vasoactve drugs≥1 h intravenously	0.6331

MPM, Mortality Probability Model; SE, Standard Error; CI, Confidence Intervals;
*Variable not significant and not included model.

New MPM_{24}(B)	
Constant ($ß_0$)	0.2599
MPMLogit	0.8169

Table 8.15 GOF tests and Area under the ROC curves for MPM_{24}

		GOF	P <	ROC	P<	Mortality Ratios	CIs
MPM_{24}	Development	53.22	0.001	0.799		1.04	1.00-1.07
	Validation	106.07	0.001	0.782		0.99	0.95-1.02
NEW MPM_{24} (A)	Development	169.34	0.001	0.816	0.009*	1.28	1.24-1.32
	Validation	134.54	0.001	0.795	0.001*	1.22	1.18-1.27
NEW MPM_{24} (B)	Development	15.11	0.057	0.807	0.003*	1.02	0.98-1.06
	Validation	30.22	0.001	0.793	0.001*	0.98	0.94-1.02

GOF, Goodness of Fit; ROC, Receiver Operating Characteristic; MPM, Mortality Probability Model; CI, Confidence Interval; * P-value of area under the ROC curve when compared to the original model.

The New MPM_{24} (A) model had significantly improved discrimination with increases in the area under the ROC curve for both the development and validation cohort (Table 8.15). However, the calibration for New MPM_{24} (A) had deteriorated and although both the original and new models had significant GOF tests the Chi Squared value was considerably increased for the new model. This calibration was reflected in the calibration curves and mortality ratios where New MPM_{24} (A) was clearly underestimating mortality (Figure 8.4).

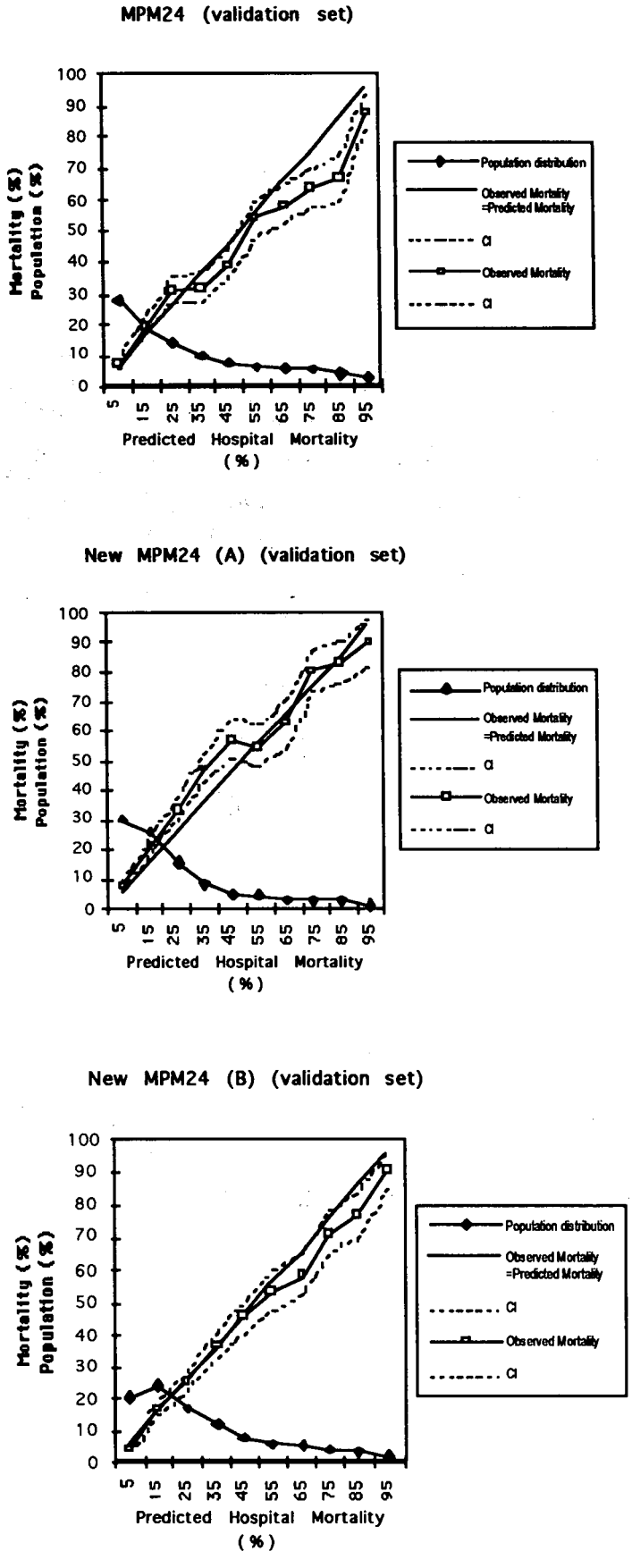
New MPM_{24} (B) (logit only model) had significantly improved areas under the ROC curve for both the development and validation cohort (Table 8.15). In contrast to the New MPM_{24} (A) (development from original MPM variables), calibration was improved when compared to the original model with reduced Chi Squared values in the GOF tests for both the developmental and validation cohort (Table 8.15).

Table 8.16 Source of admission mortality ratios for MPM_{24}

Source of Admission	N	Original MPM_{24}	χ^2 P=	New MPM_{24} (A)	χ^2 P=	New MPM_{24} (B)	χ^2 P=
A&E	510	1.08 (0.98-1.19)	0.292	1.19 (1.08-1.30)	0.020	1.01 (0.91-1.11)	0.888
Recovery/theatre	2,132	0.86* (0.80-0.93)	0.002	1.20 (1.11-1.28)	0.000	0.89 (0.82-0.96)	0.015
Ward in this hospital	1,448	1.11* (1.06-1.17)	0.008	1.29 (1.22-1.35)	0.000	1.08 (1.02-1.14)	0.063
Other ICU in this hospital	95	1.06 (0.82-1.30)	0.740	1.23 (0.95-1.50)	0.242	1.02 (0.78-1.26)	0.888
ICU in another hospital	180	0.91 (0.76-1.07)	0.475	1.14 (0.95-1.34)	0.292	0.92 (0.76-1.08)	0.502
Other area in another hospital	295	0.81 (0.67-0.96)	0.083	1.03 (0.85-1.21)	0.806	0.79 (0.64-0.94)	0.049
Home/clinic	3	2.26 (0.68-3.84)	0.237	2.11 (0.75-3.47)	0.279	2.09 (0.68-3.49)	0.288
Overall patients	4,663	0.99 (0.95-1.02)	0.671	1.22 (1.18-1.27)	0.000	0.98 (0.94-1.02)	0.410

MPM, Mortality Probibility Mortality; A&E, Accident and Emergency; ICU, Intensive Care Unit;*Significance as indicated by 95% CIs

Figure 8.4. Calibration curves for the MPM_{24} models in the validation cohort.



However, the Chi Squared value was still significant in the validation cohort. These changes were reflected in the calibration curves, with more of the line of equality lying within the CIs for observed mortality (Figure 8.4).

The New MPM_{24} models (A & B) had less variation in the mortality ratios for source of admission than the original model, with none of the CIs for patients admitted from the various areas lying outside their respective overall CIs (Table 8.16). In the original model CIs for patients admitted from Recovery/theatre and Ward in this hospital, the areas admitting the majority of patients, lay outside the CIs for the overall cohort.

Table 8.17 APACHE System mortality ratios for MPM_{24}

	n=	MPM_{24}	X^2 P=	New MPM_{24} (A)	X^2 P=	NEW MPM_{24} (B)	X^2 P=
Cardiovascular	971	1.05 (1.01-1.10)	0.145	1.18 (1.10-1.25)	0.001	1.00 (0.94-1.07)	1.000
Respiratory	993	1.20* (1.14-1.26)	0.001	1.42* (1.32-1.52)	0.001	1.16* (1.08-1.25)	0.006
Neurological	370	0.90 (0.82-0.99)	0.137	1.10 (0.96-1.25)	0.303	0.86 (0.74-0.98)	0.108
Gastrointestinal	1,598	0.95 (0.91-1.00)	0.160	1.24 (1.15-1.32)	0.001	0.95 (0.88-1.02)	0.262
Renal	138	0.75* (0.59-0.91)	0.036	0.82* (0.56-1.09)	0.322	0.65* (0.42-0.87)	0.025
Metabolic/ Endocrine	49	0.95 (0.73-1.17)	0.777	1.01 (0.64-1.38)	1.000	0.84 (0.51-1.18)	0.554
Haematological	21	1.35 (1.04-1.67)	0.154	2.04* (1.48-2.60)	0.009	1.57* (1.12-2.02)	0.101
Trauma	317	0.73* (0.58-0.87)	0.004	0.97 (0.71-1.23)	0.823	0.68* (0.47-0.89)	0.015
General	203	0.80 (0.60-1.00)	0.133	1.09 (0.74-1.44)	0.663	0.81 (0.51-1.10)	0.277
Overall patients	4,663	1.01 (0.99-1.04)	0.538	1.22 (1.18-1.27)	0.001	0.98 (0.94-1.02)	0.410

APACHE, Acute Physiology and Chronic Health Evaluation; MPM, Mortality Probability Mortality;
*Significance as indicated by 95% CIs

The Chi Squared value was, however, significant for the New MPM 24 (A) model for patients admitted from A&E, Recovery/theatre, Ward in this hospital (the three largest groups) (Table 8.16). The New MPM_{24} (B) model had a significant Chi Squared value

for patients admitted from Recovery/theatre and Other area in another hospital. The original model had significant Chi Squared values in patients admitted from Recovery/theatre and Ward in this hospital (the two largest groups).

Using the APACHE system category to group patients the New MPM_{24} model did not appear to improve the uniformity of fit (Table 8.17). The CIs for the categories of respiratory, and renal, lay outside their respective overall CIs for all MPM_{24} models. The CIs for the trauma category lay outside the overall CIs for the original MPM_{24} and the new MPM_{24} (B) model. The CIs for the haematological group also lay outside the overall CIs for both new MPM_{24} models. The numbers in this group were very small and probably had little influence on the performance of the model.

The New MPM_{24} (A) had significant Chi Squared values in patients with a cardiovascular, respiratory, gastrointestinal or haematological diagnosis (Table 8.17). The New MPM 24 (B) and original MPM_{24} has significant Chi Squared values in patients admitted with a respiratory, renal or trauma diagnosis, although the renal group represented a small number of patients.

All the models had demonstrated some improvement in at least one of the areas measured. All the models, apart from the New MPM_0 (B), had shown improved discrimination. As reported in Chapter 4-6 the APACHE II model had the superior calibration of all the models previously tested in the Scottish data . The New APACHE II (B) model had maintained this calibration (Table 8.18) but more importantly the model had improved its uniformity of fit when patients were grouped by APACHE system and source of admission (Table 8.19, 8.20). The New SAPS II (A) model was the only model that had improved its calibration to the extent that the Hosmer-Lemeshow test was no longer significant in the validation cohort. The New MPM_0 models had improved calibration as measured by the Hosmer-Lemeshow test. The New MPM_{24} (B) demonstrated improved calibration using the Chi Squared statistic and both new models showed improvement in uniformity of fit when grouped by source of admission.

Table 8.18 Results for new models in validation cohort

	GOF	P=	ROC	P=	Mortality Ratios	CIs
New APACHE II (A)	136.52	0.001	0.838	0.001*	0.88	0.94-1.00
New APACHE II (B)	44.86	0.001	0.835	0.001*	0.97	0.94-1.00
New SAPS	7.55	0.479	0.839	0.001*	0.98	0.95-1.01
New MPM_0 (A)	18.19	0.020	0.787	0.001*	0.98	0.94-1.01
New MPM_0 (B)	17.14	0.024	0.7801		0.98	0.96-1.01
New MPM_{24} (A)	134.54	0.001	0.795	0.001*	1.22	1.18-1.27
New MPM_{24} (B)	30.22	0.001	0.793	0.001*	0.98	0.94-1.02

GOF, Goodness of Fit; ROC, Receiver Operating Characteristic; CI, Confidence Interval; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Model; * Significance when compared to the original model.

Table 8.19 Uniformity of fit by source for new models in validation cohort

	A&E	CIs	Recovery theatre	CIs	Ward	CIs	Overall	CIs
New APACHE II (A)	0.84	0.78-0.90	0.92	0.86-0.98	0.91	0.87-0.95	0.88	0.85-0.91
New APACHE II (B)	0.93	0.86-1.00	1.00	0.94-1.07	0.99	0.95-1.03	0.97	0.94-0.99
New SAPS	0.82*	0.76-0.89	1.00	0.93-1.06	1.08*	1.03-1.12	0.98	0.95-1.01
New MPM_0 (A)	0.94	0.87-1.02	0.84*	0.78-0.90	1.15*	1.10-1.20	0.98	0.94-1.01
New MPM_0 (B)	0.90	0.84-0.95	0.85	0.80-0.90	1.17*	1.13-1.22	0.98	0.96-1.01
New MPM_{24} (A)	1.19	1.08-1.30	1.20	1.11-1.28	1.29	1.22-1.35	1.22	1.18-1.27
New MPM_{24} (B)	1.01	0.91-1.11	0.89	0.82-0.96	1.08	1.02-1.14	0.98	0.94-1.02

A&E, Accident and Emergency; CI, Confidence Intervals; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probibility Model;
* Mortality ratios whose CIs fall outside the CIs for the whole population.

Table 8.20 Uniformity of fit by APACHE system for new models in validation cohort

	Cardio–vascular	CIs	Respiratory	CIs	Gastro-intestinal	CIs	Overall	CIs
New APACHE II (A)	0.91	0.87-0.95	0.87	0.81-0.93	0.92	0.86-0.97	0.88	0.85-0.91
New APACHE II (B)	0.97	0.92-1.01	0.95	0.88- 1.02	0.97	0.92-1.03	0.97	0.94-0.99
New SAPS	1.03	0.98-1.08	1.02	0.95-1.09	1.07	1.01-1.13	0.98	0.95-1.01
New MPM_0 (A)	1.05	0.99-1.10	1.10*	1.02-1.18	0.96	0.90-1.02	0.98	0.94-1.01
New MPM_0 (B)	1.07*	1.03-1.12	1.08	1.01-1.15	1.00	0.95-1.06	0.98	0.96-1.01
New MPM_{24} (A)	1.18	1.10-1.25	1.42*	1.32-1.52	1.24	1.15-1.32	1.22	1.18-1.27
New MPM_{24} (B)	1.00	0.94-1.07	1.16*	1.08-1.25	0.95	0.88-1.02	0.98	0.94-1.02

APACHE, Acute Physiology and Chronic Health Evaluation; CI, Confidence Intervals; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probibility Model;
* Mortality ratios whose CIs fall outside the CIs for the whole population.

8.3.6 Selection check

Analysis for the extra random selection used 12,936 patients, 6461 patients in the developmental cohort and 6475 patients in the validation cohort. Discrimination in the APACHE II (check) model is little different compared to the new APACHE II (B) in the both the development and validation cohort (Table 8.21). However, the new APACHE II (check) model has better discrimination than the original model in both cohorts.

The calibration in the new APACHE (Check) in both the development and validation cohorts was improved in comparison to the original model (Table 8.21). However the differences in the Chi Squared statistic was small in comparison to other differences found in this study. In comparison to the new APACHE II (B) model there was little difference in calibration. The new APACHE II (check) model had a non-significant GOF test in the development cohort (P=0.088) but the same test was significant in the validation set. Although the new APACHE II (B) model was significant in the development set the actual Chi Squared values in both validation and development cohorts were very similar to those in the new APACHE II (check) model.

Table 8.21 GOF tests and Area under the ROC curves for APACHE II

		GOF	P<	ROC	P<*	Mortality Ratios (CI)
APACHE II	Development	23.88	0.002	0.8119		0.97 (0.94-1.00)
	Validation	27.52	0.001	0.8168		0.94 (0.91-0.97)
New APACHE II (check)	Development	13.74	0.088	0.8515	0.001*	0.99 (0.96-1.02)
	Validation	24.43	0.001	0.8516	0.001*	0.96 (0.94-1.00)
New APACHE II (B)	Development	19.63	0.012	0.8506	0.001*	0.99 (0.96-1.02)
	Validation	20.22	0.009	0.8476	0.001*	0.97 (0.94-1.00)

GOF, Goodness of Fit; ROC, Receiver Operating Characteristic; APACHE, Acute Physiology and Chronic Health Evaluation; CI, Confedence Intervals;

* P-value of area under the ROC curve when compared to the original model.

8.4 Discussion

For a logistic regression model to be a useful tool in assessing the differences in ICUs the probabilities produced must accurately reflect the mortality experience of the population to which the model is to be applied. Analysis from Chapter 5-7 shows that this was not the case in the Scottish Intensive Care study where the five models tested showed significant differences between observed and predicted mortality. All the models described in this paper have been developed on data collected from either an American or a mixture of American and European ICUs . The reported mortality rates in these studies have been considerably lower than those reported in the Scottish data. There is a general acceptance that this represents a difference in admission criteria rather than reflecting poorer outcomes and it is therefore not surprising that the performance of these models were not replicated in the Scottish data. Zhu et al suggested that where

> " the level of ICU care results in poor fit of the models, the models could be customised to better reflect the mortality experience among those patients" (142).

Methods for customisation of severity of illness models have recently been demonstrated (139,141,142). Zhu et al showed two methods for customising MPM_0 by using bootstrapping methods on data from the original MPM II study. The first method was to take the logit from the original MPM II model and enter it into a logistic regression analysis. The second method was to re-weight the variables from the original model. The study found that using the original variables was more successful when the difference in mortality from the original model was highest. A paper using data from the EURICUS study showed improved performance of MPM_0 after customisation of both the logit and the original variables. However, the authors did show that some large sub groups of the population had significant differences between the observed and predicted mortality suggesting that mortality ratios could be affected by the type of patients being admitted. Le Gall et al also showed improved performance of MPM II and SAPS II in patients with sepsis.

In our customisation of these models some improvement was produced in one or more of the tests either in discrimination or calibration. There is no agreed point at which a model becomes legitimate to use and the customisation process can be judged to have been successful. It is important for a model to have good discrimination, however, if a model is to be used to audit differences in different ICUs then there seems to be a general agreement that calibration is the more important measure (21). Wagner said,

> *"When you talk about performance evaluation you are fundamentally talking about calibration issues of the model. You are not as concerned about discrimination or ROC areas. Higher ROC areas are always better....."* (9th European Congress of Intensive Care medicine and personal communication).

As calibration is a measure of the differences between observed and predicted mortality measured across the "severity of illness" range, it is clear that significant differences in specific ranges may affect the apparent performance of an ICU, depending on the nature of the patients that they admit. More recently it has been argued that, as well as traditional goodness of fit tests, models need adequate "uniformity of fit" which can be defined as a model that has no significant differences between observed and predicted mortality in clinically relevant sub groups (115). If a model predicts poorly as for example neurological patients then this will effect the mortality ratios of those units admitting an unusually large number of these patients. However, it was difficult to assess the uniformity of fit of the original models as this information was not published. Therefore, any differences in the uniformity of fit in the new models may have existed in the original models.

8.4.1 APACHE II Models

Both New APACHE II models (A & B) showed improved discrimination when compared to the original model (Table 8.3). Calibration was not improved in the important validation cohort. The New APACHE II (B) model showed similar calibration to that of the original model, though still had a significant Hosmer-Lemeshow test in the validation cohort. It would appear that when accuracy in subgroups was analysed there was a real improvement in the New APACHE II (B) model (Table 8.4- 8.5). The New APACHE II (A) model also had improved performance in subgroups when the CIs from the mortality ratios were analysed. However, using the results from the Chi Squared value the New APACHE II (A) model performance was less promising (Table 8.4- 8.5). The New model (B) had only significant differences between the expected and observed mortality in patients admitted from Other area in another hospital and patients with a general diagnoses.

The improvement in the New APACHE II (B) model would seem to have created a genuine improvement in the APACHE II model. Athough the calibration results were very similar for both models (APACHE II; New APACHE (B)) analysis from Chapter 6 and 7 showed that the original APACHE II model's performance is due to the effect of averaging. Improving the performance in those patients with a neurological deficit reduces the performance of the model (Chapter 7). As the New APACHE II (B) model

had considerably improved performance in subgroups it can be argued that the overall calibration was genuine whilst that in the original model was misleading. Analysis in Chapter 7 appears to have been confirmed with the use of a pre-sedated GCS, when patients were sedated, having led to a superior model. This confirms the finding that the GCS is an important prognostic factor in the model. Also, it highlights the fact that the method used for collecting the data for a variable can have a profound effect on the accuracy of a model. The other main difference between the original APACHE II model and the New APACHE II (B) model is the use of the APACHE III diagnoses in the new model as opposed to the APACHE II diagnostic groups. This may have contributed to the improvement of the New APACHE II (B) model's performance when analysed in subgroups.

The New APACHE II (B) model still showed significant differences in the Hosmer-Lemeshow test although questions remain about the clinical significance of these results, given the large numbers of patients in the study.

