Mortality prediction and acuity assessment in critical care

Alistair E. W. Johnson

Wolfson College

Department of Engineering Science

Supervised by
Gari D. Clifford
Andrew A. Kramer

Submitted: Trinity Term, 2014

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy

Acknowledgements

The writing of this thesis, and the process of learning and research throughout, has been a transformative experience. Oxford is such a wonderful place, full of enthusiastic researchers, all of whom helped shape my world view, and who continue to shape it even today. I have been privileged to be exposed to so many great minds, kindred souls, and keen colleagues during my years here. It would be impossible to name everyone who helped shape my life at Oxford, but hopefully you can forgive me for trying.

I would like to give my sincerest thanks to my two D.Phil supervisors, Gari Clifford and Andrew Kramer. Gari, your limitless energy and passion for your research are something I hope to take on in my future career. Andrew, I could not have progressed half as far without your wealth of knowledge about all things risk adjustment and your earnest belief that I could make a difference in this field.

Of course, the people are what make the lab, and I can honestly say I had fantastic colleagues at every step. Louis, we've had such great times both near and abroad, and you've always been my role model. Thanks so much for being you, and remember, WWLD. Tingting, thanks for all the lively discussion both academic and otherwise, even though we may not always agree on everything. Actually, I don't think we agree on anything, but that was always the fun part. Joachim, your dedication to the truth in your research is truly inspiring, and it was a privilege to work with you. Marco, a good friend and a better scientist, working alongside you was a joy and reminded me of what academia was all about. Julien, thanks for teaching me all about the three "S"s: signal processing, switching kalman filters, and salsa dancing. Thanks also to "xiao" Mauro, for reminding me that not having to collect data myself was a blessing, and "da" Mauro, for your systems and database advice that I never really understood. And of course Dave, Aoife, Lisa, Arvind, Nic, Thanasis, Maxim, for all those coffee run chats and for helping to create such a great lab environment.

I've been blessed to have been mentored by other great minds in Oxford. I'd like to thank in particular Lionel Tarassenko; who taught me that engineering doesn't have to be fancy, it just needs to work well, Duncan Young; for reminding me to ask "so what?" and always taking the time to teach me about physiology, and Peter Watkinson; whose acerbic wit always lifted our spirits at work... except when it didn't. I'd also like to thank my thesis examiners, David Harrison and David Clifton, whose thorough reading and feedback were immeasurably important in making the thesis what it is today.

On a personal level, I'd like to thank my father, David Johnson, who years ago casually uttered the phrase "what about Oxford?". I'd also like to thank my mother,

Claire Johnson, for telling him to do so. Both of you have given me all I've ever needed as a son, and more. I love you both. I'll move back to Canada soon, I promise!

The rest of my family deserve a shout out - my older brother, Justin, for showing me that 30 is the new 25 (sorry, couldn't help it). My younger brother, Oliver, for living life to the fullest and always enriching my life with his antics. My sister, Louisa, for being the sweetest person I know. And finally, my grandparents, Bill and Jennie, for welcoming me into England and always making me feel at home here.

They say friends are the family you choose, but I'd be arrogant to say I could have chosen such wonderful people. Yvonne, whose excitement about everything is infectious. Johan, who showed me there's more to life in Oxford than just writing. Jon, who made me realize that I have way more free time than I thought. Nisha, Steph, Chris, Serena, Francesca, Will, we had so many great times together!

Finally, I'd like to say a few words about someone very dear to me. Penny, you are the light of my life. I don't know what I did to find someone as special as you, and I'm not just saying that because you laugh at my jokes. You're smart, funny, sometimes intentionally funny, sweet, empathetic. You called me the "perfect guy", and you stuck with me when you realized, well, maybe not *perfect*. I could go on for pages, but I'll summarize: I love you, my soul mate, and if we can get through our DPhils together, the rest of life can't be so bad, can it?

Mortality prediction and acuity assessment in critical care

Alistair E. W. Johnson
Thesis submitted for the degree of Doctor of Philosophy
Wolfson College
Trinity 2014

Abstract

Accurate mortality prediction in intensive care units (ICUs) allows for the risk adjustment of study populations, aids in patient care and provides a method for benchmarking overall hospital and ICU performance. ICU risk-adjustment models are primarily comprised of an integer severity of illness score which increases with increasing patient risk of mortality. First published in the 1980s, the improvements to these scores primarily consisted of increasing the dimensionality of the model, and hence also increasing their complexity. This thesis aims to improve upon these models. First, the field is surveyed and the major models for risk-adjusting critically ill patient cohorts are identified including the acute physiology score (APS) and the simplified acute physiology score (SAPS). A key component of model performance is data preprocessing. The effect of preprocessing ICU data is quantified on a dataset of 8,000 ICU patients, and it is shown that after preprocessing to remove extreme values a logistic regression (LR) model performed competitively (AUROC of 0.8633) with the more complex machine learning model; a support vector machine (SVM) which had an AUROC of 0.8653. For validation, model development was repeated in a larger database containing over 80,000 patients admitted to 89 ICUs in the United States. Results were similar (AUROC of 0.8895 for the LR vs 0.8917 for the SVM) but showed the performance gain when using automated outlier rejection is less pronounced in well quality controlled datasets (0.8883 for LR without rejection). It is hypothesised from this that simpler models can perform competitively with more complicated models, while having a greatly reduced burden of data collection. A severity score is developed on the large multi-center database using a Genetic Algorithm and Particle Swarm Optimisation. The severity score, named the Oxford Acute Severity of Illness Score (OASIS), is shown to outperform the APS III (AUROC 0.837 vs 0.822) and perform competitively with APACHE IV when used as a covariate in a regression model (AUROC 0.868 vs 0.881). The severity score requires only 10 variables (58% as many as APS III), reducing the burden of quality control and data collection. These variables are routinely collected in critical care by continuous monitors and do not include comorbidities, diagnosis or laboratory measurements. The severity score is then externally evaluated in an American hospital and shown to discriminate well (AUROC 0.790 vs. 0.782 for the APS III) with excellent calibration. Finally, the severity score was evaluated in an English hospital and compared to other severity scores. OASIS again had excellent calibration and discrimination (AUROC 0.776 vs 0.750 for APS III) whilst requiring a much smaller number of variables. OASIS has many applications, including both simplifying data collection for studies and improving the risk assessment therein.

Contents

List of Figures

Page

List of Tables

Page

D Detailed performance comparisons

Nomenclature

Much of the work in this thesis involves learning a model from data. These data are always presented to the model in the form of a *design matrix*. While this is not the only learning paradigm, it is the focus of this thesis. A bold upper case variable represents a matrix and usually $\mathbf{X}$ will represent the design matrix with N observations and D features. Usually vectors will not be denoted by bold font weight, unless the vector is of length N, e.g. a column vector $\mathbf{x}_j$ of length N from a design matrix $\mathbf{X}$. This allows for the easy differentiation of row vectors x_i and column vectors $\mathbf{x}_j$ from the matrix $\mathbf{X}$. In general, a bolded vector should be considered as containing values for a single column (i.e. across rows of a design matrix which represent independent instantiations of the data such as distinct patients), while a non-bolded vector will contain values across some other dimension (usually across columns of the design matrix).

Each column of the design matrix contains values related to a single *feature*. Care is taken to distinguish between a *variable*, which is a continuous parameter measurable over an indefinite period of time (such as heart rate), and a *feature*, which is a single numeric quantification of some aspect of the variable. For example, a patient's heart rate is considered as a variable, whereas the lowest patient heart rate in the first 24 hours of their ICU stay is considered a feature. The key distinction is that, for a single patient represented by a single row in the design matrix, a variable may have multiple values but a feature must have a single value.

Each row of the design matrix represents a set of instantiations of the features, referred to as an *observation*. Observations are collections of features which are derived from the same source, where the source in this work is a patient being monitored in the ICU. Concretely, we can represent observation i using a vector $x_i \in \mathbb{R}^D$, with each of the $j = 1, ..., D$ dimensions corresponding to a single feature. For example, if the first

dimension corresponds to the lowest heart rate, then $x_{i,j}$ corresponds to the lowest heart rate for observation i, and the vector x_i represents a single *observation* of design matrix $\mathbf{X}$. The term observation is used interchangeably in this work with the term "record", the term "patient" (as most observations correspond to a single patient) or explicit references to a row of the design matrix.

The Hadamard product, corresponding to element wise multiplication of two matrices of the same size, is denoted by $\circ$. Furthermore, the mean of a vector will be denoted using a bar above the vector, i.e. $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N} y_i$ where N is the number of elements in $\mathbf{y}$.

There are iterative algorithms used in this work. In order to allow for better readability, iterations of the algorithms are specified by a superscript in parentheses, while observations for a given vector (as described earlier) are specified by a subscript. For example, the k^{th} iteration of matrix $\boldsymbol{X}$ is represented by $\boldsymbol{X}^{(k)}$, while the i^{th} row and j^{th} column in matrix $\boldsymbol{X}$ is represented by $x_{i,j}$. The two schemes can be combined to reference the i^{th} row, j^{th} column and k^{th} iteration of the matrix $\boldsymbol{X}$ as follows: $x_{i,j}^{(k)}$.

The goal of severity of illness scores is to learn a functional mapping between the data, $\mathbf{X}$, and patient health. This is accomplished by explicitly learning a mapping between the data and patient mortality. The goal then becomes a two class mapping to predict an outcome of 0 or 1. It is worth noting that this is not necessarily an optimal framework for the problem. It has been suggested that instead a three class classification be used, where the three classes are severely ill (not salvageable and moribund), salvageable, and healthy (survival is guaranteed) [1]. Unfortunately, annotations for this three class problem are not readily available and consequently the two class formulation to predict mortality is utilised. The outcome for each observation is represented by the vector $\mathbf{y}$, which can take on two values (0 or 1). This problem is known as binary classification and is very common in machine learning.

Introduction

> *...the quality of my care and my confidence in its outcomes would never be better than the quality of the information behind them. The information in the room [ICU] that morning was detailed and exhaustive but was impossible to organize and interpret within the time required.*
>
> William A. Knaus, founding partner of APACHE

The Intensive Care Unit (ICU) is a hospital department to which only the most severely ill patients are admitted. The primary function of the ICU is to deliver care to patients, such as mechanical ventilation, which cannot be administered in other areas of the hospital. Patients in the ICU are the most heavily monitored patients in the entire hospital, and a one to one staff to patient ratio is recommended by the National Health Service (NHS) for England [2]. The high amount of staff and advanced technical equipment allows for continuous monitoring of all patients to ensure any deterioration in their condition is detected and corrected for before it becomes fatal. Studies have corroborated the efficacy of this approach, with higher nurse to patient ratios shown to reduce mortality, hospital length of stay and illness complications [3] [4].

The ICU is a data rich environment, even to the point of exhaustion. William A. Knaus, a principle developer of the Acute Physiology, Age, and Chronic Health Evaluation (APACHE) system [5] for predicting patient mortality, described his first day as an attending physician in the intensive care unit in a poignant reflection article [6]. Early on a Saturday morning a patient was rapidly deteriorating. The patient had been haemodynamically unstable, was in severe septic shock and despite intensive fluid rescucitation and vasopressor therapy expired that day. Dr. Knaus well articulated

the situation: "The information in the room that morning was detailed and exhaustive but was impossible to organize and interpret within the time required." At the time in 1978, there was no system for assessing a patient's illness or comparing it to similar patients who had been treated earlier. The desire to empirically model patient severity in order to derive further insight and improve patient care was a driving motivator in the development of severity of illness scores. These scores are commonly used today and synthesize various physiologic and demographic information regarding a patient into a single integer which correlates with patient mortality. The components of these scores are used in models which directly estimate a probability of mortality.

The APACHE system was the first among these systems [5], being designed to be applicable to all adult patients in the ICU and yet retain the accuracy to make clinical decisions for an individual patient. The vision was of a universal system for prognostication - one that was systematic and consistent across ICUs. The subsequent reality was, unfortunately, far from this. The third incarnation of the APACHE system [7] was made commercial, as venture capital had funded the extensive research and development costs. The critical care community strongly opposed the proprietary nature of the APACHE III model, and other alternatives gained popularity such as the Simplified Acute Physiology Score (SAPS) [8], SAPS II [9], Mortality Probability Model (MPM) [10] and the MPM II [11].