8.4.2 SAPS II Models

The New SAPS II (A) model had significantly better discrimination, however, this improvement was very small as the large numbers in the study allowed for small differences to be highlighted (Table 8.6). The differences in the validation cohort would be imperceptible if rounded to two decimal places. As the logistic regression model used to customise the SAPS II model contained only the SAPS II score and a log of the SAPS II score, there was little room for improvement of the ROC curves. This is because the ROC test is based on ranking the probabilities from the model and while the probabilities may change considerably in the new model their rank will not. It is in the calibration that the new SAPS II (A) model shows such a dramatic improvement. It is the only model that has a Hosmer-Lemeshow test which did not show significant differences between expected and observed mortality (Table 8.6). What is even more surprising is that the improvement in calibration was greatest in the validation cohort rather than the development cohort.

However, results from analysis of the New SAPS II (A) model showed that the uniformity of fit had not been improved (Table 8.7-8.8), and a considerable number of patients admitted from A&E and from Ward in this hospital showed significant differences between expected and observed mortality. When patients were grouped by admitting diagnoses and then by system there were also significant differences in some of the groups but most patients fell into groups where no significance was detected. These apparent failures of the model may be as a result of a lack of diagnostic variable.

8.4.3 MPM II Models

8.4.3.1 MPM_0

There was a significant improvement in discrimination in the new models when compared to the original (Table 8.10). However, the areas under the ROC curves in the New MPM_0 (A) and the New MPM_0 (B) models for the validation cohort was only 0.7874 and 0.7801 which was considerably lower than other models. Calibration was considerably improved with the Hosmer-Lemeshow GOF test going from 229.46 in the original model to 18.19 in the New MPM_0 (A) model and 17.7 in the validation cohort. Nevertheless, the GOF statistics still remained significant but without the same certainty that was present in the original model. The new models have shown little improvement in uniformity of fit with significant differences between expected and observed mortality still present (Table 8.11-8.12). However, as the only model to use data on admission to the ICU, thereby avoiding treatment effect, and given its ease of collection it may still remain a useful model to use. There seems to be little to choose between the new models with the logit model (New MPM_0 (B)) performing slightly better with small improvements in calibration.

8.4.3.2 MPM_{24}

Like all the models that have been customised the discrimination had improved for both the New MPM_{24} models (Table 8.12). Although the area under the curve was only 0.79 in the validation cohort for both new models and this was low in comparison to the other models in this study. This was a result of excluding patients discharged in the first 24 hours as was shown in Chapter 5. Calibration had, surprisingly, deteriorated for the New MPM_{24} (A), model, while the New MPM_{24} model (logit only) had improved its calibration. However, the GOF statistic remained statistically significant for the New MPM_{24} (B) model in the validation cohort. When analysed in subgroups none of the MPM_{24} models had good performance and many of the large sub groups showed significant differences between expected and observed mortality (Table 8.13 -8.14).

It is surprising that the model which used the logit only (New MPM_{24} (B)) was the MPM_{24} model which appeared to have the best performance. This suggest that the weights for the variables in the model, once the logit has been re-weighted, was more accurate than the weights generated from the variable only model (New MPM_{24} (A)). This may be a result of the larger numbers in the original MPM II study, which allowed for the generation of more representative weights for the variables. At the moment, the accuracy in this model may be too poor to allow it to have any meaningful use.

8.4.4 Selection check

The new APACHE II (check) model has improved discrimination and calibration compared to the original model. However, the performance of the model when compared to the new APACHE II (B) model is similar with discrimination and calibration little changed. The performance of this new model in comparison to the new APACHE II (B) model would suggest that this different random split has produced a similar model to that in the original random split. This would appear to add weight to the legitimacy of the customisation process. As the new APACHE II (check) model produces such similar results to the new APACHE II (B) model in a different random sample, this would imply that the model is relatively robust and the process has produced a model that is stable in different groups of patients

8.4.5 Overall

The process of customisation has improved the discrimination of all the new models but the changes for most appear to be small. This is not surprising as for example re-weighting the SAPS II score will not change the ranks of the predictions within an ROC curve. Larger changes were seen in the new APACHE II models as the customisation process involved re-weighting the diagnostic groups as well as the score. Therefore, depending on the changes in weights from the original score there was potential for considerable change in the ranks of probability.

All the new models except the New APACHE II (A) and the New MPM_{24} (A) had gained improved calibration when measured by Chi Squared values. The poor calibration in the New APACHE II (A) model is hard to explain but the poor performance in neurological patients may indicate that the lack of GCS scores in many of the patients had a significant effect on the accuracy of the model. The calibration for the New MPM_{24} (A) had decreased, for which there would appear to be no obvious reason. However, the numbers in each of the variables may not be large enough to provide adequate weightings. Only the New SAPS II (A) had calibration which did not demonstrate significant differences between observed and predicted mortality. Two of the reported studies, where customisation was undertaken, have reported non significant results for the Hosmer-Lemeshow test in the New MPM_0 models (139,142). The results reported for the New MPM_0 are consistent with those demonstrated in data published from the EURICUS study, with improved calibration but with some questions remaining about the "uniformity of fit" of the model.

As the original studies did not report analysis of the "uniformity of fit" it was hard to compare this aspect of performance. The original APACHE II study did not report Hosmer-Lemeshow tests and it was therefore difficult to assess if the improved New APACHE II (B) model had equivalent performance. All the new models, apart from the New APACHE II (A) and the New MPM_{24} (A), had improved performance when compared to the original models. Consequently, if severity of illness models are to be used in the continuing Audit of Intensive Care in Scotland, involving the comparison of case mix adjusted outcome, then the coefficients generated in the new customised models should be used in preference to those used in the original models. If further improvements are needed then a more fundamental process of model building needs to be undertaken. Any future model building would need to look carefully at the process and quality of the data collection while ensuring the assumptions of a multiple logistic regression model were not violated. If the new models are to be used in the comparison of case mix adjusted outcome, even with the improvements that have been made, they should still be used with caution when applied to the Scottish data. As questions remain on the ability of the models to satisfactorily adjust for case mix in the Scottish ICU population, it may be that more than one model should be applied to the data. As both the New SAPS II (A) and the New APACHE II (B) model demonstrated good areas under the ROC curves, and the New SAPS II (A) demonstrated good calibration and APACHE II good "uniformity of fit", the results from one model could be confirmed by the other model.

Chapter 9- Customisation of Models using cohorts from different time periods.

Aim: To assess the effect on customisation of using a temporal divide in the data.

Contents:

9.1 Introduction

Analysis in Chapter 8 demonstrated that the process of customisation could improve the performance of these models. The pace of change in medical science is fast with new drugs and treatments being introduced all the time. It is possible that data collected in the early part of the study might represent a different level of severity compared to data collected in the later part of the study. Lemeshow and Teres argued that the main reason for the customisation of models is to account for changes in treatment and provide more accurate coefficients (187).

In the last chapter, the data were divided using a random 50:50 split. However, splitting the data by time might highlight the differences in treatment. Coefficients generated in the early part of the study would represent the weight of variables from that time period. If outcome has changed over time, a model generated from an earlier cohort would expect to have a significant deterioration in the model's performance when tested on a later cohort. Having shown the customisation to be successful in improving the performance of these models, it should be possible to improve the accuracy of coefficients by using a larger development cohort, creating a more robust and accurate model.

To investigate the possible effect of changes over time during the period of the study and to investigate the possibility of improving the accuracy and robustness of the probabilities generated by the model, a further customisation of the APACHE II (GCS amended) and SAPS II models was undertaken. These models were chosen as they had performed best after the process of customisation. They are also the most commonly used systems in both the UK and in Europe, and therefore represent the standard models in these countries

9.2 Methods

The same data were used as described in Chapter 8. Data were split into a development cohort and validation cohort, with the 20 months representing the development cohort and the remaining 10 months the validation set. Models were customised using forward stepwise regression. Variables in the regression that were not significant were not included in the final models.

9.2.1 APACHE II

The development set was used to develop a new APACHE II model (New APACHE II (C)) which, like new APACHE II (B), incorporated a pre-sedated GCS if the patient was sedated for the first 24 hours. Three variables were entered into the regression to produce the new model, the APACHE II score, admitting diagnosis, and emergency operative status. The original APACHE II model was not calculated unless the patient stayed for longer than eight hours. This exclusion was not applied in the customisation study, and patients staying longer than an hour were included in the study if they fulfilled all other criteria.

As in Chapter 8 the APACHE III list of diagnoses (230 separate diagnoses) was used to collect admitting diagnosis.

The logistic regression equation for the new APACHE (C) model was as follows:
Logit=$ß_0$+(APACHE II score(GCS amended)*ß)+(Diagnostic ß).

9.2.2 SAPS II

The SAPS II score, the emergency operative status and as in the original published paper, a log of the score were included in the logistic regression to generate the New SAPS II (B) model.

The logistic regression equation for the new SAPS II (B) model was as follows:
Logit=$ß_0$+(SAPS IIscore*$ß_1$)+(Log(SAPS IIscore+1) *$ß_2$) +(Postemergancy surgery $ß_3$).

9.2.3 Data analysis

The methods used for measuring discrimination, calibration and GOF have been described in Chapter 3.

All logistic regression was carried out using SPSS software version 6.0 (SPSS Inc, Chicago, Illinois, US)

9.3 Results

9.3.1 APACHE II

Analysis for the APACHE II model used 12,936 patients, 8544 patients in the development cohort and 4,392 patients in the validation cohort, with 72 patients excluded because of missing data. The variable emergency surgery was not significant in the model and was therefore not included.

The coeffiecients for the new APACHE II (C) model were as follows (Table 9.1):

New APACHE II (C) constant (β_0) -3.8746
APACHE II (C) score +0.1537

Table 9.1 Diagnostic coefficients for new APACHE II (C) model

Diagnostic codes	op-nonop*	PtNo	New APACHE II (C)
Respiratory			
230 Respiratory arrest	N	106	-0.6584
1 Neoplasm-mouth/sinuses	O	60	-1.4882
2 Neoplasm-larynx/trachea	O	43	-1.6317
3 Neoplasm-lung parenchyma	O	39	0.4156
8 Other respiratory surgery	O	64	-1.2346
9 ARDS (non cardiogenic pulmonary oedema)	N	77	0.0913
10 Pneumonia-viral	N	29	0.1238
12 Pneumonia-bacterial	N	471	0
14 Pneumonia-aspiration/toxic	N	92	0.0277
18 Pulmonary embolus	N	28	-0.7427
20 Localised airway obstruction/oedema (mechanical)	N	41	-1.1188
21 Emphysema	N	75	-0.925
22 Asthma	N	165	-1.7201
23 Smoke inhalation	N	40	-0.7065
25 Other respiratory disorder	N	268	0.0846
Cardiovascular			
27 Aorto-femoral, fem-fem bypass graft	O	142	-1.0238
28 Fem-popliteal bypass graft	O	32	0.6142
29 Aortic aneurysm: pre-leak/dissection	O	231	-0.8126
30 Aortic aneurysm: dissection	O	27	-0.885
31 Aortic aneurysm: rupture	O	186	-0.425
32 Peripheral ischaemia	O	52	-0.6949
51 Other cardiovascular surgery	O	121	-0.5509
53 Aortic aneurysm	N	39	-0.046
57 Rhythm disturbance	N	38	0.0232
58 Acute myocardial infarction	N	38	-0.2279
60 Congestive heart failure	N	151	-0.6949

61 Cardiogenic shock	N	150	1.255
64 Septic shock - lungs (pneumonia)	N	62	0.6608
65 Septic shock - urinary tract infection	N	27	-0.5128
66 Septic shock - gastrointestinal tract	N	86	0.4628
67 Septic shock - unknown origin	N	83	-0.0392
69 Post cardiac arrest (± respiratory arrest)	N	320	0.7736
73 Other cardiovascular disorder	N	109	-0.3502
Neurological			
74 Subarachnoid haemorrhage/intracranial aneurysm	O	38	0.931
75 Subdural/epidural haematoma	O	39	-0.9636
82 Other neurosurgery	O	54	-0.5111
83 Subarachnoid haemorrhage/intracranial aneurysm	N	69	0.6211
85 Intracerebral haemorrhage/haematoma	N	45	0.6773
86 Cerebrovascular accident (CVA)/stroke	N	36	1.027
88 Seizures	N	99	-1.741
91 Meningitis	N	22	-0.338
92 Self-inflicted overdose	N	221	-1.9359
98 Non traumatic coma - cause unknown	N	25	-0.5654
99 Other neurological disorder	N	126	-0.4551
Gastrointestinal			
110 Bleeding - ulcer	O	114	0.226
111 Bleeding - laceration/tear	O	34	-0.4988
115 GI perforation/rupture	O	426	-0.0437
116 GI obstruction (any cause)	O	314	-0.2337
117 GI neoplasm (not perforation/obstruction)	O	595	-0.3111
118 Localised GI abscess/cyst	O	53	-0.2986
119 Peritonitis	O	97	0.2387
120 Pancreatitis	O	22	-1.0131
121 Cholangitis/cholecystitis	O	98	-1.6821
122 Diverticulosis	O	29	-0.9312
123 GI vascular insufficiency/embolism/infarction	O	69	0.8642
124 GI inflammatory disease	O	64	-0.4244
125 Liver transplant	O	50	-1.4963
128 Other GI surgery	O	212	-0.5304
129 Bleeding - ulcer	N	37	0.2599
131 Bleeding - varices	N	42	1.0967
134 GI perforation/rupture	N	98	0.2944
135 GI obstruction (any cause)	N	46	-0.5128
136 GI neoplasm (not perforation/obstruction)	N	38	-0.537
138 Peritonitis	N	33	0.8101
139 Pancreatitis	N	99	0.3909
142 GI vascular insuffieciency/ embolism/infarction	N	25	-0.1583
146 Hepatic Failure-toxin	N	26	-0.009
148 Hepatic failure - drug overdose	N	30	-0.3099
150 Other GI disorder	N	102	-0.1893
Renal			
157 Renal neoplasm	O	65	-1.287
163 Other renal surgery	O	116	-0.7372
172 Other renal disorder	N	87	-0.5613
Metabolic/endocrine			

175 Other metabolic/endocrine surgery	O	35	-4.3998
176 Diabetic ketoacidosis	N	20	-0.7175
187 Other metabolic endocrine disorder	N	56	-0.5365
Haematological			
190 Other haematological surgery	O	5	0.3585
196 Other haematological disorder	N	39	0.2884
Trauma			
197 Trauma- head/brain		21	-1.8269
199 Trauma - face	O	36	-0.9472
201 Trauma- abdomen	O	26	-0.4443
203 Trauma - extremities	O	91	-0.8699
204 Trauma - multiple sites plus head/brain	O	31	-0.259
205 Trauma - multiple site without head/brain	O	51	-1.8292
209 Trauma - head/brain	N	64	-0.4657
212 Trauma - chest	N	54	-0.4302
216 Trauma - multiple sites plus head/brain	N	103	-0.7525
217 Trauma - multiple site without head/brain	N	67	-0.6801
General			
Obs and gynae			
222 Pre-eclampsia/eclampsia		24	-4.6526
223 Hysterectomy		38	-1.0517
Elderly			
224 Fracture of hip		20	-1.416
226 Other elderly disorder		38	-0.3931
Miscellaneous			
229 Other miscellaneous		438	-0.9152

APACHE, Acute Physiology and Chronic Health Evaluation;*N, Non-operative; O, Operative GI, Gastrointestinal.

The new model (New Model APACHE II (C)) showed improvement in calibration when compared to the original model in both the development cohort and the validation cohort, with the Hosmer-Lemeshow GOF test improved and not significant in either cohort (Table 9.2). These changes are reflected in the calibration curves (Figure 9.1-9.2). The GOF statistic is smaller in the validation cohort than the development cohort with no evidence of poorer calibration in the validation cohort. Though it is important to note that the two cohorts are very different in size and this may be responsible for the differences in the GOF test. The new APACHE II model (C) has improved GOF when compared to the original customised model (new APACHE II (B)) using the Hosmer-Lemeshow GOF test (Table 9.2). The new APACHE II (B) model has significant Hosmer-Lemeshow GOF tests in both the validation and development cohorts, with this reflected in the calibration curves (Figure 9.2-9.3).

Discrimination has improved in both the development cohort and validation cohort, with significant improvement in the areas under the ROC curve when compared to the original model. There are improvements in the discrimination when compared to new APACHE II (B) model in both the validation and developmental cohorts. However, the

improvements are less marked then those with the original APACHE II model (Table 9.2)

Table 9.2 GOF tests and Area under the ROC curves for APACHE II and SAPS II

Model		GOF	P<	ROC	P<*	Mortality Ratio (CI)
APACHE II	Development	32.7	0.001	0.8068		0.96 (0.94-0.99
	Validation	29.0	0.001	0.8271		0.94 (0.91-0.98)
New APACHE II (B)	Development	29.42	0.001	0.8399	0.001	0.99 (0.96-1.02)
	Validation	13.89	0.08	0.8581	0.006	0.97 (0.91-0.98)
New APACHE II (C)	Development	10.9	0.21	0.8515	0.001	1.00 (0.97-1.02)
	Validation	6.71	0.57	0.8569	0.001	0.98 (0.95-1.02)
SAPS	Development	69.1	0.001	0.8368		0.96 (0.94-0.99)
	Validation	34.1	0.001	0.8456		0.95 (0.92-0.99)
NEW SAPS (A)	Development	5.22	0.733	0.8395	0.675	1.00 (0.97-1.02)
	Validation	4.54	0.805	0.8478	0.001	0.99 (0.95-1.02)
New SAPS (B)	Development	16.2	0.039	0.8395	0.675	1.00 (0.98-1.03)
	Validation	10.14	0.26	0.8455	0.5106	0.99 (0.96-1.03)

GOF, Goodness of Fit; ROC, Receiver Operating Characteristics; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; CI, Confidence Intervals

* P-value of area under the ROC curve when compared to the original model.

Table 9.3 Source of admission mortality ratios and P values for APACHE II for validation cohort

Source of admission	n=	Original APACHE II	X^2 P=	New APACHE II (B)	X^2 P=	New APACHE II (C)	X^2 P=
A&E	599	1.06 (0.97-1.14)	0.431	0.94 (0.87-1.02)	0.406	0.96 (0.88-1.04)	0.527
Recovery/theatre	2012	0.77* (0.70-0.84)	0.000	0.98 (0.90-1.06)	0.680	0.98 (0.90-1.06)	0.708
Ward in this hospital	1307	1.04* (0.99-1.10)	0.301	1.00 (0.95-1.05)	1.000	1.02 (0.97-1.06)	0.718
Other ICU in this hospital	75	1.19 (0.96-1.42)	0.301	1.11 (0.90-1.31)	0.538	1.08 (0.88-1.28)	0.639
ICU in another hospital	131	1.05 (0.86-1.24)	0.740	0.95 (0.78-1.13)	0.740	0.97 (0.80-1.15)	0.841
Other area in another hospital	264	0.84 (0.69-0.99)	0.153	0.77* (0.64-0.91)	0.032	0.81 (0.67-0.95)	0.078
Home/clinic	4	1.06 (0.00-2.47)	0.920	0.89 (0.11-1.71)	0.920	0.77 (0.00-1.65)	0.791
Overall patients	4392	0.94 (0.91-0.98)	0.033	0.97 (0.93-1.00)	0.260	0.98 (0.95-1.02)	0.522

APACHE,Acute Physiology and Chronic Health Evaluation, A&E, Accident and Emergency; ICU, Intensive Care Unit; *Significance as indicated by CIs

Figure 9.1 Calibration curve for original APACHE II model in validation cohort.

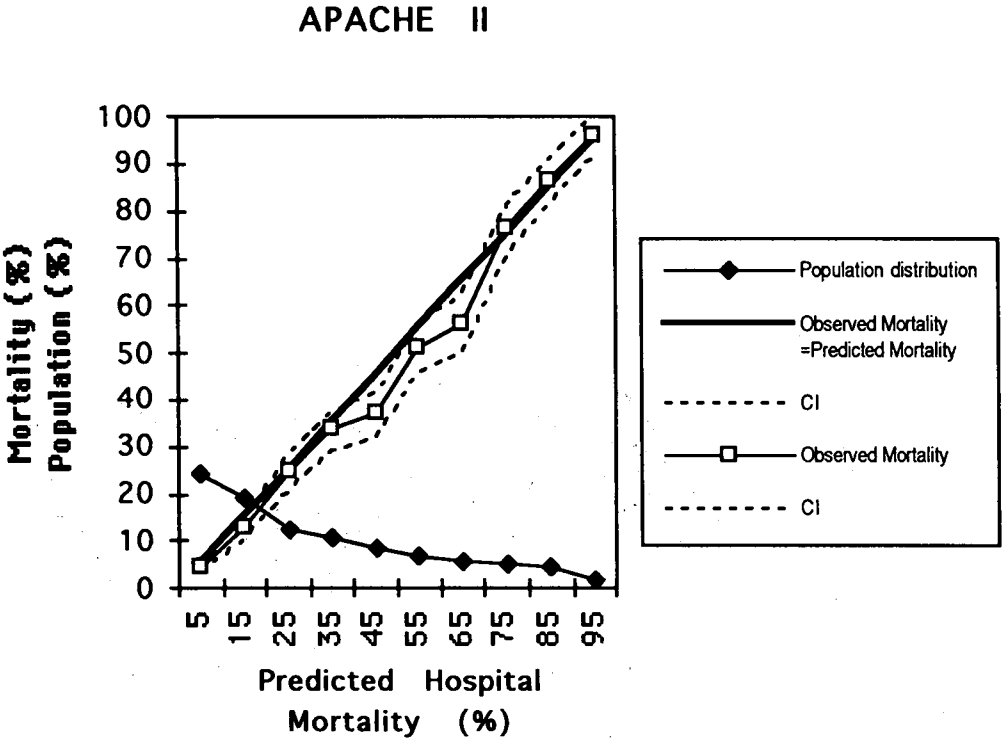


Figure 9.2 Calibration curve for New APACHE II (B) model in validation cohort.