The community became fragmented, and no single model for calculating the prognosis of an ICU patient was ever agreed upon. In the interim, many were criticising the ability of these models to accurately predict patient mortality on an individual level [12]. The consensus of the critical care community was that these models could not, and consequently should not, be used to predict mortality for an individual patient [13]. Nevertheless, the models still found use in evaluating the performance of an ICU by comparing the average mortality across a group of patients to the expected mortality for that group. Evaluating the crude mortality rates of ICUs with no regard to the severity of illness on admission unfairly penalises units which, for example, accept a higher number of terminally ill patients [14]. In 1999 the United States (US) ICU beds comprise up to 20% of total hospital beds, whilst ICU beds in the United Kingdom

(UK) comprise only up to 2% of total hospital beds [15]. UK ICUs tend to have more severely ill patients as compared to US ICUs [15], and with higher crude mortality rates an unadjusted comparison between the US and the UK would be unfairly critical of the level of care provided in the UK. A more commonly occurring difference in case-mix is that between teaching hospitals and non-teaching hospitals.

The process of risk-adjusting patients admitted to an ICU thus became a key component for evaluating the quality of care delivered at that institution. Risk adjustment has been successfully used in the past to identify hospitals with both statistically significantly higher and signficiantly lower levels of performance [16]. Hospitals are also capable of temporal self evaluation, assessing whether the facility is improving or worsening despite changing population severity and demographics. Clinicians were initially resistant to the idea of a "report card" on their performance [17], even going so far to associate the use of APACHE for predicting mortality with religious divination [6]. In 1994 the Intensive Care National Audit & Research Centre (ICNARC) established the Case Mix Programme (CMP), a national clinical audit for adult critical care, and the importance of evaluating the quality of care delivered rose to public prominence after the Bristol inquiry into poor surgical outcomes after paediatric surgery [18]. The Bristol Report [18, 19] emphasised the need for routinely assessing the quality of care delivered in all aspects of hospital care, and in the ICU the use of risk prediction models such as APACHE is integral to this task.

Furthermore, ICUs are consistently the most expensive department of the hospital, with adult intensive care units costing an estimated £1,219 (level 2 ICU bed) and £1,638 (level 3 ICU bed) per patient per bed day in the UK in 2010 [20]. Other healthcare systems have similar costs, as ICUs in the US were estimated to cost $1,699 per patient per bed day in 2005 [21]. Halpern and colleagues also reported a 32.7% increase in per day per bed patient cost from 2000 to 2005, implying a cost increase of 5.82% compounded per year in the US [21]. As a percentage of hospital costs and the gross domestic product (GDP), US ICUs represented 13.4% of hospital costs and 0.66% of the US GDP in 2005, while UK ICUs represented 5.01% of hospital costs and 0.58% of the 2005 UK GDP [20] [21]. This high cost has driven the need for performing cost-benefit

analyses and maximizing return on ICU investment [22].

Severity scores have found many uses in addition to risk adjustment. These scores are frequently used to compare the case-mix of two groups in a clinical trial to ensure they are comparable [23]. They have become especially important in trials of potential treatments for severe illnesses such as sepsis [24, 25] or acute respiratory distress syndrome [26, 27]. These scores can also be used to optimise the efficiency of care delivery [28], and identify policies with positive effects on outcomes such as a high intensivist staffing [29]. Finally, the risk of mortality can help facilitate end-of-life discussions between care givers and families [30–32].

However, there exist many shortcomings of the prognostic models currently used in clinical practice. First, the derivation of the models is a complex process in which clinical guidance is primarily used to assign varying levels of severity to patient physiology. In contrast, recent clinical practice strives for more empirical, evidence based care [33]. Furthermore, recently developed acuity models search for an optimal predictor using a univariate selection approach and selective interaction terms which are not exhaustive of all possible terms [34] [35]. Since an exhaustive search would be computationally infeasible, the result has been the development of potentially sub-optimal models; using a large number of features with little accounting for their correlation. This is partially responsible for the growth in complexity in recent severity scores and the associated increased difficulty in their implementation. Such complexity is a major barrier for clinical acceptance [35]. Certain features may be difficult to reproduce because their definition is vague (e.g. degree of heart failure), and others may not be routinely collected during ICU care (e.g. cardiopulmonary resuscitation prior to ICU admission). Due to the lack of effective data management systems, many studies which require severity of illness scores involve manually calculating the scores. This process is greatly simplified by a smaller number of features which have clear and concise definitions. Multivariate feature selection approaches offer the possibility to reduce the burden of data collection while not sacrificing performance. Additionally, recent trends in hardware and data collection have increased clinical database sizes. In 1981 the APACHE I system was validated on a data set of 581 admissions, while the APACHE IV system was validated

in 2006 on a data set of over 44,000 patients [5] [35]. Larger databases provide more flexibility in the modeling approach and allow for the use of more complicated techniques. The increase in data availability has coincided with a large amount of research into more complicated and potentially advantageous techniques for classification, such as Support Vector Machines (SVMs) [36] and Random Forests (RFs) [37]. Though these models have largely not transitioned from research into clinical practice, they remain a potential avenue for improving model performance.

This thesis aims to evaluate the estimation of patient acuity and prediction of mortality for patients admitted to the ICU. A number of machine learning techniques capable of capturing non-linear feature interactions are compared to regularised logistic regression models. The importance of preprocessing the data prior to model construction is highlighted. Finally, the potential of multivariate feature selection techniques in the field of ICU mortality prediction is demonstrated by the development of an effective parsimonious severity score.

Chapter 1

Mortality prediction in the ICU

I was not predicting the future, I was trying to prevent it.

Ray Bradbury

This chapter will first detail the general framework of developing a severity of illness score and provide a necessary introduction to the evaluation statistics used. The most common ICU severity of illness scores and systems will then be detailed. Particular emphasis will be placed on the differing methods of feature selection, but full exposition left in the references for the interested reader. The chapter concludes with comparison studies using the various severity scores.

1.1 Evaluation of model performance

Since the output predicted is throughout this work is either in-hospital patient mortality or in ICU patient mortality, the majority of the statistical tests presented are for dichotomous outcomes. Almost all binary outcome evaluation metrics will fundamentally involve interpreting the number of true positives, false positives, true negatives and false negatives. Table 1.1 describes these metrics. Patient mortality is represented as a positive outcome in the mathematical and not clinical sense, as is standard in the literature.

The efficacy of models which classify data can be easily interpreted using two con-

Metric	General Description	Model predicts patient dies	Patient dies
True Positive (TP)	Correct prediction	Yes	Yes
False Positive (FP)	Incorrect prediction	Yes	No
True Negative (TN)	Correct prediction	No	No
False Negative (FN)	Incorrect prediction	No	Yes

Table 1.1: *Description of the measurements which form the basis of many model evaluation statistics. Note that a positive outcome in this report refers to positive in the mathematical sense and corresponds to a patient expiring in the ICU or hospital.*

cepts: the calibration and the discrimination of the model [38]. The calibration of the model is how closely a model's predictions match the true patient probability of death across the range of risk from zero to one. It provides information regarding the suitability of the model to a data set, but does not infer anything about the suitability of the model to each individual observation. Usually a model is naturally calibrated to the data set on which it is trained, though a small variance is possible [38]. The discrimination of the model for dichotomous outcomes represents how well a model distinguishes the positive and negative outcome populations. Often discrimination is the aspect of a model to which most focus is given. The reasoning for this is intuitive: models with high discrimination can be rescaled to have high calibration for a given data set. For example, assume a situation with both outcomes, 0 and 1, being equally likely, with two models, A and B, attempting to classify the data as true or false. Model A has high discrimination but low calibration, while model B has high calibration but low discrimination. Table 1.2 shows an example of these two models. Model A assigns every instance with a negative outcome (0) a prediction of 0.4 and each instance with a positive outcome (1) a prediction of 0.8. Model B assigns every instance, regardless of the true outcome, a prediction of 0.5.

Though model A is much more predictively useful, as it successfully separates the two classes, it is on average less accurate than model B. Thus model A has higher discrimination and model B has higher calibration[1]. However, while model A can be

[1]With slight abuse of terminology, we have treated comparison of estimate and actual prevalence as equivalent to calibration in this toy example

Case	Model A's prediction	Model B's prediction	Observed Class
1	0.4	0.5	0
2	0.4	0.5	0
3	0.8	0.5	1
4	0.8	0.5	1
Mean	0.6	0.5	0.5

Table 1.2: *Example output from two models highlighting the difference between calibration and discrimination.*

rescaled to have higher calibration, there is no affine transform that can increase model B's discrimination. Additionally, it is clear that high calibration does not guarentee a useful model. It is for this reason that discrimination is the most commonly used criteria in evaluating classification models. The process of later rescaling the output of a model to have higher calibration is known as "recalibration" [38]. Nevertheless, calibration is an important aspect of model predictions as poorly calibrated models will provide biased estimates of actual patient risk. Calibration and discrimination are considered complementary evaluations and good performance in both is expected in high quality ICU mortality prediction models [39].

1.1.1 Operating point statistics

An operating point for a probabilistic classifier is the threshold which separates positive and negative predictions. Higher thresholds increase confidence in positive predictions being correct, and vice-versa. The operating point statistics used in this work are the accuracy, sensitivity, specificity, positive predictive value and the negative predictive value.

1.1.2 Receiver operating characteristic curve

The receiver operating characteristic curve (ROC curve) is a very popular technique that represents a model's discrimination in a simple easily interpretable 2D plot [40]. The true positive rate (sensitivity) of the model is plotted against the false positive rate (1-specificity). This is accomplished by calculating the true positives, true negatives, false positives and false negatives for every possible operating point. The sensitivity and specificity are then calculated from these values. An ideal model has a specificity

Statistic	Equation	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Proportion of correctly classified outcomes
Sensitivity	$\frac{TP}{TP+FN}$	Ability to discern positive outcomes
Specificity	$\frac{TN}{TN+FP}$	Ability to discern negative outcomes
Positive Predictive Value	$\frac{TP}{TP+FP}$	Reliability of positive outcome predictions
Negative Predictive Value	$\frac{TN}{TN+FN}$	Reliability of negative outcome predictions

Table 1.3: *List of the various operating point statistics used in this thesis and the formula for calculating them. A brief explanation of the diagnostic value of the statistic is also provided.*

and sensitivity of 1, located at (0,1) on the 2D plot. All models will contain the point, (0,0) equivalent to predicting a negative outcome for all cases, and the point (1,1), equivalent to predicting a positive outcome for all cases. A straight dashed line from (0,0) to (1,1) represents chance performance. As models improve, their ROC curve will move away from the straight dashed line toward the top left corner of the plot (which is equivalent to perfect discrimination). This curve is useful for assessing the trade off between sensitivity and specificity and selecting an operating point for the model being evaluated.

1.1.3 Area under the receiver operating characteristic curve

One useful measurement often used in model analysis is the Area Under the Receiver Operator Characteristic curve (AUROC) [40]. This metric ranges from zero to one, where higher values indicate higher model discrimination. Mathematically, the AUROC is equivalent to the Wilcoxon statistic, and can be viewed as $P(\hat{y}|y = 1 > \hat{y}|y = 0)$, where $\hat{y}$ is the prediction evaluated and y is the observed outcome (i.e. patient mortality). Intuitively, the AUROC is the probability that the predictions will rank a positive outcome higher than a negative outcome. A value of 0.5 indicates that the model does not predict any better than chance. Though the AUROC summarizes the discrimination and thus performance of the model well, it does not provide the vital information regarding the trade off between sensitivity and specificity that the ROC curve does.

1.1.4 Hosmer-Lemeshow statistic

The Hosmer-Lemeshow statistic tests a model against the null hypothesis that it is perfectly calibrated [41]. Predictions are grouped into deciles of ordered risk and each deciles' predicted positive outcome rate is compared with the deciles' observed outcome rate. The deviation from perfect fit can then be tested by approximating the sum of the deviations with a χ^2 distribution. This work exclusively uses this statistic with bins of equal sample size, denoted as the Hosmer-Lemeshow $\hat{C}$ statistic ($HL_{\hat{C}}$). The formula for its calculation is as follows:

$$HL_{\hat{C}} = \sum_{j=1}^{G} \frac{(O_j - E_j)^2}{n_j p_j (1 - p_j)} \tag{1.1}$$

Where, for decile j, O_j is the number of observed events, E_j is the number of predicted events, n_j is the number of observations and p_j is the predicted probability of a positive outcome. The statistic is sample size dependent and follows a chi-square distribution with $G-2$ degrees of freedom on the training set and G degrees of freedom on an external test set. The null hypothesis is that the model is perfectly calibrated: statistical significance (which rejects the null hypothesis) indicates the model is *not* perfectly calibrated. χ^2 values for various levels of significance are shown in Table 1.4.