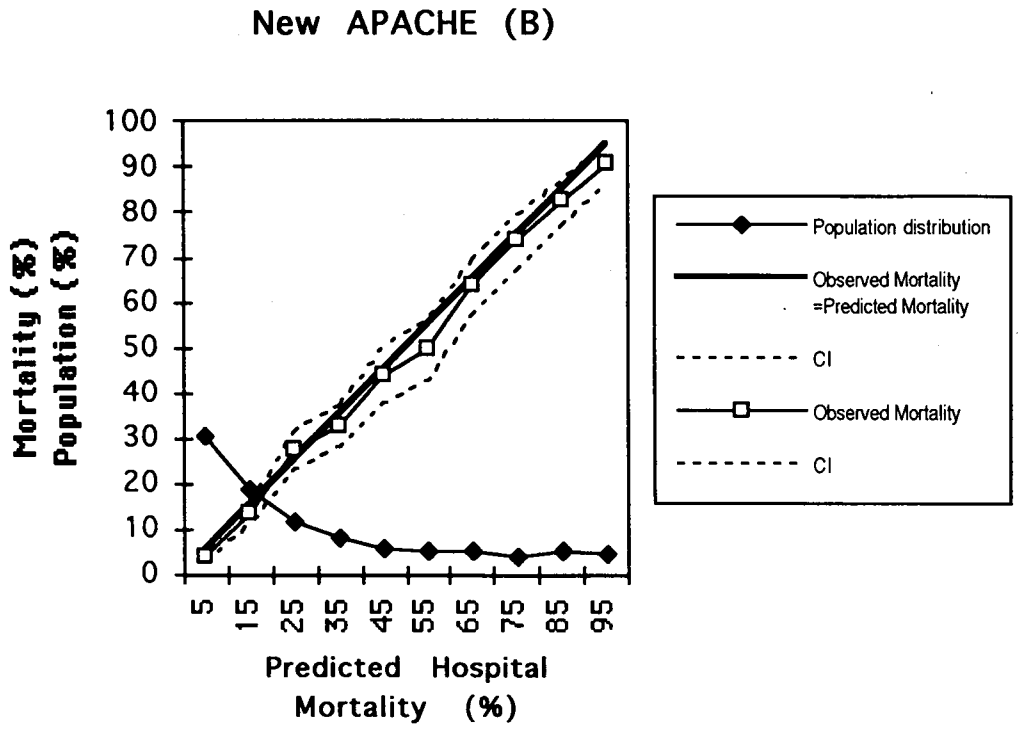


Figure 9.3 Calibration curve for New APACHE II (C) model in validation cohort.

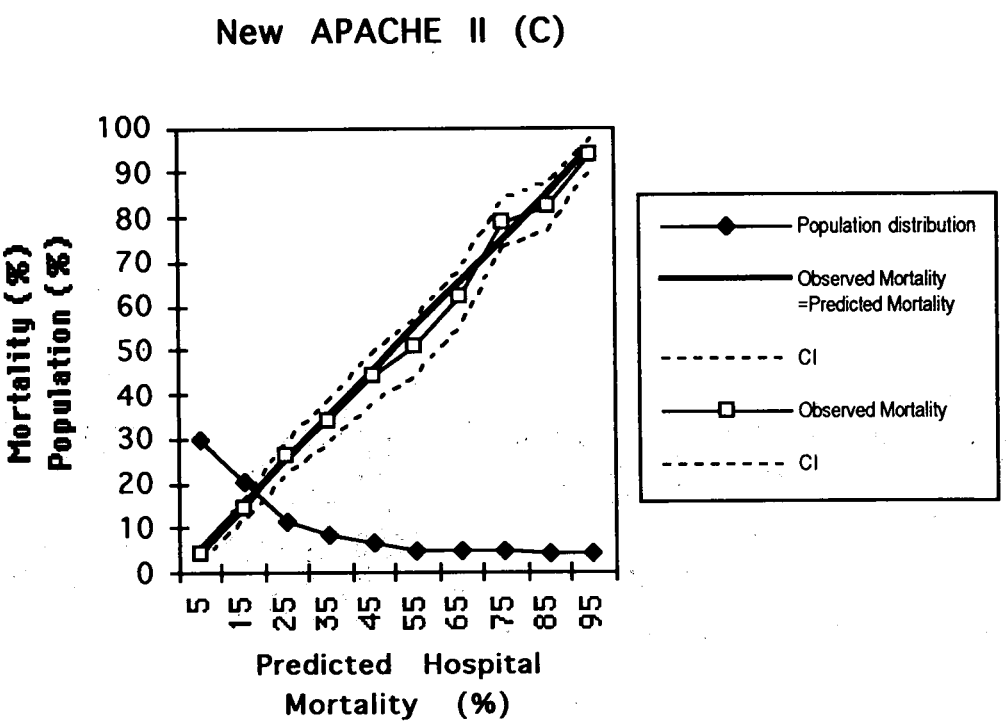
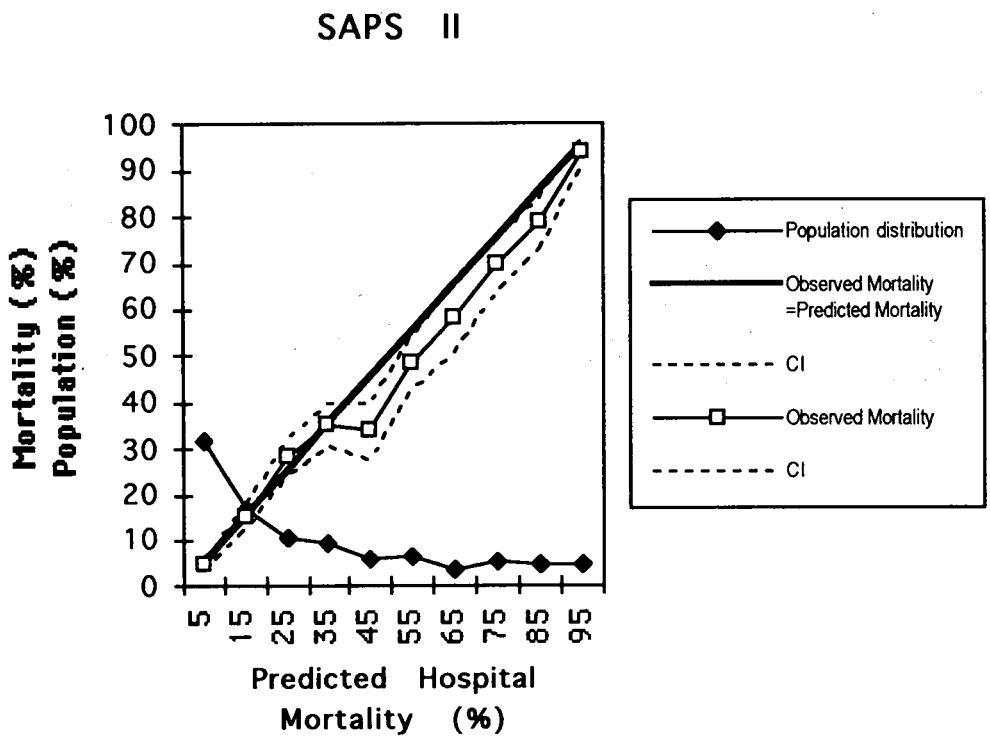


Figure 9.4 Calibration curve for original SAPS II model in validation cohort.



The Uniformity of Fit has also improved when compared to the original model. When grouped by source of admission significant differences were found in Recovery/theatre and Ward in this hospital for the original model (Table 9.3). Other area in another hospital was significant for the New APACHE II (B) model. No significance was found for any of the tests in the New APACHE II (C) model. When patients were grouped by their APACHE diagnostic system significant differences were found in the cardiovascular, neurological, gastrointestinal and general groups for the original model (Table 9.4). No significance was found using either the New APACHE II(B) or New APACHE II (C) models in any of the groups.

Table 9.4 APACHE System mortality ratios for APACHE II for validation cohort

APACHE System	n=	Original APACHE II	X^2 p=	New APACHE II (B)	X^2 p=	New APACHE II (C)	X^2 p=
Cardiovascular	959	1.06* (1.01-1.12)	0.195	0.97 (0.92-1.02)	0.480	0.98 (0.93-1.03)	0.680
Respiratory	864	1.06 (0.97-1.14)	0.354	1.00 (0.92-1.07)	1.000	1.01 (0.94-1.09)	0.823
Neurological	398	1.18* (1.04-1.33)	0.084	0.92 (0.80-1.04)	0.393	0.92 (0.80-1.05)	0.427
Gastrointestinal	1427	0.78* (0.71-0.84)	0.001	0.97 (0.90-1.05)	0.610	0.99 (0.92-1.07)	0.920
Renal	117	0.83 (0.54-1.12)	0.371	0.85 (0.57-1.12)	0.420	0.95 (0.65-1.25)	0.806
Metabolic/ Endocrine	51	1.10 (0.69-1.51)	0.718	0.98 (0.69-1.28)	1.000	1.09 (0.76-1.41)	0.764
Haematological	21	1.49 (1.02-1.94)	0.170	1.28 (0.89-1.69)	0.377	1.41 (0.98-1.84)	0.230
Trauma	286	0.91 (0.60-1.23)	0.647	0.80 (0.52-1.09)	0.254	0.75 (0.48-1.03)	0.145
General	269	0.45* (0.21-0.70)	0.001	1.12 (0.72-1.53)	0.603	0.85 (0.50-1.21)	0.484
Overall patients	4392	0.94 (0.91-0.98)	0.033	0.97 (0.93-1.00)	0.260	0.98 (0.95-1.02)	0.522

APACHE,Acute Physiology and Chronic Health Evaluation; *Significance as indicated by CIs
*Significance as indicated by CIs

9.3.2 SAPS II

Analysis for the SAPS II model used 12,944 patients, 8,552 in the development cohort and 4,392 in the validation cohort with 64 of the original 13,008 patients being excluded because of missing data.

The coeffiecients for the new SAPS II (B) model were as follows:

New SAPS II constant (β_0)	-5.1571
SAPS II score	+0.0767
Log of SAPS II score +1	+0.2796
Emergancy Operative	-0.3844

Calibration has improved in the new SAPS II (B) model when compared to the original SAPS II model with reduced Chi Squared values for both the validation and development cohort. Significant differences were found using the Hosmer-Lemeshow GOF test in the developmental cohort for the new SAPS II (B) model, while the statistic is not significant in the validation cohort (p= 0.26) (Table 9.2). This difference may be due to smaller numbers in the validation cohort. However, the new SAPS II (B) model has not improved when compared to the New SAPS (A) model, with model (A) producing smaller and non significant Chi-Squared statistics in both the development and validation cohort. These GOF results were reflected in the calibration curves (Figure 9.4-9.6). Discrimination has changed little in the new SAPS II (B) model with little difference in the area under the ROC curve when compared to the original model or the new SAPS (A) model (Table 9.2).

When uniformity of fit is considered using patients grouped by source of admission, significant differences using either the mortality ratios or the Chi Squared test were found in the original model for A&E, Recovery theatre and Ward in this hospital and Other area in another hospital. The SAPS II (A) model had significant differences in the A&E, Ward in this hospital and Other area in another hospital. The SAPS II (B) had significant differences in the A&E, Recovery/ theatre and Other area in another hospital (Table 9.5).

Table 9.5 Source of admission mortality ratios and P values for SAPS II for validation cohort

Source of admission	n=	Original SAPS II	X^2 P=	SAPS II (A)	X^2 P=	SAPS II (B)	X^2 P=
A&E	599	0.86* (0.78-0.93)	0.026	0.85* (0.78-0.93)	0.022	0.87* (0.79-0.94)	0.037
Recovery/theatre	2012	0.85 (0.78-0.92)	0.002	0.97 (0.89-1.05)	0.590	0.96 (0.88-1.04)	0.446
Ward in this hospital	1307	1.09* (1.04-1.14)	0.038	1.08* (1.03-1.13)	0.065	1.10* (1.05-1.15)	0.024
Other ICU in this hospital	75	1.12 (0.92-1.33)	0.480	1.12 (0.90-1.33)	0.517	1.13 (0.92-1.35)	0.458
ICU in another hospital	131	1.02 (0.84-1.19)	0.888	1.00 (0.82-1.18)	1.000	1.02 (0.84-1.20)	0.888
Other area in another hospital	264	0.81 (0.67-0.94)	0.071	0.79* (0.65-0.93)	0.046	0.80* (0.66-0.94)	0.064
Home/clinic	4	0.72 (0.00-1.67)	0.740	0.70 (0.00-1.75)	0.740	0.70 (0.00-1.73)	0.740
Overall patients	4392	0.95 (0.92-0.99)	0.078	0.99 (0.95-1.02)	0.597	0.99 (0.96-1.03)	0.823

SAPS, Simplified Acute Physiology Score, A&E, Accident and Emergency; ICU, Intensive Care Unit; *Significance as indicated by CIs

Figure 9.5 Calibration curve for New SAPS II (A) model in validation cohort.

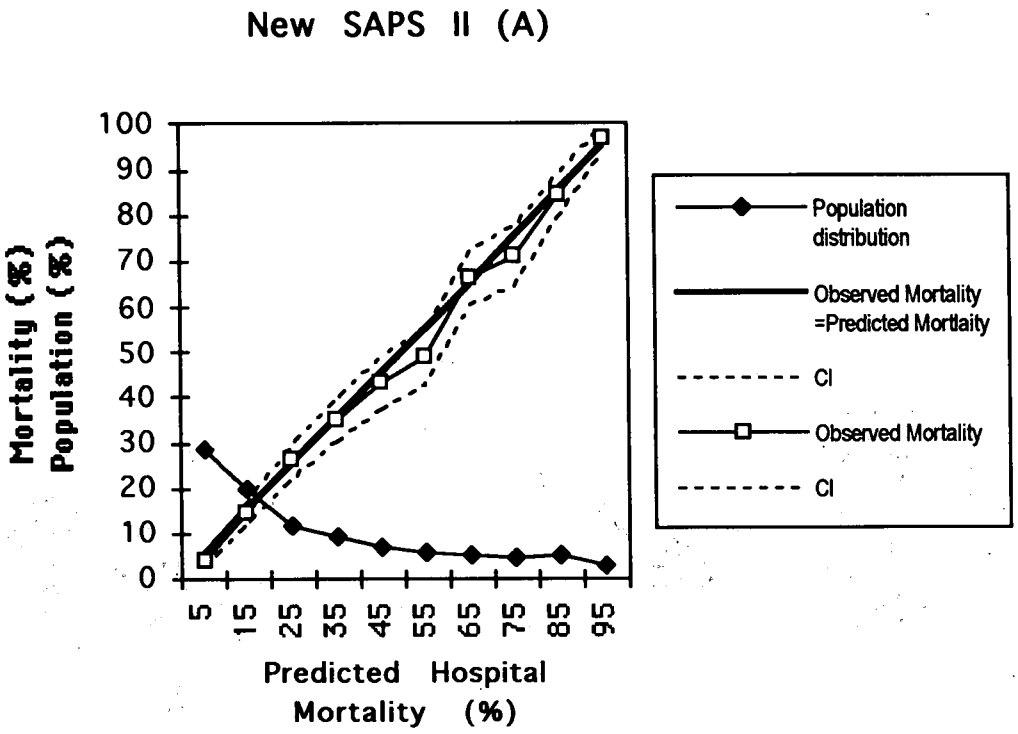


Figure 9.6 Calibration curve for New SAPS II (B) model in validation cohort.

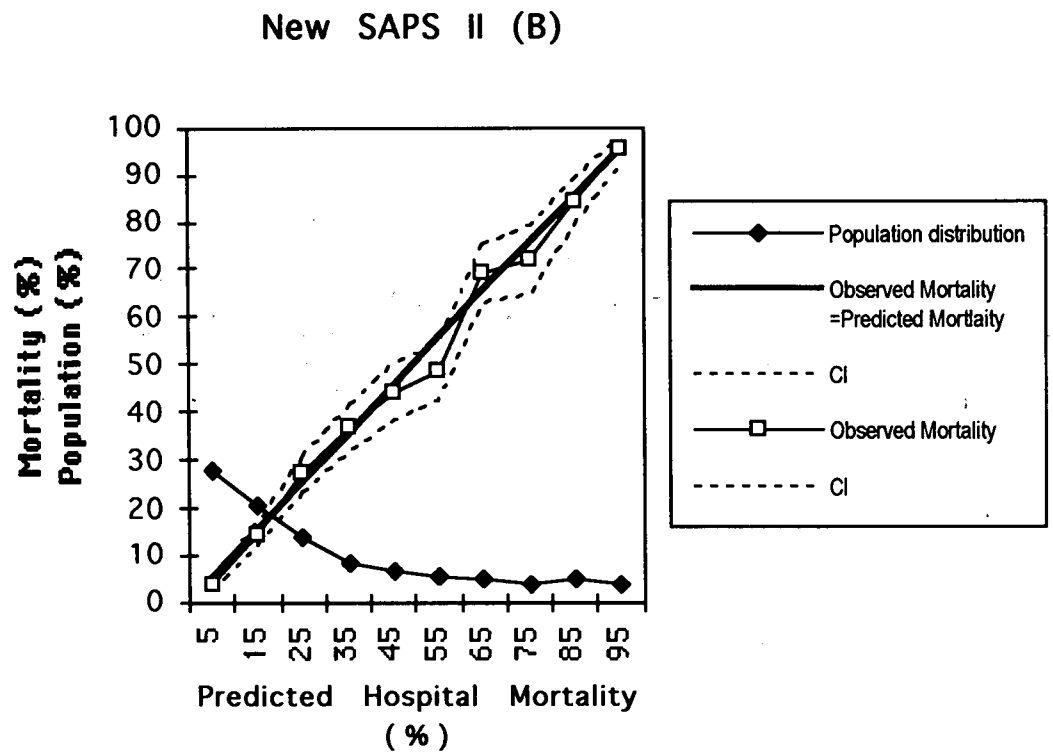


Table 9.6 APACHE System mortality ratios for SAPS II for validation cohort

APACHE System	n=	Original SAPS II	χ^2 p=	New SAPS II (A)	χ^2 p=	New SAPS II (B)	χ^2 p=
Cardiovascular	959	1.00 (0.94-1.05	0.920	1.03 (0.98-1.09)	0.507	1.04 (0.98-1.09)	0.420
Respiratory	864	1.08* (1.00-1.16)	0.188	1.05 (0.97-1.13)	0.393	1.08 (0.99-1.16)	0.219
Neurological	398	0.76* (0.65-0.87)	0.004	0.76* (0.64-0.87)	0.005	0.77* (0.66-0.89)	0.008
Gastrointestinal	1427	0.98 (0.91-1.05)	0.708	1.08 (1.01-1.16)	0.114	1.09 (1.01-1.16)	0.103
Renal	117	0.70 (0.47-0.92)	0.084	0.71* (0.47-0.94)	0.095	0.71* (0.47-0.95)	0.102
Metabolic/ Endocrine	51	0.86 (0.60-1.11)	0.554	0.86 (0.58-1.13)	0.554	0.85 (0.58-1.13)	0.554
Haematological	21	1.51* (1.08-1.92)	0.157	1.52* (1.08-1.96)	0.144	1.54* (1.09-1.99)	0.133
Trauma	286	0.51* (0.31-0.72)	0.001	0.53* (0.32-0.74)	0.001	0.53 (0.32-0.74)	0.001
General	269	0.65 (0.37-0.93)	0.052	0.61* (0.34-0.89)	0.028	0.60 (0.32-0.88)	0.020
Overall patients	4392	0.95 (0.92-0.99)	0.078	0.99 (0.95-1.02)	0.597	0.99 (0.96-1.03)	0.823

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score, *Significance as indicated by CIs

Grouping patients by the APACHE diagnostic system showed significant differences in the original model using for at least one of the CI for the mortality ratios or the Chi Squared statistic, for the respiratory, neurological and trauma groups. The neurological, renal, trauma and general groups were found to have significant tests in the new SAPS (A) model and the new SAPS II (B) model (Table 9.6). Regardless of significance the parameter estimates are similar.

9.4 Discussion

9.4.1 APACHE II

The new APACHE II (C) model has improved its performance with improved calibration, discrimination and uniformity of fit in both cohorts (developmental and validation) compared to the original model. There would appear to be no discernible change in the model's performance over time between the two time periods analysed here. The model has also improved its performance when compared to the customised GCS amended model in Chapter 8 (new APACHE II model (B)). This improvement may be due to the increased number of subjects in the development set allowing for more robust coefficients to be generated. It is important to note that the numbers in the validation cohort are considerably smaller than those of the developmental cohort and

other data sets in this thesis. Some of the apparent improved performance in the validation cohort may be due to the smaller numbers in this cohort. Caution should be used if comparing the results in the developmental cohort to those in the validation cohort. However, improvement in comparison to the original and new APACHE II (B) model would suggest that these changes are genuine.

9.4.2 SAPS II

Compared to the original SAPS II model the new model has improved its calibration and maintained its discrimination. Although significant differences between the observed mortality and that predicted by the new SAPS (B) model were found in the development set using the GOF test, the Chi Squared value is smaller than that of the original model. Significant differences are found in the original SAPS II model in the validation cohort but there is no significance in the observed and estimated mortalities using the GOF test in the new SAPS II (B) model. All SAPS models have poor uniformity of fit with many of the larger groups being significant.

The new SAPS II (B) model has improved its performance when compared to the original model. However, despite a larger developmental cohort, the model has poor performance when compared to the new SAPS II (A) model. This poorer calibration, when compared to the original customised model is difficult to explain. Time would not appear to be a factor as the performance of the new SAPS II (B) model is poorer than the new SAPS II (A) in the developmental set as well. The coefficients generated from the random 50:50 split would appear to give a better model than using the larger development cohort split temporally. The coefficients produced in the first cohort would appear to be less representative in both the first and second cohort than the original customised SAPS II model (new SAPS II (A) model). One reason may be a change in data collection over time, with data errors in the earlier cohort affecting the ability of the model to produce accurate coefficients both in the development cohort and the validation cohort. This is however conjecture as it is difficult to understand why this would have an effect on the SAPS II model but not the APACHE II model. There may have been either a change in practice over the time period or a change in type of patients being admitted. However, it would be expected that the performance in the developmental set would be good with poor performance in the validation cohort.

The results from this temporal divide in the data seem to confirm the ability of the customisation process to improve the models as both the new APACHE II (C) and new SAPS II (B) models showed improved performance when compared to the original models. There would appear to be no evidence of deterioration over time but this may be

due to the narrow time span which the data cover. It may also be, in part, due to the smaller sample size in the validation cohort.

9.4.3 Overall

The new APACHE II model appears to have improved performance when also compared to the new APACHE II (B) model from Chapter 8, with no significance in Hosmer-Lemeshow GOF tests and no significant differences found in uniformity of fit in the new APACHE II (C) model. The new SAPS II (B) model shows similar improvement compared with the original model but not in comparison to the new SAPS II (A) model.

Although no apparent change over time has been noted, this may be a result of the time span which the data cover (two and half years). It may be that comparisons over time should be made over a longer time period, allowing the two cohorts to cover non-adjacent time periods. As there is an improvement in the performance of the new APACHE II (C) model then this model should be used in preference to the other APACHE models described in this thesis, if case mix adjustments are to be carried out in Scottish ICUs. Although the new SAPS II (B) model has improved its performance when compared to the original SAPS II model, the GOF has not improved when compared to the new SAPS II (A) model. As the new SAPS II (A) model has slightly better performance with good GOF it should be used in preference to the other SAPS II models in this thesis.

Chapter 10-Discussion

Contents:

10.1 Performance of models in Scottish data

The initial analysis (Chapter 5) would suggest that all the models in this study have poor calibration, with all showing significant differences in the Hosmer-Lemeshow test. Discrimination would appear to have been maintained in most models. There is no agreed point at which these models can be said to have adequate performance, although a consensus conference agreed which statistics should be used to assess their performance (22). The significance in the Hosmer-Lemeshow test may be a result of the extra sensitivity afforded by large numbers (142), with the differences having little clinical significance. However, the Hosmer-Lemeshow GOF table (Table 5.8) seems to show quite considerable differences in the numbers of observed deaths and those expected by the models, especially in the APACHE III and MPM_0 models.

These results appear to be confirmed by the uniformity of fit (Chapter 6). In fact the models appearing to have the best fit in Chapter 5 (APACHE II and SAPS II), also had significant differences in important subgroups. The better performance of these models (APACHE II and SAPS II) seems to have resulted from the averaging used by the different tests. All models show significant differences when patients are grouped by both source of admission and their admitting diagnosis and it could be argued that they have failed to adjust adequately for case-mix.

10.1.1 Quality of care

There are a number of possible reasons for the failure of these models in the Scottish population. They highlight a difference in quality of care between the ICUs in Scotland and those making up the original databases in which the models were developed. An other possible explanation is that they fail to adequately adjust for case mix in the Scottish ICU population. There is some evidence to suggest that quality of care may be an issue. The overestimation of mortality in the UK APACHE II model (based on data from the earlier UK APACHE II study) may represent a lower standard of care than that of the original US APACHE II study. The better calibration of the APACHE II model, it could be argued, represents the standard of care in the US in the early eighties, which represents the standard of care in Scottish ICUs at the present. The underestimation of mortality in the APACHE III models would then represent a higher level of care now practised in the US. The contrast of the three models appear to provide persuasive evidence for this argument. This may be supported by the fact that a recent large study of ICUs in South West Thames showed underestimation of mortality in the APACHE III model (84). The authors pointed to two possible reasons for this underestimation, differences in the standard of care between the original US study and the UK units, and

the failure of the model to adequately adjust for case mix. The study suggests that if the underestimation of mortality is a result of poor care, this would mean that the admission of an identical group of patients in both countries would result in a 25% higher mortality in the South of England. The authors suggest that there are some explanations for differences in quality.

- Resource allocation for Intensive Care is considerably greater
- United States recognise Intensive Care as a speciality
- United States more likely to have dedicated critical care training programmes
- Resource strategies in the UK often lead to refused admissions, delay in the ICU admission, high requirement for inter hospital transfer, early ICU transfer and increased ICU readmissions.

They suggest that it is more likely that differences between the observed and estimated mortality are due to the poor fit of the equation in the South of England data. The authors argue that this could be due to the model's inability to measure international differences in case mix, admission practice or lead time bias. Other differences include systematic differences in medical definitions or diagnostic labelling, effectiveness of therapy, data collection, and failure of the APACHE III system to weight sufficiently for their impact on the UK mortality. The authors point to the fact that there is no independent standard measure of quality of care by which to verify the accuracy of the model.

However, there is evidence to contradict the argument that differences in quality of care are responsible for the underestimation of mortality by the APACHE III model. Both the MPM II and SAPS II models represent data from a similar time period to the APACHE III system and show no consistent under or overestimation. A large UK study showed similar fit in the APACHE II model to the performance of the model in this study with an underestimation of mortality at the highest end of severity of illness, and a poor estimation in certain subgroups (73). A small study in one ICU in Britain showed that both APACHE III and APACHE II underestimated mortality (83). Teres and Lemeshow argued that there have only been five studies which look at the performance of severity of illness models on new patients and that also used the correct statistics to measure performance (137). In these studies a pattern emerges where the models maintain their discrimination but show poor calibration. A similar pattern is also reflected in this study. A study evaluating the performance of the APACHE III model, in an independent US database, showed the model maintaining the discrimination but with poorer calibration (187). This poor calibration was also reflected in poor performance in some subgroups. The results from analysis of the uniformity of fit (Chapter 6) would also question the argument, that the poor performance of the models is a reflection of the quality of care in Scottish ICUs. With all models showing poor performance in two or more large

subgroups, it would seem unlikely that poor quality of care would not be reflected by poor performance in all subgroups. However, some of the groups overestimate mortality, and some underestimate mortality, which suggests that the models' failure to adjust for case mix is responsible for the models' poor performance.

10.1.2 Other reasons for poor fit

10.1.2.1 Differences in mortality

So, why do these models not maintain performance in Scottish ICUs (and other ICUs)? There is no clear answer to this, but there are a number of hypotheses. One contributing factor may be the higher mortality in Scottish units than in the original studies. The mortality in the original databases was considerably less than that reported in this study (MPM II, 20.8%; APACHE III, 17.3%; APACHE II, 19.7%; SAPS II, 21.9%; Scottish Study, 29.4%). It would be over simplistic to surmise that the much lower mortality rate in the APACHE III database is responsible for the consistent underestimation of mortality by the model in this study. However, the lower mortality rate in the APACHE III database must have some impact on the weights allocated to the variables in the model, if compared to the equivalent probabilities in Scottish ICUs. The higher mortality rate in this study probably reflects the fewer available ICU beds per head of the Scottish population.

10.1.2.2 Selection bias

It could be argued that the high mortality in this study reflects a selection bias when compared to the original databases. Selection bias must be a consideration when looking at the performance of the models. Four units in the study were HDU/ICUs who did not admit consecutive patients but rather decided when patients were "ICU type patients". It would have been preferable for these units to admit all their patients to the study but because of the large throughput of patients in these units they would not have had the resources to collect the large amount of data required by the Audit. The profile of those patients being admitted is similar to the ICUs in the study. If there is a selection bias then it is more likely to be because of the level of ICU beds available to clinicians in Scotland. The presumption would then be that the average severity of illness for units in the Scottish Audit is considerably higher than that of the units involved in the development of the original models. The original APACHE III database had an average APS of 50, with the average APS in the Scottish database 60.8, which may reflect a higher level of severity of illness.

10.1.2.3 Data quality

For severity of illness models to accurately adjust for case mix it is important that the data collection is accurate, complete and consistently collected. It is possible that the data accuracy affected the performance of the models. However, the re-scoring of 10% of patients over a nine month period showed that all models, with the exception of the MPM_0 model, had no significant over or underscoring. This suggests that although errors in collecting the data may have had an effect, errors did not cause the models to systematically over or underscore. The APACHE models all need a diagnosis before probabilities can be calculated. There were a small number of records with missing diagnoses (n=67) but this is so small it is unlikely to have had an effect on the performance of the models. All models treat missing data as normal, however, it is impossible to say whether data not recorded had an effect on the accuracy of the models. It is possible to hypothesise that, in better resourced US units, more nursing time and less constraints on the purchasing of blood tests led to more complete data with higher scores and probabilities for less severe patients. One study showed that increasing the amount of data collected by using a computerised system raised the estimated mortality by 15-25% depending on the model (APACHE II, MPM_0, SAPS II) (177). It is also very difficult for this study to assess the impact of variability of data collection on different units, or variability between different data collectors. Although the Research Nurse did report variability between units in the way data were collected this has never been quantified. Although, it is unlikely that this aspect is solely responsible for the poor performance of the models, it is not possible to rule it out. Future work, especially if it involves the comparison of different units, must address issues of data quality.

The consistent collection of data is vital if a model is to be effective in accurately measuring case mix. The Audit has attempted to ensure that this has been the case. It was hoped that a number of steps would help to ensure consistent data:

- providing software with extensive range checking
- training ICU staff both in the use of the software and data collection
- providing manuals and help screens with data collection rules
- reporting data errors to ICUs on 10% of all their records
- providing a support service for both the software and for answering questions on the rules of collection.

The Audit has attempted to collect data as consistently as possible, however, there are areas where it was impossible to determine policy on the different ICUs. Issues such as who collected the data, what blood tests were routinely collected by ICUs, where ICU treatment started, and the amount and frequency of data collection, were all decisions that were made exclusively by the different ICUs. The study made every attempt to follow the

protocols available for the different models. However, for most of the scores the information available on this subject was not complete. The APACHE III model, because of the support of AMS, has the most information on this subject. They provided a data collection manual for APACHE III and support in answering questions on rules for data collection. It is difficult to know what effect different data collection has on comparisons with the original studies, or even within the original studies themselves. Attempts are being made by the Intensive Care National Audit and Research Centre (ICNARC) in England and Wales to standardise data collection in ICUs providing data to its case mix programme.

10.1.2.4 ICU culture

It is hard to assess what actual differences there are in ICUs and their treatment in the Scottish setting compared to the ICUs contributing data to the construction of the different models. However, data in this study showed a considerable improvement in the estimated mortality for those patients for whom a GCS from the 24 hour period was not available, and a pre-sedated GCS was used (Chapter 7). This may reflect a difference in the nature of ICUs in the US and Britain. It could be surmised that patients in American ICUs are admitted at an earlier stage of their disease, before they are sedated and ventilated. This would allow for the use of a GCS score in the APACHE models. Patients in Britain may enter ICU at a later stage of their disease, when they have already been intubated and sedated and where the rules of the APACHE models will assume a normal GCS. Without knowing more about differences between the ICU cultures in the units from which the original models were developed and cultures in Scottish units, it is hard to know what impact this may have had on the performance of the models. This is an area that needs to be investigated before any meaningful comparisons can be made between Scottish ICUs and the ICUs collecting data for the original development of these models.

10.2 Improving performance in models

The models in this study were being investigated to identify the most legitimate tools for allowing valid comparisons between different Scottish ICUs. The success, illustrated in the literature by a number of authors, in improving the performance of various models made the customising of the models in this study a logical next step (139,141,142). The process of customisation in Chapter 8 has no doubt improved the performance of all the models, with all but the APACHE II (A) and the new MPM_{24} (A) models having improved performance when compared to the original analysis. However, despite having considerably reduced Chi Squared values, all except the new SAPS II had significant Hosmer-Lemeshow GOF tests. Results for the uniformity of fit showed significant

differences between observed mortality and the mortality expected by the models except in the New APACHE II (A & B) systems. This was an aspect for which the original studies did not publish any data, making it impossible to assess if this was a weakness in the original models. It was also not possible to customise the APACHE III model as the original equation and complete methodology were copyrighted. As the APACHE III model had consistent underestimation of mortality and less variation in uniformity of fit, it could be the ideal model for this approach. As the Audit had decided that it was no longer going to pay for the use of the APACHE III model it was clear that there was little to be gained for the Scottish units in pursuing this possibility. The inclusion of the diagnostic weighting may be, in part, responsible for the good uniformity of fit in the APACHE II model. The results from the second random sample and also from analysis in Chapter 9 would seem to indicate that the process is reproducible and therefore robust. The APACHE II models (new APACHE II (check), new APACHE II (C)) created from both these samples maintained similar performance to that of the original customised model (new APACHE II (B)). Though the poor performance of the SAPS II (B) model in Chapter 9 might question the stability of the original customised model (new SAPS II (A)) in other samples. When the data was split temporally there was no evidence of poorer performance of the model (new APACHE II (C)) in the latter half of the data. This may be due to the relatively short period (two and half years) covered by the sample. It would be important to repeat this analysis over a longer period , maybe with time periods that are not consecutive. Despite the improvements gained by the models using the customisation process there still remain questions about their accuracy. This has also been reflected in other studies (129,139). However, as there is no bench mark at which these models become legitimate this does not necessarily rule out their use for comparing units' performance.

Teres and Lemeshow have argued that while customisation is a legitimate method, it is important that it is used appropriately (188). Once a model has been customised it should then be used to focus on explaining the differences based on case mix or quality of care differences. They argue that the main reason for the need for customisation is the change of treatment over time.

10.3 Future of severity of illness models

10.3.1 Future research and improvement into severity of illness models

There would appear to be a presumption made about severity of illness models that, properly calibrated, they accurately adjust for case mix (21,151). However, even a well calibrated model will be unlikely to account for all the variation in mortality. Some of

this variation will be due to differences in clinicians' practice or elements of an ICU's structure and organisation. However, some unmeasured variation may be due to other factors. There is considerable evidence in a number of Scottish studies of the effect that deprivation has on outcome. Work in Glasgow on the differences in outcome from a number of surgical procedures was shown to be associated with deprivation (181). A recent report into deprivation and health in Scotland also found that there were differences in outcome associated with deprivation (179). Other work done on a large cohort of patients from the Renfrew and Paisley areas in Scotland found deprivation associated with less favourable profiles of cardiovascular risk factors (180). Some of the evidence may demonstrate differences in access to services but there is considerable evidence that point to a "miles on the clock" effect. Future work needs to be done to identify whether deprivation is associated with higher mortality, especially when adjusted for case mix. Other, less tangible variables like "will to live" may also have an impact on outcome from an ICU. It is easy to see that if issues like deprivation do have some bearing on outcome then they may have varying effect on individual ICUs. Any future research into improving severity of illness models should look into other possible determinants of outcome not already included in the models.

This study, like others, has shown that when models are implemented in different cohorts than those in which they were developed, discrimination remains high but calibration deteriorates. Although, the models' accuracy can be improved, there still remain doubts when their accuracy in subgroups is tested. There is considerable effort required in developing models from scratch. The collection of large amounts of information on large numbers of patients is required. However, all the models have maintained reasonable discrimination with the possible implication that the variables in the model remain predictive of mortality in different ICU cultures. However, there is some evidence to suggest that the areas under the curves are not as sensitive to changes in accuracy as the Hosmer-Lemeshow test (142). More complete understanding of the existing models is needed before any new major models are developed.

Teres and Lemeshow have pointed to the possible effects of having status at end of hospital stay as an outcome measure (188). As patients will be followed up for varying lengths of time, this might be a possible source of bias. They also argue that increasing numbers of patients are transferred to other hospitals for specialist treatment, giving a positive outcome at the original ICU for patients that subsequently die. They point to the use of 90 days as a less arbitrary outcome measure. The authors say that future research should also concentrate on the acute episode of care, rather than the first 24 hours of Intensive Care treatment (137,188). Many patients receive treatment in other intensive areas, and by the time they arrive on the ICU they have been sedated and stabilised. This

prevents the accurate measurement of their severity of illness. There also needs to be further understanding as to what other factors have a significant impact on variation of mortality. It would be possible to use the existing models to investigate this and Knaus et al have gone some way in doing so (98).

There has also been little investigation into data variability and quality. Like this study, most studies have had limited analysis of the effect of errors in the data they have collected. These models require the collection of a considerable amount of data and there is a need for more in-depth research into the effect of errors. Research is also needed to identify the effect on the quality of data of different data collectors, the way data are collected, where ICU treatment starts, and the consistency of data collection.

This study demonstrated that it is possible to improve the performance of the models using a process of customisation. The examination of this process in a separate random sample and in a temporal split in the data would seem to indicate that the new APACHE II models produced are both stable and robust. As there would appear to be no perceivable differences in the data over time a new model developed from the whole data set, using the methods employed in this thesis, would allow for the creation of the best possible model from the data available.

10.3.2 Future of severity of illness models in Audit

It is clear that there remain doubts in the ability of these models to adjust for case mix and therefore do not allow comparison of different ICUs. This thesis has not attempted to look at variation between ICUs and this remains an important issue. The results from this study show that even using the new Scottish coefficients, where patients are admitted from and the type of patient admitted may affect a unit's apparent performance. The effort to collect these data are considerable and it is hard to justify the resources to do this if there is no benefit to patient care. For these models to be used as an Audit tool then they must be of use in the Audit Cycle. If the data do not allow any conclusions to be drawn on quality of care and consequent changes in practice then they can not be used in Audit.

How these models have been used: As part of the Scottish Intensive Care Audit an annual Audit meeting is organised to feedback results from the Study as well as the reporting of results of local Audits. As part of this, comparisons are made and performance of ICUs reported using mortality ratios. Information is anonymised with ICUs being represented by a letter. Personnel from each ICU are made aware of which letter refers to their unit. However, as yet, no ICU has been significantly different from the rest of the units and the

rank of the different ICUs has not been investigated. So, there has been limited use of these models as a measure of quality of care. However, the Audit has attempted to use them to assess practice in some limited way. A paper by Connors et al on the effectiveness of Pulmonary Artery (PA) catheterisation in ICUs showed that this procedure was associated with poor hospital outcome even when adjusted for case mix (93). In response to this evidence the Audit retrospectively analysed data using the APACHE II model and found similar poor outcome associated with the procedure. PA catheterisation is not a therapy in itself but rather provides the means for intensive monitoring and is associated with certain types of therapy. There is strong evidence that PA catheterisation can be used to deliver beneficial therapies when associated with strict protocols (189). To stop using this technique would seem to be inappropriate when there are clear benefits. However, in the light of the evidence found by the Audit, clinicians from the ICUs in the study are now reviewing these protocols.

As well as work done centrally by the Audit, clinicians from the ICUs have reported a number of studies which have relied on case mix adjustment. Two Glasgow units used data from the APACHE II model to identify patients who died but who had low probabilities of mortality. Case note reviews were then used to investigate the reasons for these patients' deaths in an effort to identify problems in care. Another study used the APACHE II model to adjust for case mix to allow comparisons of levels of workload to be associated with outcome. The study found that even after adjustment there was a threefold increase in mortality at the highest levels of workload. The unit has since been allocated more staff and the intention is to re-run the study to assess the effect of the extra staff.

The future of severity of illness models in Audit: Any Audit in Intensive Care that uses death as the outcome measure must have some means of adjusting for case mix. The heterogeneous nature of ICU patients make it difficult to imagine how direct comparisons of mortality could be used to make conclusions about the effectiveness of care. At present these models remain the most accurate means of case mix adjustment. Despite the failure of the models in this study, even when customised, to totally satisfy all the tests of performance applied to them, it is clear from the calibration curves and the GOF table (Table 5.8), that (Figures 5.1-5.6, Figures 8.1-8.11) rising estimates of mortality are highly correlated with rising observed mortality. If comparisons are to be made between ICUs then the use of the models in this study, rather than crude mortality rates, would be preferable. All the models in this study have impressive discrimination and, after customisation, considerable agreement between observed mortality and the mortality estimated by the model. However, given the considerable effort required to collect these data, it is important that once collected this information should be used, or the data

collection abandoned. At present, although comparisons of mortality ratios are made, as none of the units have ever fallen outside the overall CIs, it has been presumed that differences between units could be explained by normal variance. If these models are only used in this way comparisons of units may only serve to hide poor care rather than help identify poorly performing ICUs. It would be possible to investigate ICUs with the lowest and highest mortality ratios to allow comparisons of the care on these units. This might help identify areas of care which make a positive or negative impact on outcome.