Degrees of freedom	χ^2 **value**								
8	3.49	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
10	4.86	7.27	9.34	11.78	13.44	15.99	18.30	23.21	29.59
p value	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Non-significant						Significant		

Table 1.4: *Table of χ^2 values used for assessing significance of Hosmer-Lemeshow test, which is approximately χ^2 distributed. Statistical significance indicates a lack of model fit. Degrees of freedom depend on whether the evaluation is on the development dataset (8 degrees of freedom) or a test dataset (10 degrees of freedom).*

Note that the dependency on sample size makes comparison of $HL_{\hat{C}}$ difficult for different data sets. This sample size dependency can be succinctly highlighted by a reorganization of Equation 1.1:

$$HL_{\hat{C}} = \sum_{j=1}^{G} \frac{n_j p_j (1 - \frac{O_j}{E_j})}{(1 - p_j)} \tag{1.2}$$

When the calibration of the model is perfect, the ratio of $\frac{O_j}{E_j}$ is 1, and therefore the sample size has no effect on the significance of the result. However, with even slight deviation from perfection, high sample sizes can cause high $HL_{\hat{C}}$ statistics, which provide an exaggerated impression of lack of fit. This is due to the statistic testing against a null hypothesis of *perfect* model fit: models at high sample sizes may be well calibrated but they are very unlikely to be perfectly calibrated. Detailed elaboration upon this phenomenon, including analysis showing that the manifestation of a 4% lack of calibration strongly depends on sample size, is provided in the references [42].

1.1.5 Standardized mortality ratio

The Standardized Mortality Ratio (SMR) is a measure of a binary or probabilistic model's calibration "in the large", meaning it assesses the concordance of the average risk of a model with the observed mortality rate. It is calculated as the number of patients who die divided by the sum of all patient predictions, as follows:

$$\text{SMR} = \frac{\sum_{j=1}^{N} y_j}{\sum_{j=1}^{N} \hat{y}_j}, \tag{1.3}$$

where y is the observed outcome (0 or 1) and $\hat{y}$ is the predicted outcome (a value between 0 and 1). Models whose prediction prevalence (i.e. mean of predictions) matches the prevalence of the outcome will have an ideal SMR value of 1.

1.1.6 Brier score

The Brier score (B), or the mean square error, was a metric used to assess the accuracy of meteorological forecasts [43]. B is calculated as the average squared error of all observations minus all predictions, as follows:

$$B = \frac{1}{N} \sum_{j=1}^{N} (\hat{y}_j - y_j)^2 \tag{1.4}$$

Perfect predictions have a B= 0.

1.1.7 Efron's pseudo R^2

One issue with the B is its dependence on the frequency of positive outcomes in the outcome vector $\mathbf{y}$, which complicates comparisons between studies with different mortality rates. Adjusted Brier score (B_{adj}) is a modified version of the B which compensates for the prevalence of the evaluated outcome [44]. The much large proportion of survivors in ICU data can bias evaluation statistics [45] and the B_{adj} has recently begun to gain popularity in evaluating mortality prediction models [46]. First the null brier score (B_n), which represents the B of a null model, is calculated as:

$$B_n = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{1}{N}\sum_{k=1}^{N} y_k\right)^2 \tag{1.5}$$

Here the null model represents predicting the incidence of the outcome (mean of $\mathbf{y}$) for all observations. The B_{adj} is then calculated as follows:

$$B_{adj} = \frac{B_n - \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{B_n} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{B_n} \tag{1.6}$$

Note that while a smaller B indicates better performance, a higher B_{adj} indicates better performance. Note that the B_{adj} is equivalent to the adjusted Brier score [46] or the sum-of-squares R^2 (see [47] for an overview of R^2 like values used to assess logistic regression models). The metric is intuitively interpretable. When $\hat{\mathbf{y}}$ is produced by the null model (i.e. the predictions are equal to the outcome prevalence), the latter term in Equation 1.6 is equal to B_n resulting in $1 - \frac{B_n}{B_n} = 1$. If all of the variance in $\mathbf{y}$ is explained by $\hat{\mathbf{y}}$ (the model predictions) then $\hat{\mathbf{y}} = \mathbf{y}$ and $\frac{1}{N}\sum_{j=1}^{N}(\hat{y}_j - y_j)^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - y_j)^2 = 0$. As a result the B_{adj} becomes $1 - \frac{0}{B_n} = 1$. Thus the B_{adj} can be thought of as the proportion of variance explained by the model which is not explained by a null model.

1.1.8 Likelihood improvement

The log likelihood is a measure commonly used to assess the fit of a model. For a binomial distribution, the log likelihood can be calculated as follows:

$$\log(\mathcal{L}(\hat{\mathbf{y}};\mathbf{y})) = \sum_{i=1}^{N} (y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i)) \tag{1.7}$$

The Likelihood Improvement ($\mathcal{I}_\mathcal{L}$) is defined as the improvement in the log likelihood when using a set of predictions, $\hat{\mathbf{y}}$, versus a null model which uses the mean of the outcome as the prediction for every observation. The $\mathcal{I}_\mathcal{L}$ is calculated as:

$$\mathcal{I}_L = \frac{\log(\mathcal{L}(\bar{\mathbf{y}};\mathbf{y})) - \log(\mathcal{L}(\hat{\mathbf{y}};\mathbf{y}))}{\log(\mathcal{L}(\bar{\mathbf{y}};\mathbf{y}))} = 1 - \frac{\log(\mathcal{L}(\hat{\mathbf{y}};\mathbf{y}))}{\log(\mathcal{L}(\bar{\mathbf{y}};\mathbf{y}))} \tag{1.8}$$

where $\mathcal{L}(\mathbf{x};\bar{\mathbf{y}})$ is the likelihood of the null model which predicts the mean outcome ($\bar{\mathbf{y}}$) for every observation. When the model is equivalent to the null model, there is no reduction in the log likelihood which leads to a $\mathcal{I}_\mathcal{L} = 0$[2]. Conversely, the ideal model perfectly predicts the binary outcome and has a log likelihood of zero. If the log likelihood of the model is zero, the numerator is equal to the null log likelihood and consequently the $\mathcal{I}_\mathcal{L} = 1$. The $\mathcal{I}_\mathcal{L}$ was proposed as a measure of model fit by McFadden [48] and has since been commonly referred to as McFadden's pseudo R^2 or the entropy based R^2 (see [47] for a review of pseudo R^2 values used to assess logistic regression models).

1.2 Assessment of generalisation performance

There exist many different methods for estimating generalisation performance of a model given a set of data. While it is possible to estimate the efficacy of a model optimised on a set of data using the same data, this provides an optimistic estimate of model performance [49, 50]. Furthermore, many more complex models gain the ability to "memorize" the training set and perform well on that specific set due to their larger degrees of freedom. These models, when performing perfectly on the development data, will usually fail to perform well on an external dataset. This process is known as overfitting, and in the statistics literature is often referred to as the model having too much *variance* [50, 51]. Many methods have been developed to guard against this phenomenon.

[2]Note that the likelihood of a model's predictions can be lower than the null likelihood resulting in a negative $\mathcal{I}_\mathcal{L}$, though this is unlikely in practice.

1.2.1 Hold out

Hold out is a relatively straight forward model evaluation technique which involves excluding a certain fraction of the dataset from the model development. That is, for a design matrix $\mathbf{X}$ with N rows, we calculate an indicator vector as $\mathbf{V} \sim \mathcal{B}(N, p)$, where p is a chosen proportion of held out data. The model is then developed using observations for which $\mathbf{V} = 1$, and evaluated on observations for which $\mathbf{V} = 0$. The technique only provides a good approximation of the generalization error if sufficient samples are available for the held out set.

1.2.2 Cross-validation

Cross-validation involves segmenting the data into subsets of equal size. Each segment is then excluded in repeated model developments, and model performance is evaluated on these held out segments. Formally, if we have a dataset $\mathbf{X}$, K-fold cross-validation involves randomly assigning each row in the design matrix $\mathbf{X}$ an index v_i where:

$$P(v_i = k) = \frac{1}{K},\ v_i \in \{1, ..., K\} \tag{1.9}$$

K models are then developed, where the k^{th} model uses rows in $\mathbf{X}$ where $\mathbf{v} \neq k$. Furthermore, the predictions of the k^{th} model are calculated for the rows in $\mathbf{X}$ where $\mathbf{v} = k$. The combination of predictions from the K models results in a single prediction for each row, where that row was held out from model development. In expectation, the performance of an algorithm across these held out predictions converges to the generalization performance of the algorithm.

1.2.3 Bootstrap

The bootstrap approach is based upon the "plug in" principle, where we assume our sample follows some underlying distribution and approximate it by the empirical distribution of our sample [52]. Mathematically, we assign probability $p_i = \frac{1}{N}$ to each observation i in the design matrix $\mathbf{X}$, where N is the number of rows. We then draw samples from this distribution, equivalent to sampling from the original data with re-

placement, and denote this new dataset as $\mathbf{X}^*$. The model is developed on the resampled data. Note that the probability that an observation $\mathbf{x}_i$ is not selected can be calculated as $Pr[\mathbf{x}_i \notin \mathbf{X}^*] = (1 - p_i)^N \approx (1 - \frac{1}{N})^N = e^{-1}$. Predictions on this incidentally held out set, commonly called "out of bag" predictions, have been shown to be unbiased [37]. The approximately equals sign is due to the possible non-uniqueness of each observation, though this is very unlikely when multiple continuous features are present in the data.

1.3 Severity scores

The following severity scores are currently used for assessing the acuity of a general ICU population. The scores primary uses include end of life discussions, benchmarking, risk-adjustment and auditing purposes.

1.3.1 Acute Physiology and Chronic Health Evaluation

The APACHE system was originally published in 1981 by Knaus *et al.* at George Washington University [5]. The system consisted of a logistic regression with hospital mortality as the independent variable and a set dependent variables including comorbidities, age, gender and a newly defined Acute Physiology Score (APS). The APS synthesised the patient's physiological measurements into a single score by assigning each measurement a score dependent on the quantile it rested within. The scores were then summed into a single metric summarising the patient's overall physiology, with higher values representing increasing severity of illness. A simplified version of the system, which aimed to improve the clinical acceptability of the system, resulted in the APS II and the APACHE II system [53]. The latest version of APS is the APS III, published alongside the APACHE III score and the APACHE III probability model [7]. The APACHE III score is based upon the APS III, except with additional points given for comorbidities and age. The APACHE III probability model includes the APACHE III score as a covariate, in addition to many diagnostic indicator variables. The latest version of the APACHE system is APACHE IV [35], which retained the APS III as it was originally published [7]. APACHE IV was developed using data from 45 hospitals in the US, whereas the APS III was developed using data from 40 hospitals. APACHE

IV's fundamental architecture remained the same to previous iterations, and consisted of a multivariate logistic regression performed on a set of clinical and demographic parameters as well as the APS. Uniquely, APACHE IV modelled age, prior length of stay and the APS in a non-linear manner using cubic spline terms. APACHE IV also used separate regression models for the patient sub-population undergoing coronary artery bypass grafting (CABG) surgery. In total there are 22 variables required to calculate the APS, for a total of 142 variables in the non-CABG model. Note that 116 of these variables correspond to indicator Boolean parameters which specify the patient's diagnosis. In Chapter 4 a recalibration of APACHE IV is calculated, and the variables used in this recalibration correspond to those used in the non-CABG model.

1.3.2 Simplified Acute Physiology Score

The SAPS was intended as a simplification of the APS, reducing the number of physiological parameters required from the original 34 to 13 plus age [8]. The variables chosen were present for 90% of patients in the initial survey used to develop APACHE [5]. Though the original publication only provided ROC curves and not the AUROC, trapezoidal integration showed that SAPS had an AUROC of 0.7697 and APS had an AUROC of 0.7661. SAPS II, published in 1993 [9], aimed to rectify two issues with SAPS. First, the variable selection process in SAPS was done by clinical judgement, whereas SAPS II utilised univariate feature selection to filter out features uncorrelated with hospital mortality. Second, there was no model for calculating a probability of mortality from SAPS. SAPS II was published with calibration coefficients which allowed conversion of the integer score into a risk of mortality which ranged between zero and one. SAPS III is the latest model published and attempts to account for poor calibration of SAPS II found in later studies [54]. While SAPS II was developed on ICUs in western Europe, SAPS III included ICUs worldwide. The parameters of the model were identified using a stepwise logistic regression, followed by statistical hypothesis testing to ensure the parameters were significantly related to patient hospital mortality [55]. The final prediction used 61 binary indicator features which dichotomised various ranges of 20 distinct variables (e.g. the highest heart rate variable was split into three binary features, one

for heart rates < 120, one for $120 \leq$ heart rates < 160 and a final feature for heart rates ≥ 160). A hierarchical model was used, specifying patient characteristics as fixed effects and different ICUs as a random effect. A logistic regression equation was then used to calculate the probability of mortality.