There have only been a few studies which have attempted to use these models to examine differences between ICUs (58,94,140). More research needs to be carried out into differences in ICUs as judged by these severity models. As there is no measurement of quality of care by which to validate severity of illness models, original and innovative ways need to be considered to judge their success. Miranda argues that although case mix adjustment is not totally accurate, the models remain the most accurate way of making comparisons (190).

As well as these direct comparisons of ICUs some independent investigation of possible determinants of outcome could be made. There have been a number of studies in the literature which have tried to identify variables that may have an effect on outcome (13,98,103). There are a number of factors which it would be possible to hypothesise have an effect on the performance of ICUs. The organisation of nursing and medical staff, numbers and skill mix of staff, availability of equipment, nursing and medical training and management structures are all in areas which it may be presumed to have an affect on the effectiveness of care. By profiling units and using severity of illness models to adjust for case mix it may be possible to identify practices which have a positive or negative effect on outcome. This might help identify standards which ICUs could use to judge their current practice against.

One way of investigating these models would be to concentrate on patients whose probabilities lie at the extremes of the severity of illness scales. Concentrating on patients who die but have low probability of mortality might allow researchers to identify problems in the care patients received or where models have failed to adequately adjust for case-mix. Alternatively by concentrating on patients who live but have high probabilities of mortality might help identify where high quality of care has worked. To do this requires carefully designed protocols with clear hypotheses and outcome measures.

A consensus conference suggested that mortality probability models concentrating on a homogenous population were more successful in adjusting for case mix than generic

models in a heterogeneous population (22). Some research has shown success in using a customised version of both SAPS II and MPM_0 in adjusting for case mix in ICU patients with sepsis (141). With the advent of powerful personal computers and the ease with which statistical analysis can be carried out, generating coefficients for the severity of illness models in specific disease categories is a relatively easy process. Rather than making judgements about clinical efficacy in heterogeneous populations, an alternative approach might be to concentrate on specific disease categories. The large numbers of patients in the Scottish Audit database would allow disease specific versions of these models to be generated. This approach might also make it easier to draw conclusions on care in different units without the confusing picture of the normal heterogeneous population.

10.4 Implications of this research

10.4.1 Implications for researchers

The evidence from this study would suggest that the severity of illness models have impressive discrimination and, after customisation, improved calibration. There remains doubt in the ability of these models to fully adjust for case mix. While no model can ever be perfect, there still remains a considerable amount unknown about severity of illness models. Clinicians remain divided on their usefulness and the question of variations within ICUs needs to be analysed to allow these models to be used with confidence.

Results from this study have provided some insight for future researchers to build on. As Teres and Lemeshow have reported, other studies have now shown that there is deterioration of these models when applied outside the cultures they where developed in (137). Studies have also shown the ability of customisation to improve these models (111,134,143). There has been considerable debate and evidence to suggest that the lack of an accurate GCS score in the APACHE models might have an effect on the models accuracy. This study has confirmed this and shown that the use of a pre-sedated GCS can improve the accuracy of probabilities generated by the model. These results have been confirmed by the customisation of the APACHE II model where the APACHE models with the best performance were those which used a pre-sedated GCS. Future researchers need to bear this in mind when developing models. When applying models outside the original ICU culture researchers need to be aware of organisational differences that might effect the performance of the model.

Other research into the customisation process have shown that even after customisation of the SAPS II and MPM II there is considerable problems with the uniformity of fit

(134,143). However, this study has shown that the APACHE II model has good uniformity of fit after customisation. This would suggest the value of the inclusion of a diagnosis element in the model and future researchers should consider this when developing new models.

Further research needs to be carried out to identify other factors significantly associated with mortality, allowing for the identification of variables like deprivation, that may need to be added to the model. Work also needs to be done on identifying aspects of care, management and organisation of ICUs that may contribute to improved outcome.

The effect of variability of data collection and the quality of data needs to be more fully understood. Any further work should help to establish standards for data collection and allow for the assessment of the robustness of the data.

Researchers into modelling of severity of illness must continue to identify weaknesses in the present approach of these models. Teres and Lemeshow have suggested that models should concentrate on the acute phase of illness and on less arbitrary outcome measures. Work needs to also look at the effect treatment in the first 24 hours may have on the accuracy of these models.

There have been a number of studies where these models have been used to adjust for case mix to allow groups of similar severity of illness to be compared. (85,93). However, if these models are to be used in this way then the caveats on their accuracy must be explicit.

10.4.2 Implications for clinicians

There has been limited use of these models in Scotland's ICUs, with most work concentrating on assessing their accuracy and improving their performance. This reluctance to use them is part due to the belief that they are not accurate enough and that differences are due to failures of the models rather than differences in care. However, with customisation there is significant improvement in the models' performance.

If these models are to have a positive impact on patient care then they need to be used. As well as direct comparisons of units there are other ways that these models could be used, including highlighting patients expected to live, examining other factors that may effect mortality, examining care in specific disease categories. These have already been described in 10.3.2.

10.4.3 Implications for managers

Managers and planners have need of good quality information to both manage and plan Intensive Care services. These models have been shown to be useful for assessing resource use and workload (88,96,97). However, managers and planners need to fully understand how to interpret this data and the caveats that apply.

When using these models to assess quality of care and current practice it is important that clinicians within ICUs do this in an atmosphere of co-operation . It is important that the quality of care of patients is measured and shown to be of a good standard. However, it also needs to be acknowledged that to allow the open investigation of their practice requires considerable courage on the part of the clinicians and nursing staff. The collection of the data for these models is done by the ICU staff themselves and unless they feel ownership of these data it is unlikely that the data will be accurate. It is possible to deliberately inflate the scores to give higher probabilities, making an ICUs mortality ratio look good.

While this study has shown that even after customisation there is doubt as to the ability of these models to adjust totally for case mix, they remain the most accurate tool for this purpose.

References

References

1. Scottish Office, *Designed to Care*, HMSO, 1998.

2. Spiby J. Intensive Care in the United Kingdom: Report from the King's Fund Panel. *Anaesthesia* 44:428-431, 1989.

3. Le Gall J-R, Lemeshow S, Saulnier F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. *Journal of the American Medical Association* 270:2957-2962, 1993.

4. Lemeshow S, Teres D, Klar J, Avrunin J, Gehibach H, Rapoport J. Mortality Probability Models(MPM II) Based on an International Cohort of Intensive Care Unit Patients. *Journal of the American Medical Association* 270:2478-2486, 1993.

5. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: A severity of disease classification system. *Critical Care Medicine* 13:818-829, 1985.

6. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III Prognostic System - Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults. *Chest* 100:1619-1636, 1991.

7. Lassen HCA. A preliminary report on the 1952 epidemic of poliomyelitis in Copenhagen with special reference to the treatment of acute respiratory insufficiency. *Lancet* i:37-41, 1952.

8. Bion JF. Rationing intensive care. *British Medical Journal* 310:682-683, 1995.

9. Halpern NA, Bettes L, Greenstein R. Federal and nationwide intensive care units and healthcare costs: 1986-1992. *Critical Care Medicine* 22:2001-2007, 1994.

10. Eddeleston J, Cheetham E, Nightingale P. Clinical audit in intensive care. *British Journal of Intensive Care* January:16-20, 1996.

11. Standards sub-committee recommendations, *Standards for Intensive Care Units*, London:Intensive Care Society (UK), Biomedica Ltd, 1984.

12. Working party recommendations (1990), *The Intensive Care Service in the UK*, Intensive Care Society (UK), 1990.

13. Rowan, K.M. *Outcome comparisons of ICU in Great Britain and Ireland using the APACHE II method*, 1993. (UnPub)

14. Hook EW, Horton CA, Schaberg DR. Failure of intensive care unit support to influence mortality from pneumococcal bacteremia. *Journal of the American Medical Association* 249:1055-1057, 1983.

15. Rogers RM, Weiler C, Ruppenthal B. Impact of the respiratory intensive care unit on survival of patients with acute respiratory failure. *Chest* 62:94-97, 1972.

16. Metcalfe, A. and McPherson, K. *Study of provision of intensive care in England, 1993*, Department of Health, 1995.

17. Secretaries of State For Health, Wales,Northern Ireland and Scotland *Working for Patients (Cmnd)*, London:HMSO, 1989.

18. Bion, J.F. Audit in intensive care. In: *Medical audit: rationale and practicalities*, edited by Frostick, S.P., Radford, P.J. and Wallace, W.A. Cambridge: Cambridge University Press, 1993, p. 307-338.

19. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An Evaluation of Outcome from Intensive Care in Major Medical Centres. *Annals of Internal Medicine* 104:410-418, 1986.

20. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland-II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. *British Medical Journal* 307:977-981, 1993.

21. Castella X, Artigas A, Bion JF, Kari A. A comparison of severity of illness scoring systems for intensive care unit patients: Results of a multicenter, multinational study. *Critical Care Medicine* 23:1327-1332, 1995.

22. Hadorn, D.C., Keeler, E.B., Rogers, W.H. and Brook, R.H. *Assessing the performance of mortality prediction models*, Santa Monica, CA:RAND/UCLA/Harvard Center for Health Care Financing Policy Research, 1993.

23. Spiegelhalter DJ, Knill-Jones R. Statistical and Knowledge-based Approaches to Clinical Decision-support Systems, with an Application in Gastroenterology. *Journal of the Royal Statistical Society* 147:35-77, 1984.

24. Fetter RB, Shin Y, Freeman JL, Averill R, Thompson JD. Case mix definition by diagnosis-related groups. *Medical Care* 18(2 Suppl):1-53, 1980.

25. DesHarnais S, Chesney J, Wroblewski M. The Risk-Adjusted Mortality Index: A New Measure of Hospital Performance. *Medical Care* 26:1129-1145, 1988.

26. Green J, Wintfield N, Sharkey P, Passman LJ. The importance of severity of illness in assessing hospital mortality. *Journal of the American Medical Association* 263:241-246, 1990.

27. Horn S, Sharkey P, Buckle J, et al . The Relationship Between Severity of Illness and Hospital Length of Stay and Mortality. *Medical Care* 29:305-317, 1991.

28. Iezzoni L, Ash A, Coffman G, et al . Predicting In-Hospital Mortality. A Comparison of Severity Measurement Approaches. *Medical Care* 30:1992.

29. Keller EB, Kahn KL, Draper D, et al . Changes in Sickness at Admission Following the Introduction of the Prospective Payment System. *Journal of the American Medical Association* 264:1962-1968, 1990.

30. Committee on Medical Aspects of Automotive Safety (CMAAS) . Rating the severity of tissue damage: I. The abbreviated scale. *Journal of the American Medical Association* 215:277-280, 1971.

31. Committee on Medical Aspects of Automotive Safety (CMAAS) . Rating the severity of tissue damage: II. The comprehensive scale. *Journal of the American Medical Association* 220:717-720, 1972.

32. Baker SP, O'Neill B, Haddon W, Long WB. The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma* 14:187-196, 1974.

33. Moore EE, Dunn EL, Moore JB, Thompson JS. Penetrating Abdominal Trauma Index. *Journal of Trauma* 21:439-445, 1981.

34. Somers RL. The Probability of Death Score: a measure of injury severity for use in planning and evaluating accident prevention. *Accident Analysis and Prevention* 15:259-266, 1983.

35. Kirkpatrick JR, Youmans RL. Trauma Index: an aide in the evaluation of injury victims. *Journal of Trauma* 11:711-714, 1971.

36. Champion HR, Sacco WJ, Hannon DS. Assessment of injury severity: the Triage Index. *Critical Care Medicine* 8:201-208, 1980.

37. Champion HR, Sacco WJ, Carnazzo AJ, Copes W, Fouty WJ. Trauma score. *Critical Care Medicine* 9:672-676, 1981.

38. Gormican SP. CRAMS scale: field triage of trauma victims. *Annals of Emergency Medicine* 11:132-135, 1982.

39. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: The TRISS method. *Journal of Trauma* 27:370-378, 1987.

40. Champion HR, Copes W, Sacco WJ, Lawnick MM, Bain LW, Gann DS, et al. A new characterization of injury severity. *Journal of Trauma* 30:539-545, 1990.

41. Durocher A, Saulnier F, Beauscart R, Dievart F, Bart F, Deturck R. A comparison of three severity score indexes in an evaluation of serious bacterial pneumonia. *Intensive Care Medicine* 14:39-43, 1988.

42. Elebute EA, Stoner HB. The grading of sepsis. *British Journal of Surgery* 70:29-31, 1983.

43. Christensen E, Schlichting P, Fauerholdt L, Gluud C, Andersen PK, Juhl E. Prognostic value of Child-Turcotte criteria in medically treated cirrhosis. *Hepatology* 4:430-435, 1984.

44. Alemi F, Rice J, Hankins R. Predicting In-Hospital Survival of Myocardial Infarction. *Medical Care* 28:762-775, 1990.

45. Teasdale G, Jennett B. Assessent of coma and impaired conciousness. A practical scale. *Lancet* ii:81-84, 1974.

46. Marion DW, Carlier PM. Problems with initial Glasgow Coma Scale assessment caused by prehospital treatment of patients with head injuries: results of a national survey. *Journal of Trauma* 36:89-95, 1994.

47. Jennett B, Teasdale G, Galbraith S, Pickard J, Grant H, Braakman R. Severe head injuries in three countries. *Journal of Neurology, Neurosurgery, and Psychiatry* 40:289-295, 1977.

48. Marion DW. The Glasgow Coma Scale Score: Contemporary Application. *Intensive Care World* 11:101-102, 1995.

49. Knaus WA. Measuring the Glasgow Coma Scale in the Intensive Care Unit: Potentials and Pitfalls. *Intensive Care World* 11:102-103, 1995.

50. Cullen DJ, Civetta JM, Briggs BA, Ferrara LC. Theraputic intervention scoring system: a method for quantitative comparison of patient care. *Critical Care Medicine* 2:57-60, 1974.

51. Keene AR, Cullen DJ. Therapeutic Intervention Scoring System: Update 1983. *Critical Care Medicine* 11:1-3, 1983.

52. Byrick RJ, Mindorff C, McKee L, Mudge B. Cost-effectiveness of intensive care for respiratory failure patients. *Critical Care Medicine* 8:332-337, 1980.

53. Zimmerman JE, Knaus WA, Judson JA, Havill JH, Trubuhovich RV, Draper EA. Patient selection for intensive care: a comparison of New Zealand and United States Hospitals. *Critical Care Medicine* 16:318-326, 1988.

54. Henning RJ, McClish D, Daly B, Nearman H, Franklin C, Jackson D. Clinical characteristics and resource utilisation of ICU patients: Implications for organisation of intensive care. *Critical Care Medicine* 15:264-269, 1987.

55. Cullen DJ, Ferrara LC, Briggs BA, Walker PF, Gilbert J. Survival, hospitalization charges and follow-up results in critically ill patients. *New England Journal of Medicine* 294:982-987, 1976.

56. Knaus WA, Wagner DP, Draper EA, Lawrence DE, Zimmerman JE. The range of intensive care services today. *Journal of the American Medical Association* 246:2711-2716, 1981.

57. Zimmerman JE, Wagner DP, Knaus WA. The use of risk predictions to identify candidates for intermediate care units: Implications for intensive care utilization and cost. *Chest* 108:490, 1995.

58. Girotti MJ, Brown SJL. Factors predicting discharge from intensive care: a Canadian experience. *Canadian Anaesthetists' Society Journal* 33:294-299, 1986.

59. Draper EA, Wagner DP, Knaus WA. The use of intensive care: A comparison of a university and community hospital. *Health Care Financing Review* 3:49-64, 1981.

60. Schwartz S, Cullen DJ. How many intensive care beds does your hospital need? *Critical Care Medicine* 9:625-629, 1981.

61. Turner JS, Potieger PD, Linton DM. Systems for scoring severity of illness in intensive care. *Critical Care Medicine* 76:17-20, 1989.

62. Knaus WA, Zimmerman JE, Wagner DP, Draper EA. APACHE- acute physiology and chronic health evaluation: a physiologically based classification system. *Critical Care Medicine* 9:591-597, 1981.

63. Knaus WA, Draper EA, Wagner DP, Zimmerman JE, Birnbaumm ML, Cullen DJ. Evaluating outcome from intensive care:A preliminary multihospital comparison. *Critical Care Medicine* 10:491-496, 1982.

64. Knaus WA, Le Gall J-R, Wagner DP, Draper EA, Loirat P, Compos RA. A comparison of intensive care in the USA and France. *Lancet* ii:642-646, 1982.

65. Wagner DP, Draper EA, Compos RA, Nikki P, Le Gall J-R, Loirat P. Initial international use of APACHE: An acute severity of disease measure. *Medical Decision Making* 4:297-313, 1984.

66. Wagner DP, Knaus WA, Draper EA. Statistical validation of a severity of illness measure. *American Journal of Public Health* 73:878-884, 1983.

67. Nicolas F, Le Gall J-R, Alperovitch A, Loirat P, Villers D. Influence of patients age on survival, level of therapy and length of stay in intensive care units. *Intensive Care Medicine* 13:9-13, 1987.

68. Meade MO, Cook DJ. A critical appraisal and systematic review of illness severity scoring systems in the intensive care unit. *Current Opinion in Critical Care* 1:221-227, 1995.

69. Dellinger EP, Wertz MJ, Meakins JL, Solomkins JS, Allo MD, Howard RJ, et al. Surgical infection stratification for intra-abdominal infection. *Archives of Surgery* 120:21-29, 1985.

70. Chang RWS, Jacobs S, Lee B. Use of APACHE II severity of disease classifications to identify intensive care unit patients who would not benefit from total parenteral nutrition. *Lancet* i:1483-1486, 1986.

71. Castella X, Gilabert J, Torner F, Torres C. Mortality prediction models in intensive care: Acute Physiology and Chronic Health Evaluation II and mortality prediction model compared. *Critical Care Medicine* 19:191-197, 1991.

72. Abbott RR, Setter M, Chan S, Choi K. APACHE II: Prediction of outcome of 451 ICU oncology admissions in community hospital. *Annals of Oncology* 2:571-574, 1991.

73. Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland-I: Variations in case mix of adult admissions to general intensive care units and impact on outcome. *British Medical Journal* 307:972-976, 1993.

74. Moreau R, Soupison T, Vauquelin P, Derrida S, Beaucour H, Sicot C. Comparison of two simplified severity scores (SAPS and APACHE II) for patients with acute myocardial infarction. *Critical Care Medicine* 17:409-413, 1989.

75. Fedullo AJ, Swinburne AJ, Wahl GW, Bixby KR. APACHE II score and mortality in respiratory failure due to cardiogenic pulmonary oedema. *Critical Care Medicine* 16:1218-1221, 1988.

76. Horst HM, Obeid FN, Sorensen VJ, Bivins BA. Factors influencing survival of elderly trauma patients. *Critical Care Medicine* 14:681-684, 1986.

77. Cerra FB, Negro F, Jerome A. APACHE II score does not predict multiple organ failure or mortality in postoperative surgical patients. *Archives of Surgery* 125:519-522, 1999.

78. Hopefl AW, Taaffee CL, Herrmann VM. Failure of APACHE II alone as a predictor of mortality in patients receiving total parenteral nutrition. *Critical Care Medicine* 17:414-417, 1989.

79. Rutledge R, Fakhry SM, Rutherford EJ, Muakkassa F, Baker CC, Koruda M, et al. Acute Physiology and Chronic Health Evaluation (APACHE II) score and outcome in the surgical intensive care unit: An analysis of multiple intervention and outcome variables in 1,238 patients. *Critical Care Medicine* 19:1048-1053, 1991.

80. Rhee KJ, Baxt WG, MacKenzie JR, Willits NH, Burney RE, O'Malley RJ, et al. APACHE II scoring in the injured patient. *Critical Care Medicine* 18:827-830, 1990.

81. Bion JF, Aitchison TC, Edlin SA, Ledingham IMcA. Sickness scoring and response to treatment as predictors of outcome from critical illness. *Intensive Care Medicine* 14:167-172, 1988.

82. Turner JS, Mudaliar YM, Chang RWS, Morgan CJ. Acute Physiology and Chronic Health Evaluation (APACHE II) scoring in a cardiothoracic intensive care unit. *Critical Care Medicine* 19:1266-1269, 1991.

83. Beck DH, Taylor BL, Millar BW, Smith GB. Prediction of outcome from intensive care: A prospective cohort study comparing Acute Physiology and Chronic Health Evaluation II and III prognostic systems in a United Kingdom intensive care unit. *Critical Care Medicine* 25:9-15, 1997.

84. Pappachan JV, Millar B, Bennett ED, Smith GB. Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. *Chest* 115:802-810, 1999.

85. Solomkin JS, Fant WK, Rivera JO, Alexander JW. Randomized trial of imipenem/cilastatin versus gentamicin and clindamycin in mixed flora infections. *American Journal of Medicine* 78:85-91, 1985.

86. Goldhill DR, Sumner A. Outcome of intensive care patients in a group of British intensive care units. *Critical Care Medicine* 26:1337-1345, 1998.

87. Pittet JF, Morel DR, Hemsen A, Gunning K, Lacroix JS, Suter PM, et al. Elevated plasma endothelin-1 concentrations are associated with the severity of illness in patients with sepsis. *Annals of Surgery* 213:261-264, 1991.

88. Wagner DP, Draper EA. Acute Physiology and Chronic health Evaluation (APACHE II) and medicare reimbursement. *Health Care Financing Review* 6:91-105, 1984.

89. Chang RWS, Jacobs S, Lee B. Predicting outcome among intensive care unit patients using computerised trend analysis of daily APACHE II scores corrected for organ failure. *Intensive Care Medicine* 14:558-566, 1988.

90. Rogers J, Fuller HD. Use of daily Acute Physiology and Chronic Health Evaluation (APACHE) II scores to predict individual patient survival rate. *Critical Care Medicine* 22:1402-1405, 1994.

91. Borlase BC, Baxter JT, Benotti PN, Stone M, Wood E, Forse RA, et al. Surgical intensive care unit resource use in a speciality referral hospital: I. Predictors of early death and cost implications. *Surgery* 109:687-693, 1991.

92. Wagner DP, Knaus WA, Harrell FR, Zimmerman JE, Watts c. Daily prognostic estimates for critically ill adults in intensive care units: Results from a prospective, multicenter, inception cohort analysis. *Critical Care Medicine* 22:1359-1372, 1994.

93. Connors Jr AF, Speroff T, Dawson NV, Thomas C, Harrell Jr FE, Wagner DP, et al. The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients. *Journal of the American Medical Association* 276:889-897, 1996.

94. Zimmerman JE, Shortell SM, Rousseau DM, Duffy J, Gillies RR, Knaus WA, et al. Improving intensive care: Observations based on organizational case studies in nine intensive care units: A prospective, multicenter study. *Critical Care Medicine* 21:1443-1451, 1993.

95. Zimmerman JE, Wagner DP, Draper EA, Knaus WA. Improving intensive care unit discharge decisions: Supplementing physician judgement with predictions of next day risk for life support. *Critical Care Medicine* 22:1373-1384, 1994.

96. Henderson A, Cleary M, Galbraith G, Hurlford R. Prospective study of costs and outcome in a major adult Australian intensive care unit utilising the APACHE III severity scoring and prediction tool. *Clinical Intensive Care* 8:58-62, 1997.

97. Becker RB, Zimmerman JE, Knaus WA, Wagner DP, Seneff MG, Draper EA, et al. The use of APACHE III to evaluate ICU length of stay, resource use, and mortality after coronary artery by-pass surgery. *Journal of Cardiovascular Surgery* 36:1-11, 1995.

98. Knaus WA, Wagner DP, Zimmerman JE. Variations in mortality and length of stay in intensive care units. *Annals of Internal Medicine* 118:1993.

99. Zimmerman JE, Shortell SM, Knaus WA. Value and cost of teaching hospitals: A prospective, multicenter, inception cohort study. *Critical Care Medicine* 21:1432, 1993.

100. Hamahata N, Nagino M, Nimura Y. APACHE III, unlike APACHE II, predicts postepatectomy mortality in patients with biliary tract carcinoma. *Critical Care Medicine* 26:1671-1676, 1998.

101. Von Bierbrauer A, Riedel S, Cassel W, Von Wichert P. Validation of APACHE III and comparison to APACHE II in a German intensive care unit. . *Anaesthetist* 47:30-38, 1998.

102. Bastos PG, Sun X, Wagner DP, Knaus WA, Zimmerman JE. Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. *Intensive Care Medicine* 22:564-570, 1996.

103. Bastos PG, Knaus WA, Zimmerman JE, Magalhaes A, Sun X, Wagner DP. The importance of technology for achieving superior outcomes from intensive care. *Intensive Care Medicine* 22:664-669, 1996.

104. Pappachan JV, Millar BW, Bennett ED, Smith GB. Outcome comparisons after case mix adjustment using the APACHE (III) (A3) system in 17 UK ICUs. *Intensive Care Society* 1997.(Abstract)

105. Pappachan JV, Millar BW, Bennett ED, Smith GB. A Study of the APACHE III (A3) prognostic system in 17 UK intensive care units- casemix and outcome. *Intensive Care Society* 1997.(Abstract)

106. Le Gall J-R, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. *Critical Care Medicine* 12:975-977, 1984.

107. Van Lanschot JJB, Feenstra BWA, Vermeij CG, Bruining HA. Outcome prediction in critically ill patients by means of oxygen consumption index and simplified acute physiology score. *Intensive Care Medicine* 14:44-49, 1988.

108. French Multicentric Group of ICU Research and the INSERM Unit 169 of Statistical and Epidemiological Studies . Factors related to outcome in intensive care: French multicenter study. *Critical Care Medicine* 17:305-308, 1989.

109. Sicignano A, Carozzi C, Giudici D, Merli G, Arlati S, Pulici M. The Influence of length of stay in the ICU on the power of discrimination of a multipurpose severity score (SAPS). *Intensive Care Medicine* 22:1048-1051, 1996.

110. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: Results from GiViTI. *Intensive Care Medicine* 22:1368-1378, 1996.

111. Collins A. MPM II and SAPS II: a review of easy-to-use severity systems. *Care of the Critically ill* 11:73-76, 1995.

112. Markgraf R, Destschinoff G, Pientka L, Scholten T. A comparison of APACHE III and SAPS II on outcome of release from hospital and six months after admission to the ICU. *International Journal of Clinical Monitoring and Computing* 14:64-65, 1997.

113. Reina A, Vazquez G, Aguayo E, Bravo I, Colmenero M, Bravo M. Mortality discrimination in acute myocardial infarction: Comparison between APACHE III and SAPS II prognosis systems. *Intensive Care Medicine* 23:326-330, 1997.

114. Moreno R, Miranda DR, Fidler V, Schilfgaarde RV. Evaluation of two outcome prediction models on an independent database. *Critical Care Medicine* 26:50-61, 1998.

115. Moreno R, Apolone G, Miranda DR. Evaluation of the uniformity of fit of general outcome prediction models. *Intensive Care Medicine* 24:40-47, 1998.

116. Brun-Buisson C, Doyon F, Carlet J, Dellamonica P, Gouin F, Lepoutre A, et al. Incidence, risk factors, and outcome of severe sepsis and septic shock in adults. A multicenter prospective study in intensive care units. French ICU Group for Severe Sepsis. *Journal of the American Medical Association* 274:968-974, 1995.

117. Cunha J, Povao P, Mourao L, Santos AL, Luis AS. Severe poisoning by organophosphate compounds. An analysis of mortality and of the value of serum cholinesterase in monitoring the clinical course. *Acta Med Port* 8:469-475, 1995.

118. Lemeshow S, Teres D, Pastides H, Avrunin J, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. *Critical Care Medicine* 13:519-525, 1985.

119. Teres D, Lemeshow S, Avrunin J, Pastides H. Validation of the mortality prediction model for ICU patients. *Critical Care Medicine* 15:208-213, 1987.

120. Lemeshow S, Teres D, Avrunin J, Gage RW. Refining intensive care unit outcome prediction by using changing probabilities of mortality. *Critical Care Medicine* 16:470-477, 1988.

121. Schafer JH, Maurer A, Jochimsen F, Emde C, Wegscheider K, Arntz HR, et al. Outcome prediction models on admission in a medical intensive care unit: Do they predict individual outcome? *Critical Care Medicine* 18:1111-1118, 1990.

122. Lemeshow S, Teres D, Klar J, Avrunin J, Gehibach H, Rapoport J. Mortality probability models for patients in the intensive care unit for 48 or 72 hours: A prospective multicenter study. *Critical Care Medicine* 22 No. 9:1351-1358, 1994.

123. Rello J, Rue M, Jubert P, Muses G, Sonora R, Valles J, et al. Survival in patients with nosocomial pnumonia: Impact of the severity of illness and the etiologic agent. *Critical Care Medicine* 25:1862-1867, 1997.

124. Murphy-Filkins RL, Teres D, Lemeshow S, Hosmer DW. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: How to distinguish a general from a speciality intensive care unit. *Critical Care Medicine* 24:1968-1973, 1996.

125. Rapoport J, Teres D, Lemeshow S, Gehibach H. A method for assessing the clinical performance and cost-effectiveness of intensive care units: A multicenter inception cohort study. *Critical Care Medicine* 22 No. 9:1385-1390, 1994.

126. Rapoport J, Teres D, Lemeshow S. Resource use implications of do not resuscitate orders for intensive care unit patients. *American Journal of Respiratory and Critical Care Medicine* 153:185-190, 1996.

127. Boyd K, Teres D, Rapoport J, Lemeshow S. The relationship between age and the use of DNR Orders in Critical Care Patients. *Archives of Internal Medicine* 156:1821-1826, 1996.

128. Teres D. Trends from the United States with end of life decisions in the intensive care unit. *Intensive Care Medicine* 19:316-322, 1993.

129. Moreno, R. Performance of the ICU: Are we able to measure it? In: *Yearbook of Intensive Care Medicine*, edited by Vincent, J.L. Berlin: Springer-Verlag, 1998, p. 729-743.

130. Cho DY, Wang YC. Comparison of the APACHE III, APACHE II and Glasgow Coma Scale in acute head injury for prediction of mortality and functional outcome. *Intensive Care Medicine* Vol 23 , 1997.:77-84, 1997.

131. Alvarez M, Nava J-M, Rue M, Quintana S. Mortality prediction in head trauma patients: Performance of Glasgow Coma Score and general severity systems. *Critical Care Medicine* 26:142-148, 1998.

132. Presterl E, Staudinger T, Pettermann M, Lassnigg A, Burgmann H, Winkler S, et al. Cytokine profile and correlation to the APACHE II and MPM II scores in patients with sepsis. *American Journal of Respiratory and Critical Care Medicine* 156:825-832, 1997.

133. Sarmiento X, Rue M, Guardiola JJ, Toboso JM, Soler M, Artigas A. Assessment of the prognosis of coronary patients: Performance and customization of the generic severity index. *Chest* 111:1666-1671, 1997.

134. Nava S. Scoring of severity in patients admitted to a respiratory intensive care unit. *Monaldi Archives for Chest Disease* 52:71-72, 1997.

135. El-Solh AA, Grant BJB. A comparison of severity of illness scoring systems for critically ill obstetric patients. *Chest* 110:1299-1304, 1996.

136. Bein T, Frohlich D, Frey A, Metz C, Taeger K. Comparison of two severity of disease classification systems (APACHE II and APACHE III). *Anaesthetist* 44:37-42, 1995.

137. Teres D, Lemeshow S. As American as apple pie and APACHE. *Critical Care Medicine* 26:1297-1298, 1998.

138. Nouira S, Belghith M, Elatrous S. Predictive value of severity scoring systems: Comparison multicenter study of four models in Tunisian adult ICUs. *Critical Care Medicine* 26:852-859, 1998.

139. Moreno R, Apolone G. The impact of different customization strategies in the performance of a general severity score. *Critical Care Medicine* 25:2001-2008, 1997.

140. Fidler, V. Analysis of ICU Performance. In: *1998 Yearbook of Intensive Care and Emergency Medicine*, edited by Vincent, J.L. Berlin: Springer-Verlag, 1998, p. 261-269.

141. Le Gall J-R, Lemeshow S, Leleu G, Klar J, Huillard J, Rue M, et al. Customized probability models for early severe sepsis in adult intensive care patients. *Journal of the American Medical Association* 273:644-650, 1995.

142. Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: A simulation study. *Critical Care Medicine* 24:57-63, 1996.

143. Rivera-Fernandez R, Vazquez-Mata G, Bravo M, Aguayo-Hoyos E, Zimmerman JE, Wagner DP, et al. The APACHE III prognostic system:customized mortality predictions for Spanish patients. *Intensive Care Medicine* 24:574-581, 1998.

144. Lemeshow S, Klar J, Teres D. Outcome prediction for individual intensive care patients: useful, misused, or abused? *Intensive Care Medicine* 21:770-776, 1995.

145. Kollef MH, Schuster DP. Predicting intensive care unit outcome with scoring systems. *Critical Care Clinics* 10:1-18, 1994.

146. Sherck JP, Shatney CH. ICU scoring systems do not allow prediction of patient outcomes or comparison of ICU performance. *Critical Care Clinics* 12:515-523, 1996.

147. Boyd O, Grounds RM. Physiological scoring systems and audit. *Lancet* 19:1573-1574, 1993.

148. Mackenzie TA, Greenaway-Coates A, Djurfeldt MS, Hopman WM. Use of severity of illness to evaluate quality of care. *International Journal for Quality of Care* 8:125-130, 1996.

149. O'Connor GT, Plume SK, Olmstead EM, Coffin LH, Morton JR, Maloney CT, et al. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. The Northern New England Cardiovascular Disease Study Group. *Journal of the American Medical Association* 267:932-933, 1991.

150. Teres D, Lemeshow S. Why Severity Models Should Be Used With Caution. *Critical Care Clinics* 10:93-109, 1994.

151. Becker RB, Zimmerman JE. ICU scoring systems allow prediction of patient outcomes and comparison of ICU performance. *Critical Care Clinics* 12:503-514, 1996.

152. Rowan KM. The reliability of case mix measurement in intensive care. *Current Opinion in Critical Care* 2:209-213, 1996.

153. Clinical Outcomes Working Group (1994), *Clinical Outcome Indicators Report, December 1994*, Clinical Resource and Audit Group, The Scottish Office, NHS in Scotland, 1994.

154. Clinical Outcomes Working Group (1995), *Clinical Outcome Measures Report, December 1995*, Clinical Resource and Audit Group. The Scottish Office, NHS in Scotland, 1995.

155. Clinical Outcomes Working Group (1996), *Clinical Outcomes Measures Report, July 1996*, Clinical Resource and Audit Group. The Scottish Office, NHS in Scotland, 1996.

156. Clinical Outcomes Working Group (1998), *Clinical Outcome Measures Report, March 1998*, Clinical Resource and Audit Group. The Scottish Office, NHS in Scotland, 1998.

157. Parry GJ, Gould CR, McCabe CJ, Tarnow-Mordi WO. Annual league tables of mortality in neonatal intensive care units: longitudinal study. *British Medical Journal* 316:1931-1935, 1998.

158. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal* 316:1701-1704, 1998.

159. Kendrick S, Cline D, Finlayson A. Clinical Outcome Indicators in Scotland: Lessons and prospects. *Third International Conference on Strategic Issues in Health Care Management* 1998.(Abstract)

160. Goldstein H, Spiegelhalter DJ. League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society* 159:1-25, 1997.

161. Leyland AH, Pritchard CW, McLoone P, Boddy FA. Measures of performance in Scottish maternity hospitals. *British Medical Journal* 303:389-393, 1991.

162. Livingston, B.M. *Design, Implementation, and Evaluation of Training for an Intensive Care Computer System*, 1994. (UnPub)

163. Hosmer, D.W. and Lemeshow, S. Assessing fit. In: *Applied Logistic Regression*, edited by Hosmer, D.W. and Lemeshow, S. New York: John Wiley & Sons, 1998, p. 140-145.

164. Hanley JA, McNeil BJ. The Meaning and use of the area under a Receiver Operating Characteristics (ROC) curve. *Radiology* 143:29-36, 1982.

165. Detrano R. Accuracy curves: an alternative graphical representation of probability data. *Journal of Clinical Epidemiology* 42:983-986, 1989.

166. Katz D, Foxman B. How well do prediction equations predict? Using receiver operating characteristic curves and accuracy curves to compare validity and generalizability. *Epidemiology* 4:319-326, 1993.

167. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44:837-845, 1988.

168. Hosmer, D.W. and Lemeshow, S. Assessing fit. In: *Applied Logistic Regression*, edited by Hosmer, D.W. and Lemeshow, S. John Wiley & Sons, 1998, p. 146-147.

169. Lemeshow S, Hosmer DW. A Review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology* 115:92-106, 1982.

170. Gardiner, M.J. and Altman, D.G. *Statistics with Confidence*, London:BMJ, 1989. Ed. 7 pp. 28-29.

171. Rothman, K.S. *Modern Epidemiology*, Boston:Little Brown & Co, 1986. pp. 45-49.

172. Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine* 14:2161-2172, 1995.

173. Altman, D.G. *Practical Statistics for Medical Research*, London:Chapman & Hall, 1994. Ed. 3 pp. 243-244.

174. Altman, D.G. *Practical Statistics for Medical Research*, London:Chapman & Hall, 1994. Ed. 3 pp. 184-185.

175. Altman, D.G. *Practical Statistics for Medical Research*, London:Chapman & Hall, 1994. Ed. 3 pp. 397-400.

176. Holt AW, Bury LK, Bersten AD, Skowronski GA, Vedig AE. Prospective evaluation of residents and nurses as severity score data collectors. *Critical Care Medicine* 20:1688-1691, 1992.

177. Bosman RJ, Oudemans van Straaten HM, Zandstra DF. The use of intensive care information systems alters outcome prediction. *Intensive Care Medicine* 24:953-958, 1998.

178. Civetta, J.M. The clinical limitations of ICU scoring systems. In: *Problems in Critical Care*, edited by Farmer, J.C., Kirby, R.R. and Taylor, R.W. Philadelphia: JB Lippincott, 1989, p. 140-145.

179. McLaren, G. and Bain, M. *Deprivation and health in Scotland: Insights from NHS data*, Edinburgh:Information & Statistics Division, National Health Service in Scotland, 1998.

180. Smith GD, Hart C, Watt G, Hole D, Hawthorne V. Individual social class, area-based deprivation, cardiovascular disease risk factors, and mortality: the Renfrew and Paisley Study. *Journal of Epidemiology Community Health* 52:399-405, 1998.

181. Burns, H. *The effects of social inequalities on the outcome of surgery*, 1990. (UnPub)

182. Bastos PG, Xiaolu S, Wagner DP, Knaus WA. Glasgow coma scale in the evaluation of outcome in the intensive care unit: Findings from the Acute Physiology and Chronic Health Evaluation III study. *Critical Care Medicine* 21:1459-1465, 1993.

183. Teres D, Brown RB, Lemeshow S. Predicting mortality of intensive care patients: the importance of coma. *Critical Care Medicine* 10:86-95, 1982.

184. Dragsted L, Jorgensen J, Jenson N-H, Bonsing E, Jacobsen E, Knaus WA, et al. Interhospital comparisons of patient outcome from intensive care: Importance of lead time bias. *Critical Care Medicine* 17:418-422, 1989.

185. Kendrick S, Clarke J. The Scottish Record Linkage System. *Health Bulletin* 51:72-79, 1993.

186. Hanley JA, McNeil BJ. A Method of Comparing the Areas under Receiver Operating Characteristics Curves Derived from the Same Cases. *Radiology* 148:839-843, 1983.

187. Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA. Evaluation of Acute Physiology and Chronic Health Evaluation III predictions of hospital mortality in an independent database. *Critical Care Medicine* 26:1317-1326, 1998.

188. Teres D, Lemeshow S. When to customize a severity model. *Intensive Care Medicine* 25:140-142, 1999.

189. Shoemaker WC, Appel PL, Kram HB, Waxman K, Lee TS. Prospective trial of supranormal values of survivors as therapeutic goals in high-risk surgical patients. *Chest* 94:1176-1186, 1988.