1.3.3 Mortality Probability Model

The Mortality Probability Model (MPM) consists of two models of interest: one upon ICU admission (MPM_0-I) and one after 24 hours (MPM_{24}-I) [10]. These two variations have been developed concurrently and distinctly since the model's inception. The current version of the model, MPM_0-III, does not differ significantly from the prior versions and is essentially a recalibration of MPM_0-II [11] to a larger, more recent dataset of 74,578 patients [56]. The parameters for the model were selected if there existed a significant univariate relationship between the parameter and mortality, quantified as a Student's t-test p-value ≤ 0.2. Interaction terms were considered through a series of stepwise regressions. The final model, MPM_0-III, included 3 physiology variables, 3 chronic health indicators, 5 acute diagnoses, age, 5 binary indicator variables and 7 interaction terms between age and various other variables already present in the model. The advantage of this model is that a highly specific diagnosis is not required, and the prediction can be obtained after the first hour of ICU admission [56].

1.3.4 Intensive Care National Audit & Research Centre model

The ICNARC model was developed to improve risk prediction on United Kingdom ICU admissions [34]. Due to the reduced size, patient throughput and different admission criteria, UK ICUs tend to have more severely ill patients and as such models developed on US databases do not normally perform as well on UK databases [15]. The ICNARC model used stepwise backward feature removal with 1000 bootstrap repetitions performed at each step until a final model was developed. The data was repeatedly split into development and validation data sets each bootstrap repetition by randomly selecting one third of ICUs as the only contributors of data to the validation set. The model includes a physiology score similar to the APS used in the risk prediction, calculated by

	Pulmonary PaO_2/FiO_2‡	Renal Creatinine μmol/L	Hepatic Bilirubin μmol/L	Cardiovascular -	Hematologic Platelets 10^{-3}	Neurologic GCS
0	≥ 400	0-109	0-19	MAP† ≥ 70 No treatment	≥ 150	15
1	300-399	110-170	20-32	MAP† < 70 or dobutamine	100-149	13-14
2	200-299	171-299	33-101	Dopamine ≤ 5 (nor)epinephrine ≤ 0.1	50-99	10-12
3	100-199	300-440	102-204	Dopamine > 5 (nor)epinephrine > 0.1	20-49	6-9
4	0-99	>440	>204		0-19	<6

Table 1.5: *Severity score construction for SOFA. The organ system represented and primary physiologic parameters utilised (if applicable) are shown in the top two rows. The dysfunction for each organ system is usually assessed by the single physiological value listed and mapped into the five possible scores ranging from 0-4 (shown on the far left).*
‡Pressure of oxygen in the arteries over fraction of inspired oxygen.
† Mean arterial pressure measured in mmHg.

scaling and rounding the log odds from the regression models developed.

1.3.5 Sequential Organ Failure Assessment

The Sepsis-related Organ Failure Assessment score was first developed by a consensus meeting of the ESICM in October 1994, though it eventually became known as the Sequential Organ Failure Assessment (SOFA) as it was applied outside of septic populations [57]. The purpose of the score was to provide the clinical community with an objective measure of the severity of organ dysfunction in a patient. It is stressed that the score is not meant as a direct predictor of mortality but rather a measure of morbidity, or the level of the diseased state, in a patient. As such, direct comparisons to severity metrics presented in this report are not representative of the true design intent of the score. The score is evaluated as shown in Table 1.5 for 6 organ systems: pulmonary, renal, hepatic, cardiovascular, haematologic and neurologic. Note that organ-specific morbidity scores are provided.

	Patients			Performance		
Model	**Total**	**Training**	**Test**	AUROC	$HL_{\hat{C}}$	SMR
APACHE II	5,815	-	5,815 (100%)	0.863	-	-
APACHE III	17,440	8,720 (50.0%)	8,720 (50.0%)	0.900†	-	-
APACHE IV	110,558	66,335 (60.0%)	44,223 (40.0%)	0.880	16.80	0.997
SAPS	16,784	13,427 (80.0%)	3,357 (20.0%)	0.770	-	-
SAPS II	12,997	8,369 (64.4%)	4,628 (35.6%)	0.860	15.85	-
SAPS III	16,784	13,427 (80.0%)	3,357 (20.0%)	0.848†	14.29†	1.000†
MPM_0 II	19,124	12,610 (65.9%)	6,514 (34.1%)	0.824	11.40	-
MPM_{24} II	19,124	10,357 (54.1%)	5,568 (29.1%)	0.836	12.87	-
MPM_0 III	124,885	74,578 (59.7%)	50,307 (40.3%)	0.823	11.62	1.018
ICNARC‡	216,626	170,037 (79.5%)	46,589 (21.5%)	0.863	62.4	-

Table 1.6: *Comparison of severity scores most frequently used in the literature.*
†Optimistic due to the use of training data in the final evaluation.
‡Optimistic due to the use of a physiology score trained on final evaluation data.

Models	Observed mortality, %	Predicted mortality, %	SMR (95% CI)	AUROC	H-L
APACHE IV	13.51	13.55	0.997 (p=0.79)	0.880	16.8 (p=0.08)
SAPS III	17.7	17.7	1 (0.98-1.02)	0.848	14.29 (p=0.16)
MPM III	13.80	14.05	1.018 (0.996-1.040)	0.823	11.62 (p=0.174)
ICNARC	31.1	31.1	-	0.863	64.2 (p<0.001)

Table 1.7: *Comparison of the latest generation severity score performance, for both calibration and discrimination, on their respective validation data sets. 95% CI refer to the 95% confidence intervals.*

1.4 Severity score comparisons

1.4.1 Original performance

The following compares the results presented in each of the models' respective publishing. First, the size of the data sets used for development and validation are compared in Table 1.6. This is followed by the statistical evaluations of the final models on their respective validation data sets in Table 1.7.

1.4.2 External evaluation

Since their original publication, the presented severity scores have been analysed in many validation studies. A recent review by Strand *et al.* [58] surveyed the literature and provided performance comparisons. The inclusion criteria for a validation study was

Author, year	Country	Years	No. of patients	No. of ICUs	AUROC	SMR	HL
Schneider [59]	Australasia‡	2001-2010	636,428	190	0.842		
Harrison [60]	UK	1995-2003	141,106	163	0.804	1.20	2947.0
Brinkman [61]	Denmark	2006-2010	44,112	59	0.840		
Peek [62]	Netherlands	1999-2003	42,139	29	0.818	0.83	881.3
Beck [63]	UK	1993-1996	16,646	15	0.835	1.18	232.1
Ho [64]	Australia	1993-2003	11,107	1	0.846	0.84	189.3
Livingston [65]	Scotland	1995-1996	9,848	22	0.763	0.95	67.4
Markgraf [66]	Germany	1991-1994	2,661	1	0.832	1.07	11.8
Vassar [67] ⋆	US	1990-1991	2,414	6	0.870		92.6
Sakr [68]	Germany	2004-2005	1,851	1	0.800		1417.0
Duke [69]	Australia	2005-2007	1,843	1	0.820	0.52	
Bastos [70]	Brazil	1990-1991	1734	10	0.790	1.66	
Capuzzo [71]	Italy	1994-1997	1,721	2	0.805	0.98	5.1
Nouira [72]	Tunisia	1994-1995	1,325	3	0.820	1.19	26.0
Khwannimit [73]	Thailand	2004-2005	1,316	2	0.888	0.77	66.7
Ho [74]	Australia	2005	1,311	1	0.858		10.0
Beck [75]	UK	1993-1996	1144	1	0.806	1.23	98.6
Moreno [76]	Portugal	1994-1995	982	19	0.787	0.96	49.7
Kim [77]	Korea	2009	826	15	0.729	0.76	56.0
Katsaragakis [78]	Greece	1992-1997	661	1	0.839	1.14	18.1
Christensen [79]	Denmark	2007	469	1	0.730		13.7
Patel [80]	US	1996-1997	302	1	0.702	1.03	14.3

Table 1.8: *Comparison of studies which evaluated APACHE II in an independent sample.*
‡Refers to Australia and New Zealand only.
⋆ *Patients were admitted to trauma ICUs.*

a cohort with at least 100 patients, a general ICU population (i.e. no disease specific evaluation though specialised ICUs were permissible) and evaluation of the model on an independent dataset. This review has been augmented with other studies on severity score comparisons using the same inclusion criteria as Strand *et al.* and further including studies published after the publication of their review (2007). The following tables report the performance of various severity scores in validation studies: APACHE II is shown in Table 1.8, APACHE III is shown in Table 1.9, APACHE IV is shown in Table 1.10, SAPS II is shown in Table 1.11 and SAPS III is shown in Table 1.12. Publications which make use of the MPM models, for all three versions (MPM-I, MPM-II, MPM-III), are shown Table 1.13.

The most extensively evaluated iteration of APACHE was APACHE II. The discrimination of the model in the larger datasets ranged from 0.804-0.846. In UK ICUs the APACHE II predictions tended to be lower than observed patient mortality. APACHE III overall had higher discrimination in the large databases (AUROCs of 0.832-0.890

Author, year	Country	Years	No. of patients	No. of ICUs	AUROC	SMR	HL
Schneider [59]	Australasia‡	2001-2010	636,428	190	0.854		
Paul [81]	Australasia‡	2004-2009	152,456	147	0.885	0.84	1596.6
Harrison [60]	UK	1995-2004	141,107	164	0.832	1.36	10883.0
Zimmerman [82]	US	1993-1996	37,668	285	0.890	1.01	35.8
Beck [63]	UK	1993-1996	16,646	15	0.867	1.24	443.3
Shann [83]	Australia	2005-2006	16,356	21	0.880	0.85	28.9
Livingston [65]	Scotland	1995-1996	10,326	22	0.795	1.23	365.7
Markgraf [66]	Germany	1991-1994	2,661	1	0.846	1.23	48.4
Keegan [84]	US	2006	2,596	3	0.868	0.66	33.7
Vassar [67] ⋆	US	1990-1991	2,414	6	0.890		7.0
Duke [69]	Australia	2005-2007	1,843	1	0.910	0.66	
Bastos [70]	Brazil	1990-1991	1734	10	0.820	1.67	400.3
Beck [75]	UK	1993-1996	1144	1	0.847	1.36	129.8
Pettila [85]	Finland	1995	520	1	0.825	1.62	9.3

Table 1.9: *Comparison of studies which evaluated APACHE III in an independent sample.*
‡Refers to Australia and New Zealand only.
⋆ *Patients were admitted to trauma ICUs.*

Author, year	Country	Years	No. of patients	No. of ICUs	AUROC	SMR	HL
Brinkman [61]	Denmark	2006-2009	55,661	59	0.870	0.87	822.7
Kuzniewicz [86]	US	1999-2003	11,300	35	0.892	1.03	22.4
Keegan [84]	US	2006	2,596	3	0.861	0.74	31.0

Table 1.10: *Comparison of studies which evaluated APACHE IV in an independent sample.*
‡Refers to Australia and New Zealand only.

Author, year	Country	Years	No. of patients	No. of ICUs	AUROC	SMR	HL
Harrison [60]	UK	1995-2005	141,108	165	0.822	1.10	2664.0
Brinkman [61]	Denmark	2006-2011	44,112	59	0.850		
Peek [62]	Netherlands	1999-2003	42,139	29	0.831	0.90	879.4
LeGall [87]	France	1998-1999	38,745	106	0.858	0.84	1162.9
Reiter [88]	Austria	1998-2001	30,099	31	0.870	0.93	290.1
Metnitz [54]	Global	2002	16,784	303	0.830	1.10	184.70
Beck [63]	UK	1993-1996	16,646	15	0.852	1.17	287.5
Aegerter [89]	France	1999-2000	13739	32	0.870	0.79	
Kuzniewicz [86]	USA	1999-2003	11,300	35	0.873	1.04	18.1
Livingston [65]	Scotland	1995-1996	10,334	22	0.784	0.97	142.0
Haaland [90]	Norway	2008-2010	10,135	42	0.830	0.73	689.1
Moreno [91]	Europe	1994-1995	10,027	89	0.822	0.90	208.4
Poole [92]	Italy	2007	3,661	103	0.830	0.87	
Metnitz [93]	Austria	1997-1998	2,901	13	0.830	0.90	69.1
Markgraf [66]	Germany	1991-1994	2,661	1	0.846	1.13	20.5
Strand [94]	Norway	2006-2007	1,873	2	0.820	0.82	27.4
Sakr [68]	Germany	2004-2005	1,851	1	0.830		452.0
Duke [69]	Australia	2005-2007	1,843	1	0.870	0.91	
Metnitz [95]	Austria	1997	1,733	9	0.810	0.85	85.7
Capuzzo [71]	Italy	1994-1997	1,721	2	0.816	0.89	9.3
Apolone [96]	Italy	1994	1393	99	0.800	1.14	71.0
Nouira [72]	Tunisia	1994-1995	1,325	3	0.840	1.27	73.8
Khwannimit [73]	Thailand	2004-2005	1,316	2	0.911	0.77	71.4
Moreno [76]	Portugal	1994-1995	982	19	0.817	0.98	28.3
Soares [97] *	Brazil	2003-2005	952	1	0.880	1.28	32.1
Capuzzo [98]	Italy	2006-2007	684	2	0.851	0.89	12.1
Katsaragakis [78]	Greece	1992-1997	661	1	0.870	1.62	60.5
Lim [99]	Korea	2008-2009	633	1	0.760	0.84	23.5
Christensen [79]	Denmark	2007	469	1	0.740		4.4
Patel [80]	US	1996-1999	304	3	0.672	1.10	22.6

Table 1.11: *Comparison of studies which evaluated SAPS II in an independent sample.*
* *Evaluated in a surgical-oncologic ICU.*

Author, year	Country	Years	No. of patients	No. of ICUs	AUROC	SMR	HL
Poole [100]	Italy	2007	28,357	147	0.855	0.73	
Poole [92]	Italy	2007	3,661	103	0.830	0.63	
Keegan [84]	US	2006	2,596	3	0.801	0.66	36.6
Metnitz [101]	Austria	2006-2007	2,060	22	0.820	0.79	90.3
Khwannimit [102]	Thailand	2007-2009	1,873	1	0.933	0.86	101.2
Strand [94]	Norway	2006-2007	1,873	2	0.810	0.71	17.4
Sakr [68]	Germany	2004-2005	1,851	1	0.840		208.5
Duke [69]	Australia	2005-2007	1,843	1	0.880	0.85	
Soares [97] *	Brazil	2003-2005	952	1	0.870	1.19	13.6
Capuzzo [98]	Italy	2006-2007	684	2	0.835	0.83	22.4
Lim [99]	Korea	2008-2009	633	1	0.780	0.74	3.2
Christensen [79]	Denmark	2007	469	1	0.690		9.2
Khwannimit [102]*	Thailand	2007-2009	1,873	1	0.933	0.92	96.2
Duke [69]*	Australia	2005-2007	1,843	1	0.880	0.93	
Lim [99]*	Korea	2008-2009	633	1	0.780	0.79	3.3
Silva Junior [103]†	Brazil	2008-2009	1,310	2	0.860	1.04	10.5
Soares [97] *†	Brazil	2003-2005	952	1	0.870	0.95	9.1
Strand [94]‡	Norway	2006-2007	1,873	2	0.810	0.74	18.3
Poole [100]$^{\triangle}$	Italy	2007	28,357	147	0.860	0.73	

Table 1.12: *Comparison of studies which evaluated SAPS III in an independent sample. A subset of studies used coefficients specific to geographic regions which were published by the authors of SAPS III. Unless otherwise stated, the general SAPS III coefficients are used in the studies.*

* *Australian coefficients were used.*

* *Evaluated in a surgical-oncologic ICU.*

† *Central and South American coefficients were used.*

‡ *Northern European coefficients were used.*

$^{\triangle}$ *Mediterranean coefficients were used.*

Author, year	MPM	Country	Years	No. of patients	No. of ICUs	AUROC	SMR	HL
Moreno [91]	I	Europe	1994-1995	10,027	89	0.785	0.85	368.2
Nouira [72]	I	Tunisia	1994-1995	1,325	3	0.850	0.91	36.7
Duke [69]	II	Australia	2005-2007	1,843	1	0.860	0.82	
Peek [62]	II	Netherlands	1999-2003	42,139	29	0.796	0.99	371.1
Livingston [65]	II	Scotland	1995-1996	10,393	22	0.741	1.00	451.9
Rue [104]	II	Spain	1995	1,441	15	0.800	0.90	67.9
Kuzniewicz [86]	III	US	1999-2003	11,300	35	0.809	1.04	9.8
Keegan [84]	III	US	2006	2,596	3	0.721	0.78	21.8
Nouira [72]	I	Tunisia	1994-1995	1,325	3	0.880	1.09	29.6
Harrison [60]	II‡	UK	1995-2006	141,109	166	0.815	1.11	1598.0
Peek [62]	II	Netherlands	1999-2003	42,139	29	0.822	0.90	206.1
Livingston [65]	II	Scotland	1995-1996	7,343	22	0.791	1.02	100.8
Rue [104]	II	Spain	1995	1,441	15	0.840	0.91	28.3
Patel [80]	II	US	1996-1998	303	2	0.695	1.02	20.7

Table 1.13: *Comparison of studies which evaluated MPM_0 (above the double lines) or MPM_{24} (below the double lines) in an independent sample. The version of MPM used is shown for each publication. Note that there is no third incarnation of the MPM_{24} model.*
‡These authors evaluated MPM_0 for patients staying less than 24 hours and MPM_{24} for patients staying longer than 24 hours.

except for Scotland where the AUROC was 0.795) but also tended to underpredict mortality in UK ICUs. The AUROC was highest in the hospitals which had earlier participated in the data collection (AUROC of 0.890) [82]. Finally, APACHE IV had the highest range of discrimination with AUROCs of 0.861-0.892. In two studies APACHE IV underpredicted mortality but retained excellent calibration in the third.

SAPS II had a good range of discrimination in studies with more than 1,000 patients (AUROC of 0.784-0.911) and this good discrimination was also evident in studies with more than 10,000 patients (AUROC of 0.784-0.873). Among the larger studies, SAPS II tended to overpredict patient mortality, as expected due to model drift, except in the UK where it underpredicted mortality. SAPS III had only one independent study evaluating the model on more than 10,000 patients. The AUROC reported in this study was 0.855 and the model underpredicted mortality (SMR < 1). Relaxing the evaluation requirement to studies with more than 1,000 patients, we find a reasonable range of AUROCs between 0.801-0.933. Whereas subsequent iterations of the APACHE system appeared to have better discrimination, there does not seem to be a consistent shift upward between SAPS II and SAPS III, though a comprehensive systematic review

would be needed to confirm this proposition. The calibration of SAPS III indicated a large bias towards underpredicting mortality (between 15-37%) in larger studies. The likely explanation for this is due to differing patient demographics, and likely not model drift, as the data in the studies was collected within 2-3 years of the original SAPS III publication.

The MPM models have had much fewer independent evaluations as compared to the SAPS and APACHE models. Both MPM_0-II and MPM_{24}-II had an extensive evaluation in UK hospitals with a database of 141,109 patients [34]. In this study MPM had an AUROC of 0.815 and underpredicted mortality as all other risk prediction models did in the UK. MPM-II was the most extensively evaluated and has a reported AUROC of 0.741-0.860 for the admission model and 0.791-0.840 for the 24 hour model (excluding the low sample size study [105]). The MPM_0-III model had reasonable discrimination of 0.721-0.809 given its limited window for data collection.

Overall, it appears that the performance of a risk prediction model varies due to the changing study populations and changing risk prediction model. Overall the APACHE IV model had the highest range of AUROCs, though was only evaluated in three studies. SAPS II had a consistently high discrimination given the large number of diverse locations it was evaluated. The AUROCs for US hospitals was 0.873 for the SAPS II model, 0.801 for the SAPS III model, 0.870 for the APACHE II model and 0.861-0.892 for APACHE IV (excluding a single low sample size study [105]). Overall, it appears an AUROCs of 0.873-0.892 can be considered state of the art performance for US hospitals. For UK hospitals, the best risk adjustment performance is likely achieved by the ICNARC model. Among the models presented here however, state of the art appears to be an AUROC between 0.804-0.867. These studies provide the needed context for assessing the discrimination achieved by models in this thesis, and facilitates their comparison to the performance already obtained by models which have been clinically validated.

Chapter 2

Preprocessing and machine learning in the intensive care unit

> *On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.*
>
> Charles Babbage

A key aspect of model development is data preprocessing. Preprocessing is performed to accomplish a wide variety of goals: to transform data to a more convenient distribution, to remove undesirable noise from a signal (filtering), decorrelate features (whitening) or to remove extreme values which occur due to measurement artefact. In this work, preprocessing refers to the identification and removal of values present in the data that are hypothesised to be artefacts and not actually reflect patient health, i.e. outliers. For example, a pH value of 810 is almost certainly an artefact, as pH has a plausible physiological range of 6.5 - 8.0. The clinical impetus for preprocessing data in this fashion is the frequent occurrence of errors in medical data, either due to humans or sensors [106].

As most prediction models utilised for prediction of ICU mortality are based off a linear combination of the data, extreme values will have a large impact on the performance

of these models. In the past, a large proportion of data collected for the development and evaluation of severity scores was done so by trained personnel. This process can be considered as a "human filter", rejecting unreasonable values during data collection. However, human data collection and annotation is a labour intensive task, and an equivalently effective automatic preprocessing technique would be worthwhile. Furthermore, given the increased digitisation of hospitals, there is an increasing possibility of automatically extracting data from clinical data management systems for use in predictive models. The use of electronic data management systems as training databases is secondary to their core purpose, and preprocessing becomes increasingly important in this framework. The first goal of this chapter is to evaluate the benefit of automated outlier rejection on mortality prediction models.

Modern mortality risk models have two components: i) a single feature which synthesises physiology in a non-linear fashion (e.g. APS III), and ii) an additional set of features not directly based upon physiology (e.g. patient admission urgency, diagnosis). These two components have traditionally been combined linearly (e.g. APACHE III). Yet, more recent approaches have incorporated non-linear terms, such as cubic splines in APACHE IV and interaction terms in the ICNARC model. The second goal of this chapter is to quantify the performance of machine learning models which are capable of implicitly modelling a non-linear mapping from the data to risk of mortality.

2.1 Physionet/Computing in Cardiology 2012 Challenge database ($\mathbf{PN_{db}}$)

Each year PhysioNet[1] [107], in cooperation with the annual Computing in Cardiology conference, announces an international "challenge" which aims to encourage research into clinically relevant problems which have not been well solved or are worthy of further research. The PhysioNet/Computing in Cardiology 2012 challenge focused on in-hospital mortality prediction for ICU patients, and data corresponding to 4,000 anonymised patient records in the ICU were released to the general public on PhysioNet with corre-

[1] http://www.physionet.org

sponding patient outcomes (length of ICU stay, number of days between ICU admission and death and in-hospital mortality) [108]. An additional set of 4,000 patient records were also released, however the outcomes for this second set were not made publicly available and model performance on this set is only available upon contacting organisers of the PhysioNet/Computing in Cardiology 2012 challenge [108]. The first set of data available with outcomes is referred to as set a (PN_a) and the second set of data available without outcomes is referred to as set b (PN_b). All patients were admitted to an ICU at the Beth Israel Deaconess Medical Center in Boston, USA between 2001 and 2008. Each row of a patient record file contains: a time value in minutes since admission, a descriptor indicating the variable measured and the value of the variable measured. Records pertain to the first 48 hours of a patient's ICU stay, and all records strictly ended at 48 hours post ICU admission. Consequently all patients both stayed a minimum of 48 hours and survived to 48 hours. Given the openly available nature of the dataset and the relevance to clinical prediction it provides an excellent platform for evaluating models for predicting in-hospital mortality.

2.1.1 Variable description

A total of 37 physiologic variables and 5 static descriptor variables (i.e. one value per patient) were present in the dataset. For PN_a, an additional 6 outcomes of interest were provided. A list of the variables available as a time series is shown in Table 2.1, while the remaining descriptor and outcome variables are shown in Table 2.2.

An example of the time-series measurements available is shown in Figure 2.1. As the figure shows, the frequency of measurement varies between hourly observations (such as blood pressure and heart rate) to daily observations (such as most lab values).