190. Miranda DR. Scoring systems in the measurement of performance of ICUs. *Intensive Care Medicine* 25:418-419, 1999.

Appendix 1

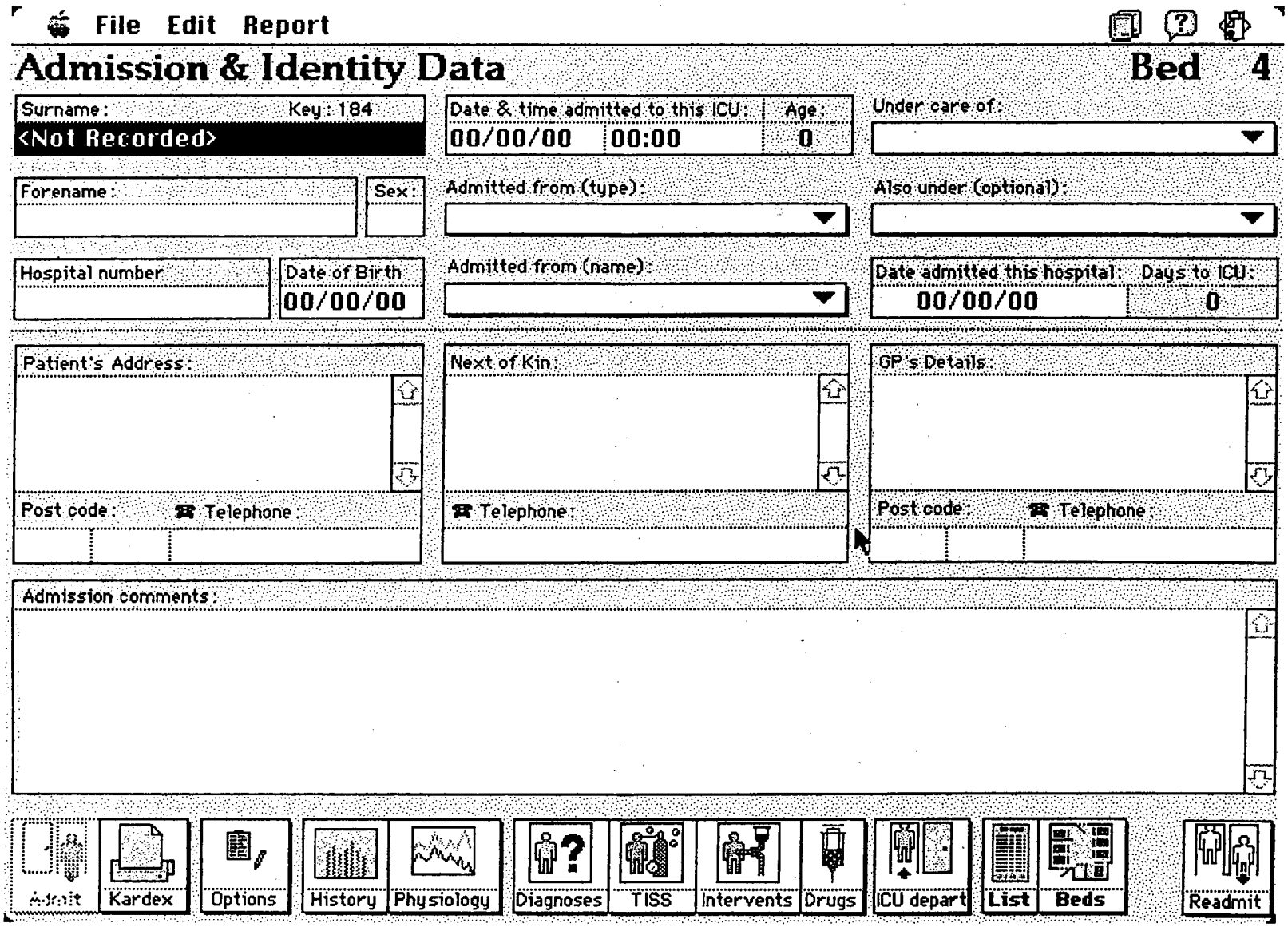
Participating hospitals and co-ordinating consultant

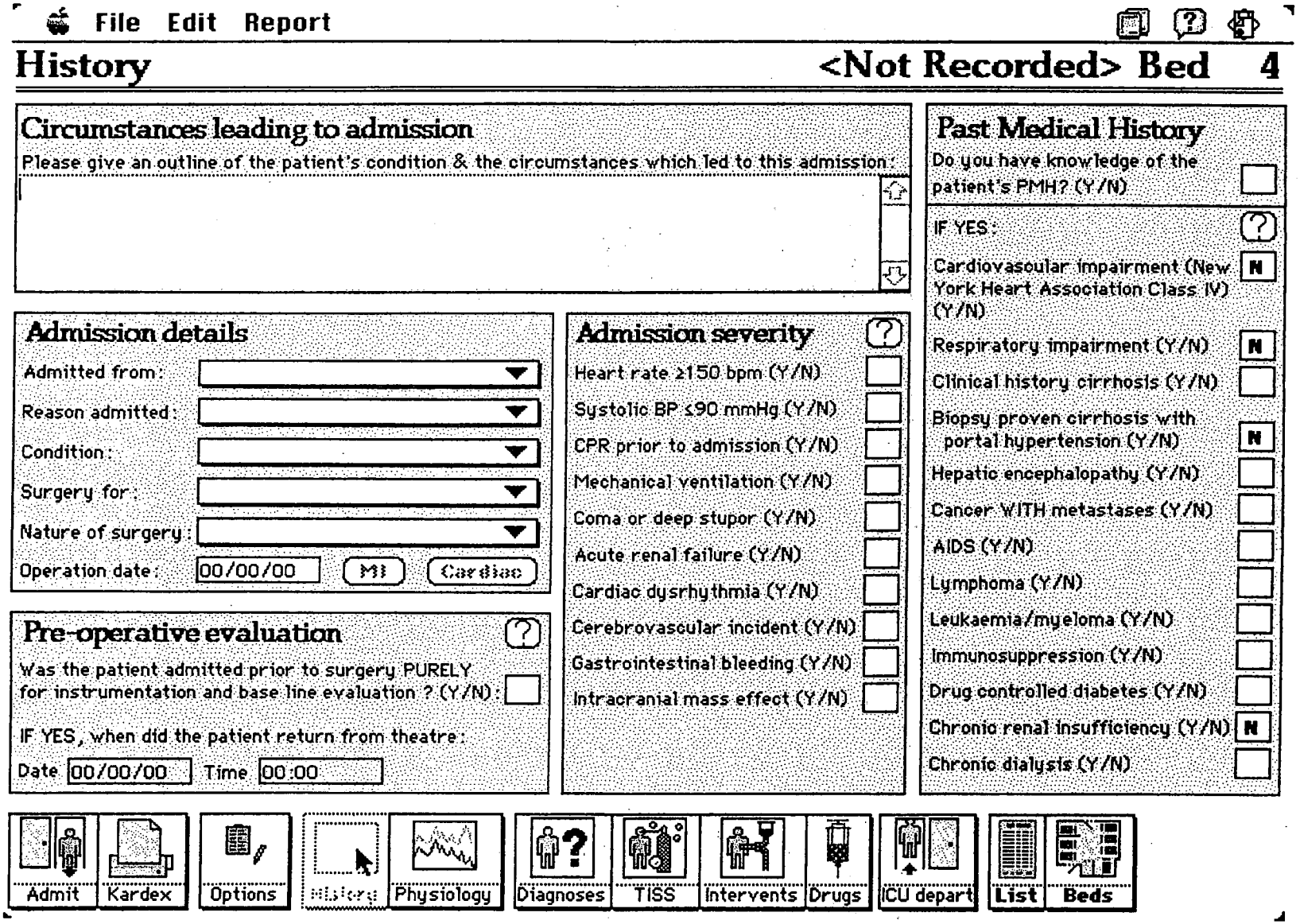
List of participating hospitals and co-ordinating consultant

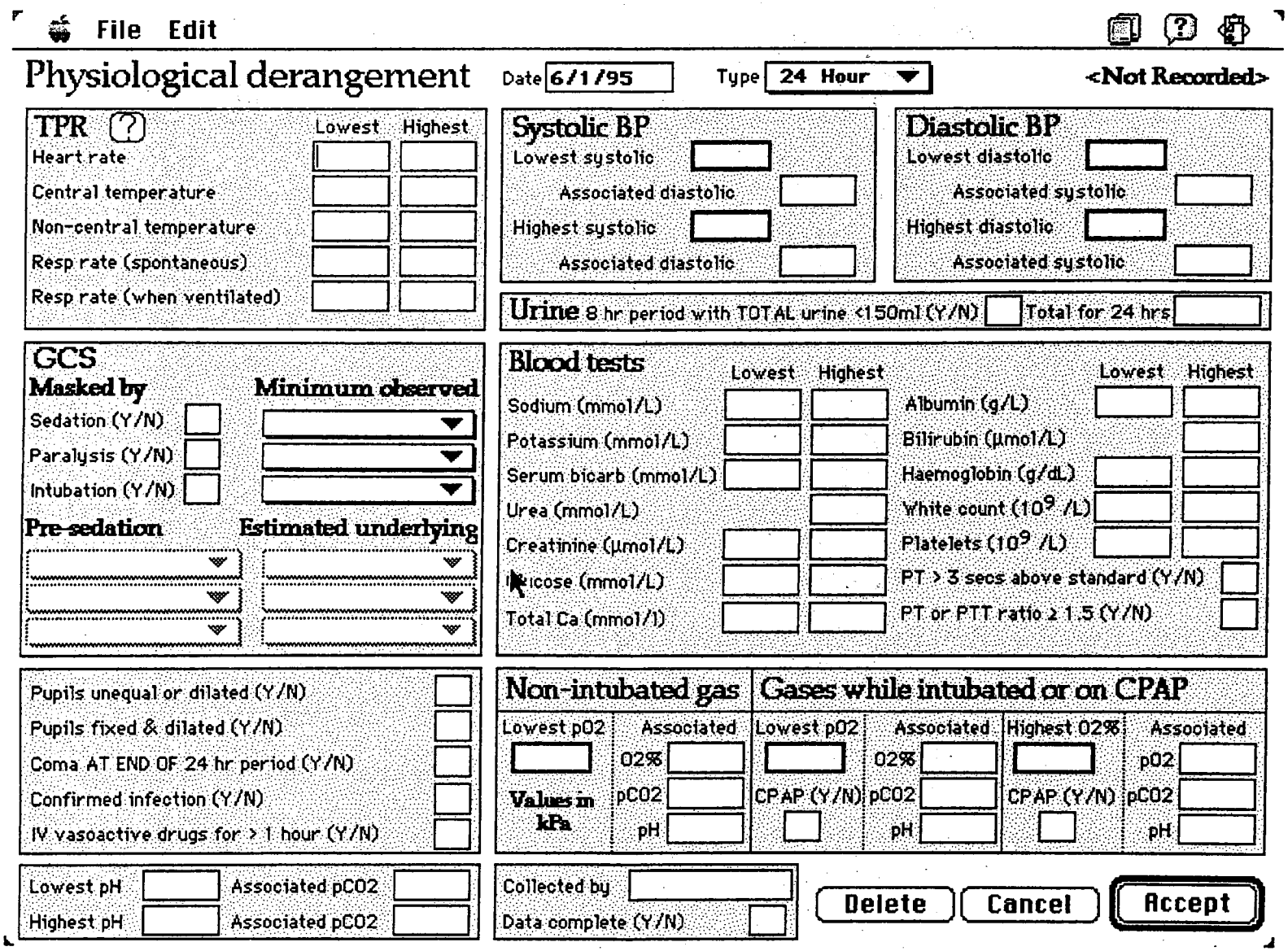
Participating Hospital	Co-ordinating consultant
Aberdeen Royal Infirmary	Dr M.S.P. MacNab
Borders General Hospital	Dr N. Leary;
Crosshouse Hospital	Dr R. White
Edinburgh Royal Infirmary	Dr S. Mackenzie
Falkirk and District Royal Infirmary	Dr D. Simpson
Inverclyde Royal Hospital	Dr T. Winning
Law Hospital	Dr D. MacLean
Monklands District Hospital	Dr M. Inglis
Ninewells Hospital	Dr A.J. Shearer
Perth Royal Infirmary	Dr F.D. Magahy
Queen Margaret Hospital	Dr P. Nicholas
Raigmore Hospital	Dr I. Skipsey
Royal Alexandra Hospital	Dr S. Madsen
Southern General Hospital	Dr J.C. MacDonald
St John's Hospital at Howden	Dr P. Armstrong
Stirling Royal Infirmary	Dr M. Worsley
Stobhill Hospital	Dr C. Miller
Vale of Leven District General Hospital	Dr W. Easy
Victoria Hospital	Dr A. Mowbray
Victoria Infirmary	Dr J.C. Howie
Western General Hospital	Dr I. Grant
Western Infirmary	Dr L. Plenderleith

Appendix 2

Main Ward Watcher data collection screens.


File Edit Report
Admission & Identity Data
Bed 4
Surname: Key: 184
<Not Recorded>
Date & time admitted to this ICU: Age:
00/00/00 00:00 0
Under care of:
Forename:
Sex:
Admitted from (type):
Also under (optional):
Hospital number
Date of Birth
00/00/00
Admitted from (name):
Date admitted this hospital: Days to ICU:
00/00/00 0
Patient's Address:
Next of Kin:
GP's Details:
Post code: Telephone:
Telephone:
Post code: Telephone:
Admission comments:
Kardex
Options
History
Physiology
Diagnoses
TISS
Intervents
Drugs
ICU depart
List
Beds
Readmit


File Edit Report
History
<Not Recorded> Bed 4
Circumstances leading to admission
Please give an outline of the patient's condition & the circumstances which led to this admission:
Admission details
Admitted from:
Reason admitted:
Condition:
Surgery for:
Nature of surgery:
Operation date: 00/00/00
MI
Cardiac
Pre-operative evaluation
Was the patient admitted prior to surgery PURELY for instrumentation and base line evaluation ? (Y/N):
IF YES, when did the patient return from theatre:
Date 00/00/00 Time 00:00
Admission severity
Heart rate ≥150 bpm (Y/N)
Systolic BP ≤90 mmHg (Y/N)
CPR prior to admission (Y/N)
Mechanical ventilation (Y/N)
Coma or deep stupor (Y/N)
Acute renal failure (Y/N)
Cardiac dysrhythmia (Y/N)
Cerebrovascular incident (Y/N)
Gastrointestinal bleeding (Y/N)
Intracranial mass effect (Y/N)
Past Medical History
Do you have knowledge of the patient's PMH? (Y/N)
IF YES:
Cardiovasoular impairment (New York Heart Association Class IV) (Y/N)
N
Respiratory impairment (Y/N)
N
Clinical history cirrhosis (Y/N)
Biopsy proven cirrhosis with portal hypertension (Y/N)
N
Hepatic encephalopathy (Y/N)
Cancer WITH metastases (Y/N)
AIDS (Y/N)
Lymphoma (Y/N)
Leukaemia/myeloma (Y/N)
Immunosuppression (Y/N)
Drug controlled diabetes (Y/N)
Chronic renal insufficiency (Y/N)
N
Chronic dialysis (Y/N)
Admit
Kardex
Options
Physiology
Diagnoses
TISS
Intervents
Drugs
ICU depart
List
Beds


File Edit
Physiological derangement
Date 6/1/95
Type 24 Hour
<Not Recorded>
TPR
Lowest Highest
Heart rate
Central temperature
Non-central temperature
Resp rate (spontaneous)
Resp rate (when ventilated)
Systolic BP
Lowest systolic
Associated diastolic
Highest systolic
Associated diastolic
Diastolic BP
Lowest diastolic
Associated systolic
Highest diastolic
Associated systolic
Urine 8 hr period with TOTAL urine <150ml (Y/N)
Total for 24 hrs
GCS
Masked by
Minimum observed
Sedation (Y/N)
Paralysis (Y/N)
Intubation (Y/N)
Pre-sedation
Estimated underlying
Blood tests
Lowest Highest
Sodium (mmol/L)
Potassium (mmol/L)
Serum bicarb (mmol/L)
Urea (mmol/L)
Creatinine (µmol/L)
Glucose (mmol/L)
Total Ca (mmol/l)
Albumin (g/L)
Bilirubin (µmol/L)
Haemoglobin (g/dL)
White count (10^9 /L)
Platelets (10^9 /L)
PT > 3 secs above standard (Y/N)
PT or PTT ratio ≥ 1.5 (Y/N)
Pupils unequal or dilated (Y/N)
Pupils fixed & dilated (Y/N)
Coma AT END OF 24 hr period (Y/N)
Confirmed infection (Y/N)
IV vasoactive drugs for > 1 hour (Y/N)
Non-intubated gas
Gases while intubated or on CPAP
Lowest pO2
Associated
O2%
pCO2
pH
Values in kPa
CPAP (Y/N)
Highest O2%
Associated
pO2
Lowest pH
Associated pCO2
Highest pH
Associated pCO2
Collected by
Data complete (Y/N)
Delete
Cancel
Accept

Appendix 3

APACHE III diagnostic groups and APACHE III to APACHE II Mapping.

Condition and operation codes

Respiratory

Operative

1 Neoplasm-mouth/sinuses
2 Neoplasm-larynx/trachea
3 Neoplasm-lung parenchyma
4 Broncho-pleural fistula
5 All other pleural disease
6 Respiratory infection/abscess
7 Tracheal-oesophageal fistula
8 Other respiratory surgery

Non-operative

9 ARDS (non cardiogenic pulmonary oedema)
10 Pneumonia-viral
11 Pneumonia-parasitic
12 Pneumonia-bacterial
13 Pneumonia-fungal
14 Pneumonia-aspiration/toxic
15 Neoplasm-mouth/sinuses
16 Neoplasm-larynx/trachea
17 Neoplasm-lung parenchyma
18 Pulmonary embolus
19 Pulmonary hypertension (primary)
20 Localised airway obstruction/oedema (mechanical)
21 Emphysema
22 Asthma
23 Smoke inhalation
24 Cystic fibrosis
230 Respiratory arrest
25 Other respiratory disorder

Cardiovascular

Operative

26 Carotid endarterectomy
27 Aorto-femoral, fem-fem bypass graft
28 Fem-popliteal bypass graft
29 Aortic aneurysm: pre-leak/dissection
30 Aortic aneurysm: dissection
31 Aortic aneurysm: rupture
32 Peripheral ischaemia
33 Gangrenous extremity
34 Cellulitis
35 Septic shock - lungs (pneumonia)
36 Septic shock - urinary tract infection
37 Septic shock - gastrointestinal tract
38 Septic shock - unknown origin
39 Pericardial effusion
40 Valvular repair/replacement
41 Valvular repair/replacement with CABG
42 Coronary artery bypass graft(s)
43 Heart transplant ± lungs
44 Fibrosarcoma (leg, shoulder)
45 Atrial myxoma
46 Congenital anomaly
47 Ventricular aneurysm
48 Automatic implantable cardiac defibrillator
49 Vena cava clipping
50 Vena cava filter
51 Other cardiovascular surgery

Non-operative

52 Carotid disease/TIAs
53 Aortic aneurysm
54 Peripheral ischaemia
55 Hypertension
56 Unstable angina
57 Rhythm disturbance
58 Acute myocardial infarction
59 Rule out MI
60 Congestive heart failure
61 Cardiogenic shock
62 Gangrenous extremity
63 Cellulitis
64 Septic shock - lungs (pneumonia)
65 Septic shock - urinary tract infection
66 Septic shock - gastrointestinal tract
67 Septic shock - unknown origin
68 Pericardial tamponade
69 Post cardiac arrest (± respiratory arrest)
70 Heart transplant rejection

71 Cardiomyopathy
72 Anaphylaxis
73 Other cardiovascular disorder

Neurological

Operative

74 Subarachnoid haemorrhage/intracranial aneurysm
75 Subdural/epidural haematoma
76 Intracerebral haemorrhage/haematoma
77 Craniotomy for neoplasm
78 Transphenoidal removal of neoplasm
79 Neurological abscess
80 Encephalitis/inflammation
81 Spinal cord surgery
82 Other neurosurgery

Non-operative

83 Subarachnoid haemorrhage/intracranial aneurysm
84 Subdural/epidural haematoma
85 Intracerebral haemorrhage/haematoma
86 Cerebrovascular accident (CVA)/stroke
87 Neurological neoplasm
88 Seizures
89 Neurological abscess
90 Encephalitis/inflammation
91 Meningitis
92 Self-inflicted overdose
93 Myaesthenia gravis
94 Guillain Barré
95 Other neuromuscular disorder
96 Non traumatic coma - metabolic disturbance
97 Non traumatic coma - anoxia/ischaemia
98 Non traumatic coma - cause unknown
99 Other neurological disorder

Gastrointestinal

Operative

110 Bleeding - ulcer
111 Bleeding - laceration/tear
112 Bleeding - varices
113 Bleeding - diverticulosis
114 Bleeding - angiodysplasia
115 GI perforation/rupture
116 GI obstruction (any cause)
117 GI neoplasm (not perforation/obstruction)
118 Localised GI abscess/cyst
119 Peritonitis
120 Pancreatitis
121 Cholangitis/cholecystitis
122 Diverticulosis
123 GI vascular insufficiency/embolism/infarction
124 GI inflammatory disease
125 Liver transplant
126 Portal-systemic shunt surgery
127 Surgery for obesity
128 Other GI surgery

Non-operative

129 Bleeding - ulcer
130 Bleeding - laceration/tear
131 Bleeding - varices
132 Bleeding - diverticulosis
133 Bleeding - angiodysplasia
134 GI perforation/rupture
135 GI obstruction (any cause)
136 GI neoplasm (not perforation/obstruction)
137 Localised GI abscess/cyst
138 Peritonitis
139 Pancreatitis
140 Cholangitis/cholecystitis
141 Diverticulosis
142 GI vascular insufficiency/embolism/infarction
143 GI inflammatory disease
144 Rejection of liver transplant
145 Hepatic failure - virus
146 Hepatic failure - toxin
147 Hepatic failure - drug reaction
148 Hepatic failure - drug overdose
149 Acute corrosive injury
150 Other GI disorder

Renal

Operative

156 Kidney transplant
157 Renal neoplasm
158 Renal infection/abscess
159 Renal bleeding
160 Renal vascular insufficiency/infarction/embolism
161 Transurethral resection
162 Renal obstruction
163 Other renal surgery

Non-operative

164 Kidney transplant rejection
165 Renal neoplasm
166 Renal infection/abscess
167 Renal bleeding
168 Renal vascular insufficiency/infarction/embolism
169 Nephrotoxic injury
170 Renal obstruction
171 Hepato-renal syndrome
172 Other renal disorder

Metabolic/endocrine

Operative

173 Adrenal neoplasm
174 Thyroid neoplasm
175 Other metabolic/endocrine surgery

Non-operative

176 Diabetic ketoacidosis
177 Adrenal neoplasm
178 Thyroid neoplasm
179 Myxoedema
180 Hypoadrenal crisis
181 Cushing's syndrome/disease
182 Hyperthyroid storm/crisis
183 Acid-base/electrolyte disturbance: diuretic induced
184 Acid-base/electrolyte disturbance: diarrhoea induced
185 Acid-base/electrolyte disturbance: GI fistula
186 Hypothermia/hyperthermia
187 Other metabolic endocrine disorder

Haematological

Operative

188 Bone marrow transplant
189 Haematological neoplasm
190 Other haematological surgery

Non-operative

191 Sickle cell crisis
192 Neutropenia
193 Thrombocytopenia
194 Blood transfusion reaction
195 Coagulopathy
196 Other haematological disorder