2.1.2 Feature extraction

The majority of data available is in the form of a time-series of measurements. An example of heart rate and blood pressure measurements from a single patient is shown in Figure 2.2. Information contained in these time series, including variability or trend based metrics, could allow for more accurate predictions for individual patients.

Variable	Description (Unit)
Albumin	Serum Albumin (g/dL)
ALP	Alkaline phosphatase (IU/L)
ALT	Alanine transaminase (IU/L)
AST	Aspartate transaminase (IU/L)
Bilirubin	Serum Bilirubin (mg/dL)
BUN	Blood urea nitrogen (mg/dL)
Cholesterol	Serum Cholesterol (mg/dL)
Creatinine	Serum creatinine (mg/dL)
DiasABP	Invasive diastolic arterial blood pressure (mmHg)
FiO2	Fraction of inspired oxygen (0-1)
GCS	Glasgow Coma Scale (3-15)
Glucose	Serum glucose (mg/dL)
HCO3	Serum bicarbonate (mmol/L)
HCT	Hematocrit (%)
HR	Heart rate (beats per minute, bpm)
K	Serum potassium (mEq/L)
Lactate	Serum Lactate (mmol/L)
Mg	Serum magnesium (mmol/L)
MAP	Invasive mean arterial blood pressure (mmHg)
MechVent	Mechanically ventilated (0:false or 1:true)
Na	Serum sodium (mEq/L)
NIDiasABP	Non-invasive diastolic arterial blood pressure (mmHg)
NIMAP	Non-invasive mean arterial blood pressure (mmHg)
NISysABP	Non-invasive systolic arterial blood pressure (mmHg)
PaCO2	Partial pressure of arterial CO_2 (mmHg)
PaO2	Partial pressure of arterial O_2 (mmHg)
pH	Arterial pH (0-14)
Platelets	Platelet count (cells/nL)
RespRate	Respiration rate (breaths per minute, bpm)
SaO2	Oxygen saturation in hemoglobin (%)
SysABP	Invasive systolic arterial blood pressure (mmHg)
TropI	Troponin-I (μg/L)
TropT	Troponin-T (μg/L)
Urine	Urine output (mL)
Temp	Temperature (°C)
WBC	White blood cell count (cells/nL)
Weight[1]	Measured or estimated weight (kg)

Table 2.1: *Variables with available time-series measurements in the PN_{db}.*
[1]Note weight is both a temporal variable (for estimating fluid balance) and a demographic variable.

Variable	Description (Unit)
RecordID	Unique value for each patient (unitless integer)
Age	Age at admission (years)
Gender	If the patient is female (binary)
Height	Measured or estimated height (cm)
Weight[1]	Measured or estimated weight (kg)
ICUType	Coronary Care Unit, Cardiac Surgery Recovery Unit, Medical ICU or Surgical ICU
In hospital mortality	Binary flag
In ICU mortality	Binary flag
Days until mortality	Positive integer[2]
Length of stay	Fractional days
SAPS-I	Simplified Acute Physiology Score [8]
SOFA	Sequential Organ Failure Assessment [57]

Table 2.2: *Descriptive and outcome variables available in the PN_{db}.*
[1]Note weight is both a temporal variable (for estimating fluid balance) and a static variable recorded once at admission.
[2]Missing values indicate patient did not die after hospital discharge.

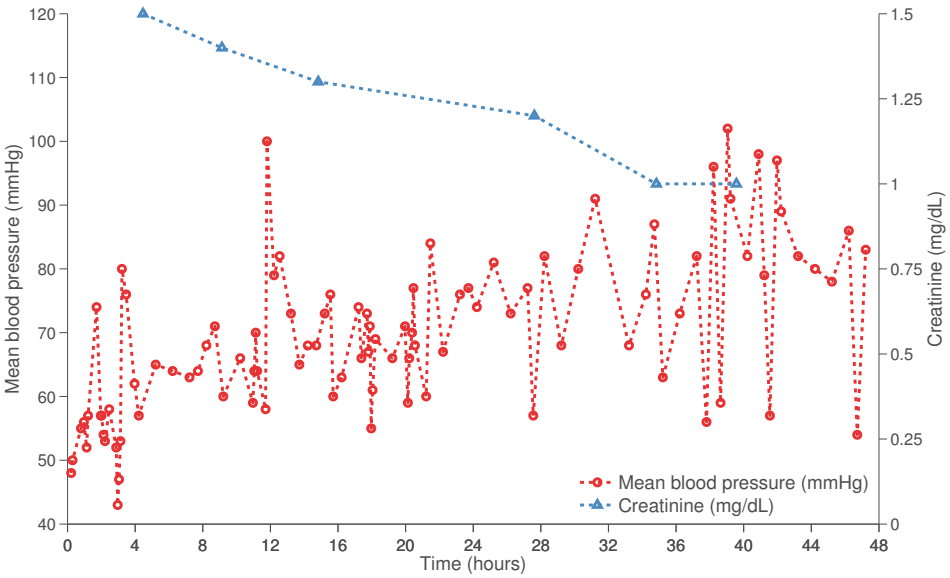

Figure 2.1: *Example of time-series measurements available over the first 24 hours for patient record 132540. The patient's mean blood pressure and serum creatinine levels are shown across the first 48 hours.*

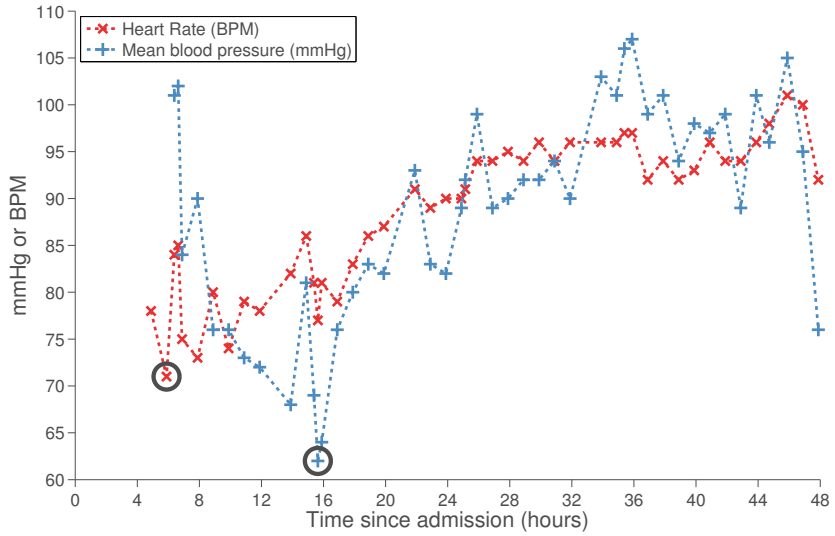

Figure 2.2: *Example data for patient record 132547. The mean blood pressure and heart rate are seen to increase, indicating increased cardiac output. The values circled are the worst values in the first 24 hours, and would be the only inputs from the time series in a severity of illness score such as the APS or SAPS.*

In order to capture information of the patient's entire 48 hour stay the median, highest, lowest, first and last measurement for the time-series variables were extracted. This was the generic method of feature extraction used for the time-series variables.

Three time-series features were extracted in a unique fashion due to the nature of the measurement. For urine output, The sum of all measurements was extracted as the total urine output of a patient is of more clinical relevance than the individual measurements. As troponin-T and troponin-I are very infrequently measured, an additional single feature was extracted which indicated if a measurement for either troponin was present for the given patient. Finally, the mechanical ventilation variable was parsed specially, but with a similar intention as the generic feature extraction. The five features extracted were: presence of mechanical ventilation, presence of mechanical ventilation in the first four hours, presence of mechanical ventilation in the last four hours (hours 44-48), the time elapsed before the initial mechanical ventilation and the duration of mechanical ventilation. The duration of mechanical ventilation is estimated from the first and last instances of mechanical ventilation and is consequently an approximation.

A set of static variables including age, weight (on admission), height and gender were added to the design matrix. Note that the weight variable occurs twice: once as an admission variable and again as a time-series measurement (as it is used to monitor

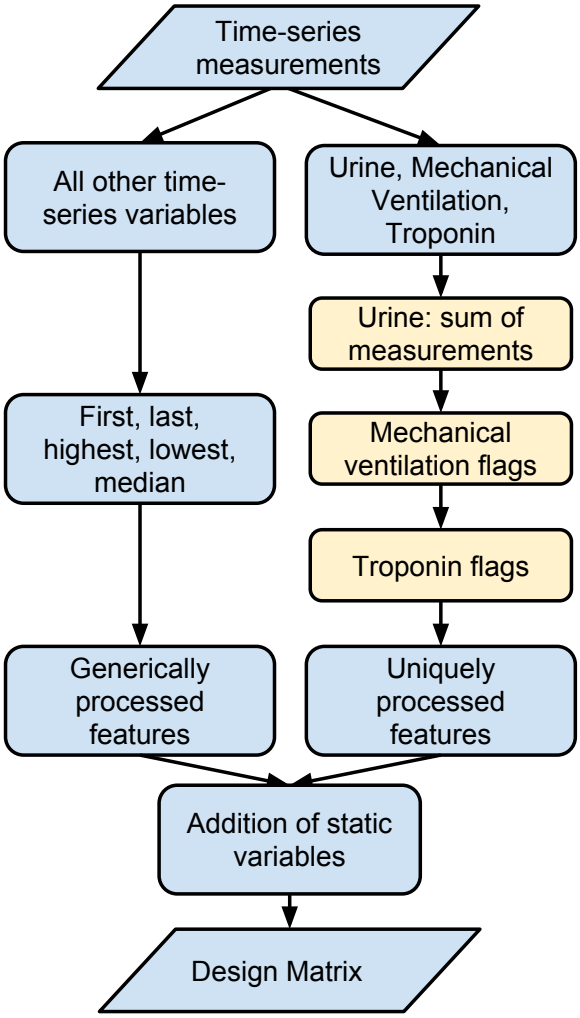

Figure 2.3: *Overview of data extraction performed for most time-series features (left) and for special cases (right, detail in text).*

fluid balance). Finally, the type of care unit the patient was admitted to was converted into a set of four binary indicator variables, each of which had value 1 if and only if the patient was admitted to the corresponding care unit. Thus for each patient only one of the medical unit, cardiac surgery recovery unit, surgical unit or coronary care unit features were positive.

After concatenating these features, the final number of features used in the model development was 198. A visualization of the data extraction, with particular emphasis on the unique feature extractions performed, is shown in Figure 2.3. The result of this process is a design matrix $\mathbf{X}$ which contains data to be used for predicting patient in-hospital mortality.

2.2 Box-Cox outlier rejection

The method introduced in this section aims to: i) remove artefactual observations from the input features and ii) transform variables to more resemble a normal distribution. Implicit in the second aim is the hypothesis that the linear relationships learned between independent variables and the dependent variable are more predictive if the independent variables are normally distributed, a hypothesis which must be empirically evaluated.

Box-Cox Outlier Rejection (BCOR) is a preprocessing technique designed to remove artefactual data. The algorithm as proposed in this thesis was developed by the author and has been previously applied successfully to develop and evaluate mortality prediction models [109, 110]. The BCOR process proceeds iteratively and univariately. Each feature is Box-Cox transformed [111] to increase its similarity to a normal distribution. Thresholds are then determined using a critical value at the 0.01 significance level ($\alpha = 0.01$) with application of the Bonferroni correction [112]. Formally, given a data vector $\mathbf{x}$, the Box-Cox transform is as follows:

$$\tilde{\mathbf{x}} = \begin{cases} \frac{\mathbf{x}^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(\mathbf{x}) & \lambda_1 = 0. \end{cases} \tag{2.1}$$

where $\tilde{\mathbf{x}}$ is the transformed value of $\mathbf{x}$. The optimal choice for the parameter λ_1 is determined in this work using a maximum likelihood approach [111]. If we define $\tilde{\mu}$ as the mean of the vector $\tilde{\mathbf{x}}$ and $\tilde{\sigma}$ as the standard deviation of the vector $\tilde{\mathbf{x}}$ then we select λ_1 as follows:

$$\hat{\lambda}_1 = \underset{\lambda_1}{\operatorname{argmax}} \frac{N}{2} \log(\tilde{\sigma}^2) + (\lambda_1 - 1) \sum_{i=1}^{N} \log(\tilde{x}_i). \tag{2.2}$$

where $\hat{\lambda}_1$ is the maximum likelihood estimate for λ_1.

Note that Equation 2.1 requires $\mathbf{x} > 0$. If this is not satisfied, then the transform is trivially extended to:

$$\tilde{\mathbf{x}} = \begin{cases} \frac{(\mathbf{x} + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(\mathbf{x} + \lambda_2) & \lambda_1 = 0. \end{cases} \tag{2.3}$$

where λ_2 is appropriately selected to ensure $\mathbf{x} + \lambda_2 > 0$.

Once transformed, two thresholds used for outlier detection are calculated. The lower threshold l and the upper threshold h are defined as:

$$\begin{aligned} l &= \Phi^{-1}\left(\frac{\alpha}{2N}\right), \quad \Phi \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}) \\ h &= \Phi^{-1}\left(1 - \frac{\alpha}{2N}\right), \quad \Phi \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}) \end{aligned} \tag{2.4}$$

where $\tilde{\mu}$ is the mean of the transformed data $\tilde{\mathbf{x}}$, $\tilde{\sigma}$ is the standard deviation of $\tilde{\mathbf{x}}$, $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for the normal distribution defined by Φ and N is the number of data points observed. Data not residing within the thresholds, i.e. data that is deemed to be too extreme, are replaced with a missing value. This removes values which, given the number of samples, are extremely unlikely to have been drawn from the overall distribution. For each feature, the Box-Cox transform and threshold calculation steps are repeated until no values are removed. All transformation parameters and associated thresholds are saved for later application to the validation and test set. Binary, nominal and ordinal variables are not preprocessed in this manner.

2.3 Models

The most commonly employed clinical risk-adjustment model is logistic regression. However, there are many more sophisticated techniques capable of capturing higher order interactions in the data. The models selected for comparison in this work were Support Vector Machines (SVMs), Random Forests (RFs), Regularized Logistic Regressions (RLRs) and Regularized Logistic Regressions with square terms (RLR^2s). In this section a brief overviews of each model is provided.

2.3.1 Regularised logistic regression

Logistic regression aims to maximise the conditional likelihood of the observations given the data. The target, or outcome, is assumed to follow a binomial distribution. Given a

set of coefficients β, a vector of predictions $\hat{\mathbf{y}}$ can be generated as follows:

$$\hat{\mathbf{y}} = \frac{1}{1 - e^{\mathbf{X}\beta^T}}, \tag{2.5}$$

where β^T is the transpose of β. The log likelihood of the targets given these predictions can be written as:

$$\log(\mathcal{L}(\hat{\mathbf{y}}; \mathbf{y})) = \sum_{i=1}^{N} \left(y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i) \right), \tag{2.6}$$

where y_i is the observed outcome for the i^{th} row of the design matrix and $\hat{y}_i$ is the predicted outcome for the i^{th} row of the design matrix. Note this is the log likelihood assuming $\mathbf{y}$ is a Bernoulli random variable (and identical to the $\mathcal{I}_{\mathcal{L}}$ introduced in Chapter 1).

Maximising the log-likelihood is usually replaced with minimising the negative log-likelihood for computational convenience. This minimisation can be performed using gradient descent methods.

In general, one often faces the issue of additional model complexity increasing performance on the training data but reducing model generalisation performance. One technique which can counteract this phenomenon is regularisation. In the context of logistic regression, regularisation refers to the addition of a term to the likelihood function which is dependent on the coefficient values. The general form of the log likelihood function with an added penalty term can be written as follows:

$$\log(\mathcal{L}(\hat{\mathbf{y}}; \mathbf{y})) = \sum_{i=1}^{N} \left(y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i) \right) + \lambda f(\beta), \tag{2.7}$$

where β represents a vector of coefficient values and λ is a hyperparameter which controls how strongly regularised the model is. As $f(\beta)$ is usually chosen to be a positive monotonic function of β, large values of lambda force all model coefficients (β) to zero. In this situation, any reduction in the negative log likelihood due to predictively useful covariates would be outweighed by the increase due to $\lambda f(\beta)$. Conversely, a value of zero for λ implies no constraint on the model coefficients.

Certain functional forms of $f(\beta)$ are common. L^2 regularization refers to when

the penalty term is the sum of the square of each coefficient (i.e. the L^2 norm). L^1 regularisation refers to when the penalty term consists of the sum of the absolute value of each coefficient (i.e. the L^1 norm). While both L^1 and L^2 regularisation ensure the coefficient values are not excessively large, L^1 regularisation tends to produce models which have more coefficients equal to zero as compared to L^2 regularisation, i.e. L^1 regularisation tends to produce sparser models. See for example Bishop *et. al* [50] for a good tutorial on the difference between L^1 and L^2 regularisation. Note that the intercept term is not penalised in the regularisation term.

Regularized logistic regression can be solved in a similar way to logistic regression using gradient descent methods or non-linear programming approaches [113]. In this work, a faster algorithm referred to as the least absolute shrinkage and selection operator (LASSO) is used to learn the coefficients [113]. This method involves restating the problem as minimising the negative log likelihood subject to the condition that $\sum f(\beta) \leq \alpha$. It can be shown that optimising the negative log likelihood with this constraint and minimising Equation 2.7 are equivalent, and the reformulation is performed for primarily computational gain.

2.3.2 Regularised logistic regression with square terms

As logistic regression only applies linear transformations to each individual feature, it will fail to capture potentially non-linear, higher order interactions. An obvious example of such is the belief that extreme values are indicative of abnormality. Hypotension, or blood pressure which is too low, is an indicator of severity of illness for patients with sepsis [114]. Conversely, hypertension is a common risk factor for cardiovascular disease [115]. Modelling different risks for the lower and upper tails of a distribution with a linear model can be accomplished by the addition of a square term for the given covariate. The additional square term allows for a single point of inflection, and thus both low and high values for the given covariate can be assigned a high risk[2]. This is illustrated in Figure 2.4.

A straightforward extension to the addition of square terms would be the inclusion

[2]It is also possible to assign low risk to the extreme values, though this is less intuitive clinically.

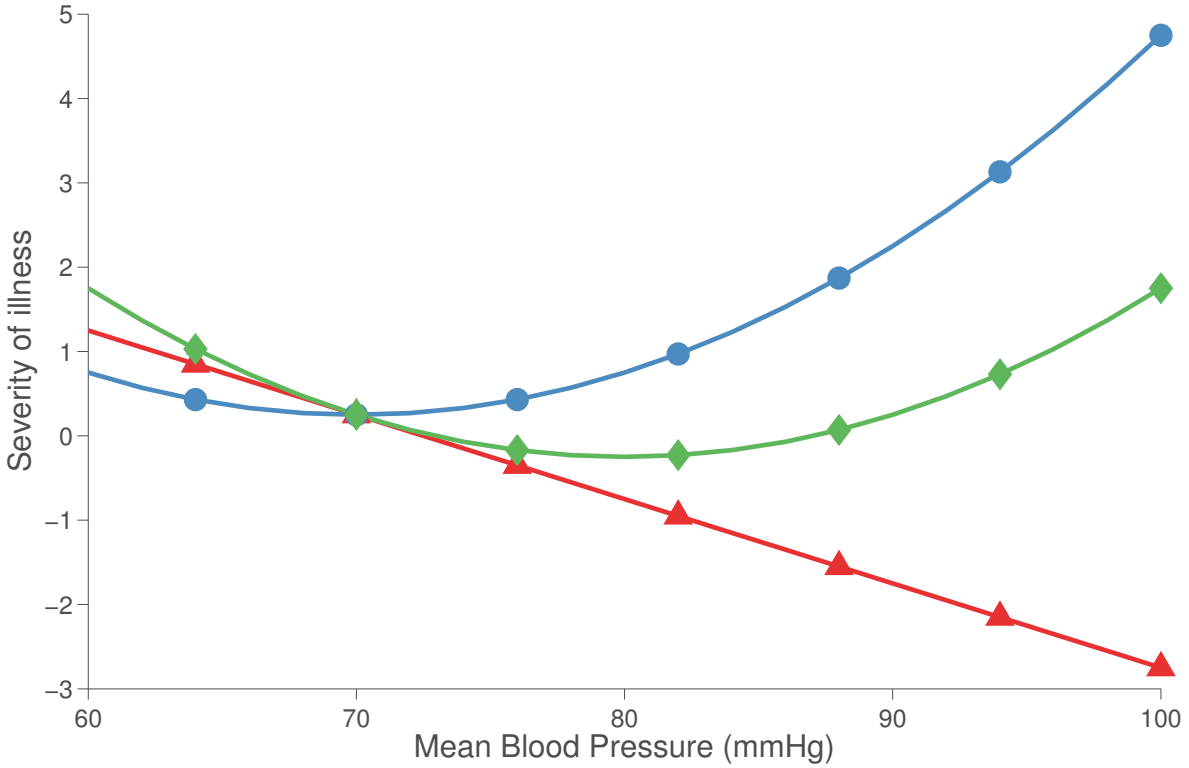

Figure 2.4: *Example of a regression using a single feature: the patient's mean blood pressure. The red line with triangles represents an increasing risk of mortality proportional to blood pressure. The green line with diamonds represents a decreasing risk of mortality proportional to the blood pressure squared. The green line with diamonds represents the addition of these two contributions. There appears to be a "normal" region with a low estimate of severity, and excursions in either direction from this region of normality increase the severity of illness.*

of interaction terms between various covariates. One clinically used example of these interactions is an approximation of cardiac output, which involves multiplying heart rate with blood pressure. While the addition of square terms increases the regression model's dimensionality from $D + 1$ to $2D + 1$, the addition of interaction terms would increase the dimensionality to $\binom{D}{2} = \frac{D!}{2!(D-2)!}$. As D^2 grows very quickly as D increases, and since it is desirable to have a large number of observations for each parameter in a model to prevent overfitting, the inclusion of these terms would require an unreasonably large amount of data. These reasons motivated the inclusion of square terms, but not interaction terms. RLR^2 was treated as a distinct model in the empirical evaluation in this and subsequent chapters. Note that the squaring of terms is performed after mean normalisation in order to ensure that the features are not perfectly correlated (which would cause the optimisation technique to select only one of the features).

2.3.3 Random forest

RFs are tree based classifiers first detailed by Brieman *et al.* in 2001 [37]. The model is based on the principle of combining weak learners into a single strong classifier, where a weak learner is typically a simple model which is predictive but not very accurate. In the context of an RF, the weak learner is a decision tree, which have the desirable property of high variance but low bias. That is, the predictive value of the trees varies but there is no systematic component consistent across all the trees. There are two key aspects to RFs: the use of bootstrap resampling to generate approximately independent trees, and an improved implementation of "bagging" (bootstrap aggregation) to reduce the variance of the average prediction across all trees.

Each tree in a RF is trained using a resampled version of the training dataset (with replacement). This resampling procedure, known as bootstrapping (see Section 1.2.3), is equivalent to approximating the population cumulative distribution function with an empirical distribution function that assigns $\frac{1}{N}$ probability to each observed data point. That is, a new matrix $\mathbf{X}^*$ of size N is generated from $\mathbf{X}$ with $P(x_k^* = x_j) = \frac{1}{N}$ where N is the number of rows in $\mathbf{X}$. Sampling from the empirical cumulative distribution in this way provides a good approximation to sampling from the actual population distribution. This process creates new design matrices $\mathbf{X}^*$ drawn from the distribution of the training data $\mathbf{X}$, and these individual design matrices are independently used to train decision trees T.

RFs have an additional subsampling for the features included in the bootstrap sample $\mathbf{X}^*$. A user specified number of features, m, are selected contemporaneously to the bootstrap resampling. Only these features and the bootstrap observations are used to develop the decision tree. The bootstrap resampling process is repeated for each tree B times, resulting in B trees $T_b(\mathbf{X})$.

In order to determine the cut points for a single split in an individual tree $T_b(\mathbf{X})$, some measure of performance is needed for the two resultant leaf nodes. In this work the Gini index is used for this purpose. The Gini index is calculated as follows:

$$G(x) = \sum_{i=1}^{N} x_i(1 - x_i) \tag{2.8}$$

for $n = 1 \rightarrow N_t$ **do**
  Sample from the data $\mathbf{X}$ with replacement to generate $\mathbf{X}^*$
  Generate a tree T_b as follows:
  repeat
    Select m of D variables from $\mathbf{X}^*$
    Determine the split that minimizes the Gini index for these variables
    Split the data into two daughter nodes
  until $N_n \leq N_m$
end for

Figure 2.5: *Pseudocode for the RF algorithm. The Gini index is a measure of agreement (see Equation 2.8). N_n is the number of observations in the leaf nodes. N_m is a user specified minimum number of nodes for any leaf node in the forest. N_t is the user defined number of trees in the forest.*

where $G(x)$ represents the frequency at which an element x_i would be incorrectly classified using the prevalence of the classes in the given set. It is equivalent to the null Brier score described in Section 1.1.7 applied to the observations only in the current leaf. The splits in a single tree of the RF are selected to minimise $G(x)$. The result of the RF training procedure is a committee of (ideally) uncorrelated trees which vote on the most likely class label to assign a given observation. The final output of a RF for classification is:

$$\hat{\mathbf{y}} = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{X}). \tag{2.9}$$

where $T_b(\mathbf{X})$ is the b^{th} tree.