Trauma

Operative

197 Trauma - head/brain
198 Trauma - spine
199 Trauma - face
200 Trauma - chest
201 Trauma - abdomen
202 Trauma - pelvis
203 Trauma - extremities
204 Trauma - multiple sites plus head/brain
205 Trauma - multiple site without head/brain

Non-operative

209 Trauma - head/brain
210 Trauma - spine
211 Trauma - face
212 Trauma - chest
213 Trauma - abdomen
214 Trauma - pelvis
215 Trauma - extremities
216 Trauma - multiple sites plus head/brain
217 Trauma - multiple site without head/brain

General

Obs and gynae

221 Septic abortion
222 Pre-eclampsia/eclampsia
223 Hysterectomy

Elderly

224 Fracture of hip
225 Fracture of extremity
226 Other elderly disorder

Miscellaneous

227 Snake bite
228 Food/plant/mushroom poisoning
229 Other miscellaneous

AP3 No	APACHE III Diagnosis	Type	WW code	AMS Label	APACHE II Diagnosis	WW AP2 Code
1	Neoplasm-mouth/sinuses	Post-op	77	SRESPLAR	Thoracic surgery for neoplasm (op)	38
2	Neoplasm-larynx/trachea	Post-op	77	SRESPLAR	Thoracic surgery for neoplasm (op)	38
3	Neoplasm-lung parenchyma	Post-op	79	SRESPCA	Thoracic surgery for neoplasm (op)	38
4	Broncho-pleural fistula	Post-op	80	SRESOTH	Other respiratory (op)	48
5	All other pleural disease	Post-op	80	SRESOTH	Other respiratory (op)	48
6	Respiratory infection/abscess	Post-op	78	SRESPINF	Other respiratory (op)	48
7	Tracheal-oesophageal fistula	Post-op	80	SRESOTH	Other respiratory (op)	48
8	Other respiratory surgery	Post-op	80	SRESOTH	Other respiratory (op)	48
9	ARDS (non cardiogenic pulmonary oedema)	Medical	39	PULEDEM	Pulmonary oedema (noncardiogenic) (med)	3
10	Pneumonia-viral	Medical	32	BACVPNEU	Respiratory failure (infection) (med)	7
11	Pneumonia-parasitic	Medical	33	PARAPNEU	Respiratory failure (infection) (med)	7
12	Pneumonia-bacterial	Medical	32	BACVPNEU	Respiratory failure (infection) (med)	7
13	Pneumonia-fungal	Medical	33	PARAPNEU	Respiratory failure (infection) (med)	7
14	Pneumonia-aspiration/toxic	Medical	34	ASPPNEU	Respiratory failure (Aspiration) (med)	5
15	Neoplasm-mouth/sinuses	Medical	37	RESPCA	Respiratory neoplasm (med)	8
16	Neoplasm-larynx/trachea	Medical	37	RESPCA	Respiratory neoplasm (med)	8
17	Neoplasm-lung parenchyma	Medical	37	RESPCA	Respiratory neoplasm (med)	8
18	Pulmonary embolus	Medical	38	PULEMB	Pulmonary embolus (med)	6
19	Pulmonary hypertension (primary)	Medical	40	RESPOTH	Other respiratory (med)	26
20	Localised airway obstruction/oedema (mechanical)	Medical	35	AIROBS	Other respiratory (med)	26
21	Emphysema	Medical	42	COPDD	COPD (med)	2
22	Asthma	Medical	41	ALLERGY	Asthma/allergy (med)	1
23	Smoke inhalation	Medical	40	RESPOTH	Other respiratory (med)	26
24	Cystic fibrosis	Medical	40	RESPOTH	Other respiratory (med)	26
230	Respiratory arrest	Medical	36	RESPARR	Postrespiratory arrest (med)	4
25	Other respiratory disorder	Medical	40	RESPOTH	Other respiratory (med)	26
26	Carotid endarterectomy	Post-op	81	SCAROTID	Peripheral vascular surgery (op)	32
27	Aorto-femoral, fem-fem bypass graft	Post-op	82	SFEMAORT	Peripheral vascular surgery (op)	32
28	Fem-popliteal bypass graft	Post-op	85	SPERISC	Peripheral vascular surgery (op)	32
29	Aortic aneurysm: pre-leak/dissection	Post-op	83	SELAORT	Peripheral vascular surgery (op)	32
30	Aortic aneurysm: dissection	Post-op	84	SAORTDIS	Peripheral vascular surgery (op)	32

31	Aortic aneurysm: rupture	Post-op	84	SAORTDIS	Peripheral vascular surgery (op)	32
32	Peripheral ischaemia	Post-op	85	SPERISC	Peripheral vascular surgery (op)	32
33	Gangrenous extremity	Post-op	85	SPERISC	Peripheral vascular surgery (op)	32
34	Cellulitis	Post-op	86	SCARDOTH	Peripheral vascular surgery (op)	32
35	Septic shock - lungs (pneumonia)	Post-op	43	SEPSIS	Sepsis (op)	14
36	Septic shock - urinary tract infection	Post-op	45	SEPTICUT	Sepsis (op)	14
37	Septic shock - gastrointestinal tract	Post-op	43	SEPSIS	Sepsis (op)	14
38	Septic shock - unknown origin	Post-op	43	SEPSIS	Sepsis (op)	14
39	Pericardial effusion	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
40	Valvular repair/replacement	Post-op	88	SVALVE	Heart valve surgery (op)	33
41	Valvular repair/replacement with CABG	Post-op	88	SVALVE	Heart valve surgery (op)	33
42	Coronary artery bypass graft(s)	Post-op	72	GENOTH	Other cardiovascular (op)	47
43	Heart transplant ± lungs	Post-op	72	GENOTH	Other cardiovascular (op)	47
44	Fibrosarcoma (leg, shoulder)	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
45	Atrial myxoma	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
46	Congenital anomaly	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
47	Ventricular aneurysm	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
48	Automatic implantable cardiac defibrillator	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
49	Vena cava clipping	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
50	Vena cava filter	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
51	Other cardiovascular surgery	Post-op	86	SCARDOTH	Other cardiovascular (op)	47
52	Carotid disease/TIAs	Medical	51	PERIART	Other cardiovascular (med)	28
53	Aortic aneurysm	Medical	48	MEDAORT	Dissecting t/a aneurysm (med)	17
54	Peripheral ischaemia	Medical	51	PERIART	Other cardiovascular (med)	28
55	Hypertension	Medical	49	HYPERT	Hypertension (med)	9
56	Unstable angina	Medical	46	ACUTMI	Coronary artery disease (med)	13
57	Rhythm disturbance	Medical	50	RHYTHM	Rhythm disturbance (med)	10
58	Acute myocardial infarction	Medical	46	ACUTMI	Coronary artery disease (med)	13
59	Rule out MI	Medical	46	ACUTMI	Coronary artery disease (med)	13
60	Congestive heart failure	Medical	47	CVOTH	Congestive heart failure (med)	11
61	Cardiogenic shock	Medical	52	CARDIOG	Cardiogenic shock (med)	16
62	Gangrenous extremity	Medical	51	PERIART	Other cardiovascular (med)	28
63	Cellulitis	Medical	51	PERIART	Other cardiovascular (med)	28

64	Septic shock - lungs (pneumonia)	Medical	43	SEPSIS	Sepsis (med)	14
65	Septic shock - urinary tract infection	Medical	45	SEPTICUT	Sepsis (med)	14
66	Septic shock - gastrointestinal tract	Medical	43	SEPSIS	Sepsis (med)	14
67	Septic shock - unknown origin	Medical	43	SEPSIS	Sepsis (med)	14
68	Pericardial tamponade	Medical	47	CVOTH	Other cardiovascular (med)	28
69	Post cardiac arrest (± respiratory arrest)	Medical	44	CARDARR	Postcardiac arrest (med)	15
70	Heart transplant rejection	Medical	47	CVOTH	Other cardiovascular (med)	28
71	Cardiomyopathy	Medical	47	CVOTH	Other cardiovascular (med)	28
72	Anaphylaxis	Medical	47	CVOTH	Other cardiovascular (med)	28
73	Other cardiovascular disorder	Medical	47	CVOTH	Other cardiovascular (med)	28
74	Subarachnoid haemorrhage/intracranial aneurysm	Post-op	92	SSAH	Craniotomy for ICH/SDH/SAH (op)	39
75	Subdural/epidural haematoma	Post-op	93	SSDH	Craniotomy for ICH/SDH/SAH (op)	39
76	Intracerebral haemorrhage/haematoma	Post-op	90	SICH	Craniotomy for ICH/SDH/SAH (op)	39
77	Craniotomy for neoplasm	Post-op	89	SCRANNEO	Craniotomy for neoplasm (op)	34
78	Transphenoidal removal of neoplasm	Post-op	89	SCRANNEO	Other neuro (op)	46
79	Neurological abscess	Post-op	94	SNEUROTH	Other neuro (op)	46
80	Encephalitis/inflammation	Post-op	94	SNEUROTH	Other neuro (op)	46
81	Spinal cord surgery	Post-op	91	SLAMINE	Laminectomy and other spinal cord surgery (op)	40
82	Other neurosurgery	Post-op	94	SNEUROTH	Other neuro (op)	46
83	Subarachnoid haemorrhage/intracranial aneurysm	Medical	60	SAHMED	ICH/SDH/SAH (med)	21
84	Subdural/epidural haematoma	Medical	60	SAHMED	ICH/SDH/SAH (med)	21
85	Intracerebral haemorrhage/haematoma	Medical	59	ICHMED	ICH/SDH/SAH (med)	21
86	Cerebrovascular accident (CVA)/stroke	Medical	61	STROKE	Other neuro (med)	27
87	Neurological neoplasm	Medical	53	NEONEUR	Other neuro (med)	27
88	Seizures	Medical	62	SEIZ	Seizure disorder (med)	20
89	Neurological abscess	Medical	54	NEURINF	Other neuro (med)	27
90	Encephalitis/inflammation	Medical	54	NEURINF	Other neuro (med)	27
91	Meningitis	Medical	54	NEURINF	Other neuro (med)	27
92	Self-inflicted overdose	Medical	58	OD	Drug overdose (med)	22
93	Myaesthenia gravis	Medical	55	NEURMUSC	Other neuro (med)	27
94	Guillain Barré	Medical	55	NEURMUSC	Other neuro (med)	27
95	Other neuromuscular disorder	Medical	55	NEURMUSC	Other neuro (med)	27
96	Non traumatic coma - metabolic disturbance	Medical	57	COMAMETU	Other neuro (med)/Other metabolic renal (med)	27/25

97	Non traumatic coma - anoxia/ischaemia	Medical	44	CARDARR	Other neuro (med)	27
98	Non traumatic coma - cause unknown	Medical	57	COMAMETU	Other neuro (med)	27
99	Other neurological disorder	Medical	56	NEUROTH	Other neuro (med)	27
110	Bleeding - ulcer	Post-op	100	SGIBLEE	GI bleeding (op)	42
111	Bleeding - laceration/tear	Post-op	100	SGIBLEE	GI bleeding (op)	42
112	Bleeding - varices	Post-op	100	SGIBLEE	GI bleeding (op)	42
113	Bleeding - diverticulosis	Post-op	100	SGIBLEE	GI bleeding (op)	42
114	Bleeding - angiodysplasia	Post-op	100	SGIBLEE	GI bleeding (op)	42
115	GI perforation/rupture	Post-op	101	SGIPERF	GI perforation/obstruction (op)	45
116	GI obstruction (any cause)	Post-op	95	SGIOBS	GI perforation/obstruction (op)	45
117	GI neoplasm (not perforation/obstruction)	Post-op	99	SGICA	GI surgery for neoplasm (op)	43
118	Localised GI abscess/cyst	Post-op	96	SGIINFL	Other GI (op)	49
119	Peritonitis	Post-op	96	SGIINFL	Other GI (op)	49
120	Pancreatitis	Post-op	96	SGIINFL	Other GI (op)	49
121	Cholangitis/cholecystitis	Post-op	97	SGICHOL	Other GI (op)	49
122	Diverticulosis	Post-op	96	SGIINFL	Other GI (op)	49
123	GI vascular insufficiency/embolism/infarction	Post-op	96	SGIINFL	Other GI (op)	49
124	GI inflammatory disease	Post-op	96	SGIINFL	Other GI (op)	49
125	Liver transplant	Post-op	105	SLIVERTR	EXCLUDED	* *
126	Portal-systemic shunt surgery	Post-op	98	SGIOTH	Other GI (op)	49
127	Surgery for obesity	Post-op	98	SGIOTH	Other GI (op)	49
128	Other GI surgery	Post-op	98	SGIOTH	Other GI (op)	49
129	Bleeding - ulcer	Medical	63	GIBLEED	GI bleed (med)	24
130	Bleeding - laceration/tear	Medical	63	GIBLEED	GI bleed (med)	24
131	Bleeding - varices	Medical	65	GIBLVAR	GI bleed (med)	24
132	Bleeding - diverticulosis	Medical	64	GIBLEUL	GI bleed (med)	24
133	Bleeding - angiodysplasia	Medical	64	GIBLEUL	GI bleed (med)	24
134	GI perforation/rupture	Medical	66	GIPERF	Other GI (med)	29
135	GI obstruction (any cause)	Medical	66	GIPERF	Other GI (med)	29
136	GI neoplasm (not perforation/obstruction)	Medical	68	GIOTHER	Other GI (med)	29
137	Localised GI abscess/cyst	Medical	67	GIINLFA	Other GI (med)	29
138	Peritonitis	Medical	67	GIINLFA	Other GI (med)	29
139	Pancreatitis	Medical	67	GIINLFA	Other GI (med)	29

140	Cholangitis/cholecystitis	Medical	67	GIINLFA	Other GI (med)	29
141	Diverticulosis	Medical	67	GIINLFA	Other GI (med)	29
142	GI vascular insufficiency/embolism/infarction	Medical	67	GIINLFA	Other GI (med)	29
143	GI inflammatory disease	Medical	67	GIINLFA	Other GI (med)	29
144	Rejection of liver transplant	Medical	72	GENOTH	EXCLUDED	* *
145	Hepatic failure - virus	Medical	69	HEPATF	Other GI (med)	29
146	Hepatic failure - toxin	Medical	69	HEPATF	Other GI (med)	29
147	Hepatic failure - drug reaction	Medical	69	HEPATF	Other GI (med)	29
148	Hepatic failure - drug overdose	Medical	69	HEPATF	Other GI (med)	29
149	Acute corrosive injury	Medical	67	GIINLFA	Other GI (med)	29
150	Other GI disorder	Medical	68	GIOTHER	Other GI (med)	29
156	Kidney transplant	Post-op	104	SRENTRAN	Renal transplant (op)	36
157	Renal neoplasm	Post-op	103	SRENCA	Renal surgery for neoplasm (op)	35
158	Renal infection/abscess	Post-op	102	SRENOTH	Other metabolic/renal (op)	50
159	Renal bleeding	Post-op	102	SRENOTH	Other metabolic/renal (op)	50
160	Renal vascular insufficiency/infarction/embolism	Post-op	102	SRENOTH	Other metabolic/renal (op)	50
161	Transurethral resection	Post-op	102	SRENOTH	Other metabolic/renal (op)	50
162	Renal obstruction	Post-op	102	SRENOTH	Other metabolic/renal (op)	50
163	Other renal surgery	Post-op	102	SRENOTH	Other metabolic/renal (op)	50
164	Kidney transplant rejection	Medical	73	RENOTH	Other metabolic/renal (med)	25
165	Renal neoplasm	Medical	73	RENOTH	Other metabolic/renal (med)	25
166	Renal infection/abscess	Medical	73	RENOTH	Other metabolic/renal (med)	25
167	Renal bleeding	Medical	73	RENOTH	Other metabolic/renal (med)	25
168	Renal vascular insufficiency/infarction/embolism	Medical	73	RENOTH	Other metabolic/renal (med)	25
169	Nephrotoxic injury	Medical	73	RENOTH	Other metabolic/renal (med)	25
170	Renal obstruction	Medical	73	RENOTH	Other metabolic/renal (med)	25
171	Hepato-renal syndrome	Medical	69	HEPATF	Other metabolic/renal (med)	25
172	Other renal disorder	Medical	73	RENOTH	Other metabolic/renal (med)	25
173	Adrenal neoplasm	Post-op	111	HEMAMISC	Other metabolic/renal (op)	50
174	Thyroid neoplasm	Post-op	111	HEMAMISC	Other metabolic/renal (op)	50
175	Other metabolic/endocrine surgery	Post-op	111	HEMAMISC	Other metabolic/renal (op)	50
176	Diabetic ketoacidosis	Medical	76	DIABETIC	Diabetic ketoacidosis (med)	23
177	Adrenal neoplasm	Medical	68	GIOTHER	Other metabolic/renal (med)	25

178	Thyroid neoplasm	Medical	70	METAMISC	Other metabolic/renal (med)	25
179	Myxoedema	Medical	70	METAMISC	Other metabolic/renal (med)	25
180	Hypoadrenal crisis	Medical	70	METAMISC	Other metabolic/renal (med)	25
181	Cushing's syndrome/disease	Medical	70	METAMISC	Other metabolic/renal (med)	25
182	Hyperthyroid storm/crisis	Medical	70	METAMISC	Other metabolic/renal (med)	25
183	Acid-base/electrolyte disturbance: diuretic induced	Medical	70	METAMISC	Other metabolic/renal (med)	25
184	Acid-base/electrolyte disturbance: diarrhoea induced	Medical	70	METAMISC	Other metabolic/renal (med)	25
185	Acid-base/electrolyte disturbance: GI fistula	Medical	70	METAMISC	Other metabolic/renal (med)	25
186	Hypothermia/hyperthermia	Medical	70	METAMISC	Other metabolic/renal (med)	25
187	Other metabolic endocrine disorder	Medical	70	METAMISC	Other metabolic/renal (med)	25
188	Bone marrow transplant	Post-op	72	GENOTH	Other cardiovascular (op)	47
189	Haematological neoplasm	Post-op	111	HEMAMISC	Other cardiovascular (op)	47
190	Other haematological surgery	Post-op	111	HEMAMISC	Other cardiovascular (op)	47
191	Sickle cell crisis	Medical	111	HEMAMISC	Other cardiovascular (med)	28
192	Neutropenia	Medical	71	COAGTHRO	Other cardiovascular (med)	28
193	Thrombocytopenia	Medical	71	COAGTHRO	Other cardiovascular (med)	28
194	Blood transfusion reaction	Medical	71	COAGTHRO	Other cardiovascular (med)	28
195	Coagulopathy	Medical	71	COAGTHRO	Other cardiovascular (med)	28
196	Other haematological disorder	Medical	111	HEMAMISC	Other cardiovascular (med)	28
197	Trauma - head/brain	Post-op	107	SHEADTR	Head trauma (op)	37
198	Trauma - spine	Post-op	108	SMULTR	Multiple trauma (op)	30
199	Trauma - face	Post-op	108	SMULTR	Multiple trauma (op)	30
200	Trauma - chest	Post-op	108	SMULTR	Multiple trauma (op)	30
201	Trauma - abdomen	Post-op	108	SMULTR	Multiple trauma (op)	30
202	Trauma - pelvis	Post-op	108	SMULTR	Multiple trauma (op)	30
203	Trauma - extremities	Post-op	108	SMULTR	Multiple trauma (op)	30
204	Trauma - multiple sites plus head/brain	Post-op	107	SHEADTR	Head trauma (op)	37
205	Trauma - multiple site without head/brain	Post-op	108	SMULTR	Multiple trauma (op)	37
209	Trauma - head/brain	Medical	74	HEADTR	Head trauma (med)	19
210	Trauma - spine	Medical	75	MULTRAUM	Multiple trauma (med)	18
211	Trauma - face	Medical	75	MULTRAUM	Multiple trauma (med)	18
212	Trauma - chest	Medical	75	MULTRAUM	Multiple trauma (med)	18
213	Trauma - abdomen	Medical	75	MULTRAUM	Multiple trauma (med)	18

214	Trauma - pelvis	Medical	75	MULTRAUM	Multiple trauma (med)	18
215	Trauma - extremities	Medical	75	MULTRAUM	Multiple trauma (med)	18
216	Trauma - multiple sites plus head/brain	Medical	74	HEADTR	Head trauma (med)	19
217	Trauma - multiple site without head/brain	Medical	75	MULTRAUM	Multiple trauma (med)	18
221	Septic abortion	Medical/Post-op	72	GENOTH	Other cardiovascular (med)/(op)	28/47
222	Pre-eclampsia/eclampsia	Medical/Post-op	49	HYPERT	Other cardiovascular (med)/(op)	28/47
223	Hysterectomy	Medical/Post-op	106	SOBHYST	Other cardiovascular (med)/(op)	28/47
224	Fracture of hip	Medical/Post-op	87	SHIPS	Other cardiovascular (med)/Admission due to chronic CV disease(op)	28/31
225	Fracture of extremity	Medical/Post-op	87	SHIPS	Other cardiovascular (med)/Admission due to chronic CV disease(op)	28/31
226	Other elderly disorder	Medical/Post-op	86	SCARDOTH	Other cardiovascular (med)/(op)	28/47
227	Snake bite	Medical/Post-op	58	OD	Other cardiovascular (med)/(op)	28/47
228	Food/plant/mushroom poisoning	Medical/Post-op	58	OD	Other cardiovascular (med)/(op)	28/47
229	Other miscellaneous	Medical/Post-op	72	GENOTH	Other cardiovascular (med)/(op)	28/47