The overall procedure for training a RF is shown in Figure 2.5.

2.3.4 Support vector machine

SVMs were originally proposed by Cortes and Vapnik [36]. SVMs are based upon statistical learning theory [116, 117], which laid the theoretical foundation for evaluating both a learning algorithm's capacity (its ability to perfectly classify an arbitrary number of points) and performance (generalisation accuracy as the number of test set points tends to infinity). The notation used to describe the SVM in this work largely follows that of Cortes, Vapnik and Burges [36, 118]. Assume there are N pairs of x_i (a set of design matrix features)[3] and y_i (the outcome of interest) with $x_i \in \mathbb{R}^D$ and $y_i \in \{-1, +1\}$.

[3]Recall that row vectors are not bolded to distinguish them from column vectors.

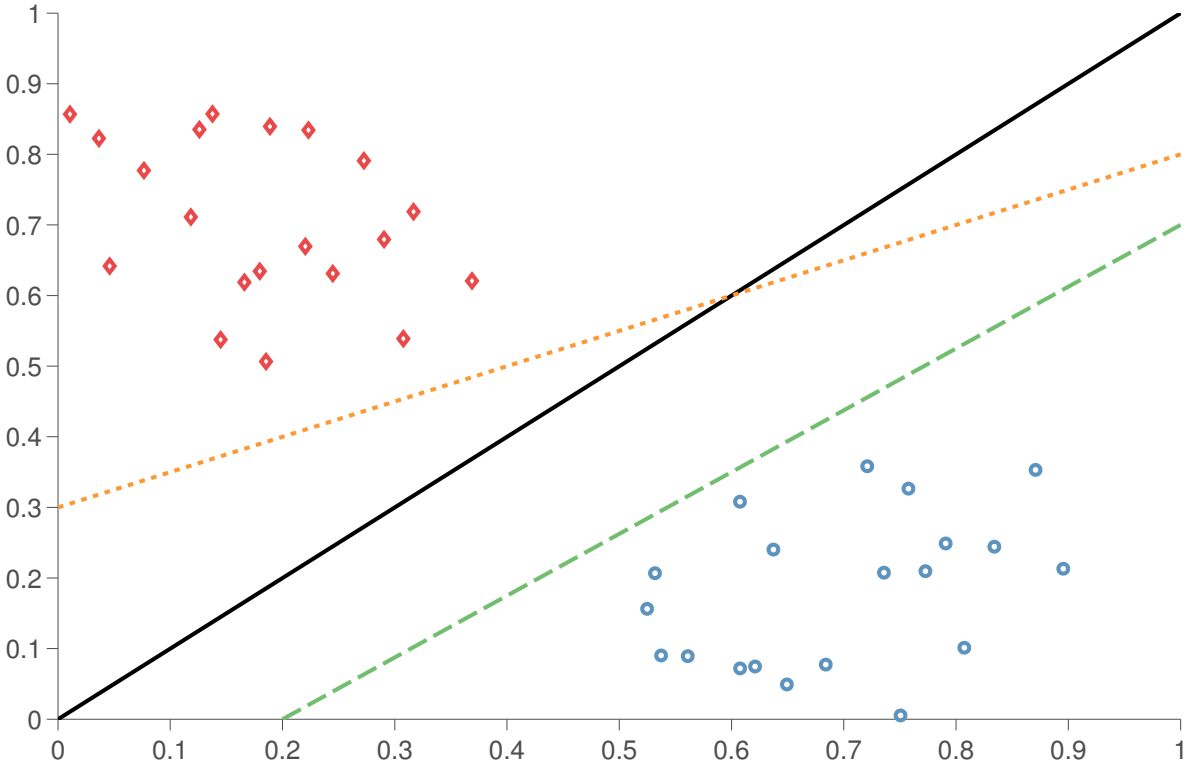

Figure 2.6: *Example of the linearly separable case. Two classes are separated by three lines: the solid black line provides intuitively ideal separation and maximum margin. The green dashed line and the orange dotted line are closer to one of the two respective classes and may not generalise as well as the black solid line.*

For the moment, also assume that all vectors x_i associated with $y_i = +1$ are linearly separable from all vectors associated with $y_i = -1$. We can then construct a "separating hyperplane" which separates these two classes defined by $x \cdot w + b = 0$. While a separating hyperplane is in itself useful for classification, we would also like to formalise the notion of a *good* separating hyperplane, i.e. one that will generalise well. Consider the two lines separating the data in Figure 2.6, where the lines are equivalent to 2 dimensional hyperplanes.

Intuitively from Figure 2.6 we would separate our data using the solid black line as opposed to the dashed green line or dotted orange line. This is formalised by defining the *margin*: the distance between the closest example of each class and the hyperplane. Let us define these distances as d_+ and d_-, and constrain each vector i to satisfy the following conditions:

$$x_i \cdot w + b \geq +1 \, , y_i = +1 \quad \forall \; i, \tag{2.10}$$

and

$$x_i \cdot w + b \leq -1\,, y_i = -1 \;\; \forall\; i. \tag{2.11}$$

These conditions are equivalent to:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \;\; \forall\; i \tag{2.12}$$

Given this hyperplane, one can define a margin which represents the distance from the vectors to the given separating hyperplane. Assume that all points lie at least a distance of ± 1 units from this hyperplane, and that at least one vector satisfies the equation $y_i(x_i \cdot w + b) - 1 = 0$ for $y_i = +1$ and $y_i = -1$. This vector's distance from the origin can be calculated from $x_i \cdot w + b = \pm 1$, giving $\frac{1-b}{\|w\|}$ for $y_i = +1$ and $\frac{-1-b}{\|w\|}$ for $y_i = -1$ (since w and x_i are perpendicular). Since the hyperplane itself is $\frac{|b|}{\|w\|}$ from the origin, the distances can be calculated as $d_+ = d_- = \frac{1}{\|w\|}$ and the size of the margin $(d_+ + d_-)$ is consequently $\frac{2}{\|w\|}$. Support vectors are those vectors which satisfy $y_i(x_i \cdot w + b) - 1 = 0$ for either $y_i = +1$ or $y_i = -1$.

The SVM thus solves the following optimisation problem:

$$\min \|w\| \;\text{ subject to }\; y_i(x_i \cdot w_i + b) \geq 1. \tag{2.13}$$

Recall the assumption that the data be perfectly separable by a hyperplane. This is an impractical assumption, and is relaxed by the addition of slack variables ξ. The earlier constraint in Equation 2.13 becomes:

$$y_i(x_i \cdot w_i + b) \geq 1 - \xi_i, \tag{2.14}$$

and the minimisation becomes

$$\min \|w\| \;\text{ subject } to \; \begin{cases} y_i(x_i \cdot w_i + b) \geq 1 - \xi_i \,\forall i \\ \xi \geq 0, \sum \xi \leq C, \end{cases} \tag{2.15}$$

where we have introduced a hyperparameter C, the capacity. This parameter controls how many misclassifications are allowable in determining the maximum margin hyper-

plane.

The method of Lagrange multipliers (well described by Bishop [50]) is used to transform the constrained optimisation problem into a convex formulation which is easier to solve using quadratic programming techniques.

Extension of the SVM to the non-linear case is achieved through use of the "kernel" trick. The trick relies on the principle that when data are projected into a higher dimensional space they are (ideally) linearly separable. The maximum margin separating hyperplane can then be learnt in this higher dimensional space as previously described. For a feature transformation $\phi(x)$, the existence of a kernel function:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j). \tag{2.16}$$

As the only occurrence of the data in the Lagrange dual formulation of the SVM [50] is as an inner product (i.e. $x_i \cdot x_j$), the use of the kernel function would allow for application of a feature transformation $\phi(x)$ without the computationally expensive inner product of $\phi(x) \cdot \phi(x')$.

If the kernel function satisfies Mercer's condition ($\int \int K(x, y)g(x)g(y)dxdy \geq 0$) [36] then the result of the kernel function is guaranteed to be equivalent to projecting the data into the higher dimensional space and calculating the dot product. For the SVM, common kernels which satisfy this condition are:

$$K(x_i, x_j) = e^{\frac{\|x_i - x_j\|^2}{2\gamma^2}} \tag{2.17}$$

$$K(x_i, x_j) = (x_i) \cdot x_j + 1)^p \tag{2.18}$$

$$K(x_i, x_j) = tanh(\kappa(x_i \cdot x_j) - \delta) \tag{2.19}$$

Equation 2.17 shows the radial basis function kernel, Equation 2.18 shows the polynomial kernel and Equation 2.19 shows the sigmoidal kernel. The kernel chosen for the SVM in this work was an RBF kernel (Equation 2.17). The RBF kernel has two associated hyperparameters, γ and C, optimised during model development. γ is a hy-

perparameter which controls the influence of an individual training point: smaller values extend the influence of the data point, while larger values restrict it. As γ gets larger, the decision boundary becomes more complicated and in the limit becomes many small decision boundaries circumscribing each training observation. C is the same as in Equation 2.13, and controls the number of misclassifications allowed by the separating hyperplane. The software package used to develop SVMs in this work was LIBSVM [119].

2.4 Model configurations

2.4.1 Preprocessing

There are two stages at which removal of artefactual data can be applied: on a time-series of values or on individual features of the design matrix. The first method potentially provides more information as features extracted from a patient time series which would normally be artefactual would be replaced by an true measurement. For example, given a set of heart rate measurements $h = \{0, 65, 70, 68\}$, we can calculate $\min h = 0$. However, the value of 0 here is certainly an outlier, and removal of this value would result in $h = \{65, 70, 68\}$ and $\min h = 65$. However, patient time series are not always available for analysis, and thus a comparison of the preprocessing method applied to only the design matrix is of interest. Furthermore, any given preprocessing method could be applied twice: once to time-series so as to remove values for *variables* which are outliers (e.g., artefactual heart rates), and afterwards to the design matrix so as to remove values for *features* which are outliers (e.g., artefactual highest heart rate). Table 2.3 summarizes the various combinations of model development performed for each model.

2.4.2 Handling of missing values

Most models applied in the comparison require either observations with missing values to be removed (case deletion) or imputation of a numeric value for the missing value. Case deletion is extremely wasteful of data as it is rare to have completely observed all parameters relating to a patient's stay. This is particularly the case with medical data, where missing values can occur because a caregiver perceived it was not necessary or

Preprocessing stage	BCOR	Missing value flagging
(none)	×	×
Timeseries	✓	×
Timeseries	×	✓
Timeseries	✓	✓
Design Matrix	✓	×
Design Matrix	×	✓
Design Matrix	✓	✓
Both	✓	×
Both	×	✓
Both	✓	✓

Table 2.3: *Various permutations of model development for missing value flagging and data preprocessing. Note blank rows indicate no design matrix preprocessing and no missing value flagging.*

important to record a value (which frequently would require ordering a special test). Furthermore, the use of case deletion alters the patient cohort used for model development and subsequently the patient cohort which is suitable for model application. A model which is only applicable to patients for whom all measurements have been taken is extremely impractical. For this reason missing values were imputed prior to model development.

Univariate mean imputation was used to replace missing values. This involves calculating the mean value of all non-missing measurements for a single feature and replacing missing measurements for that feature by the calculated mean. This process is repeated for all features. One obvious objection would be the bias incorporated into the model by this approach. As clinicians often do not order measurements when they believe the value would be normal [120], the imputed values would be biased towards abnormal values. To assess the impact of this, a second method of missing value imputation was used which involved an additional step: concatenation of an additional feature which indicated if the original feature was missing. This is done for each feature which had missing data. The addition of this binary feature allows modelling of the risk associated with the j^{th} feature not being present missing. The downside of this approach is the added dimensionality as in the worst case it would double the number of features in the design matrix. An example of the two methods of missing value imputation is shown in Figure 2.7.

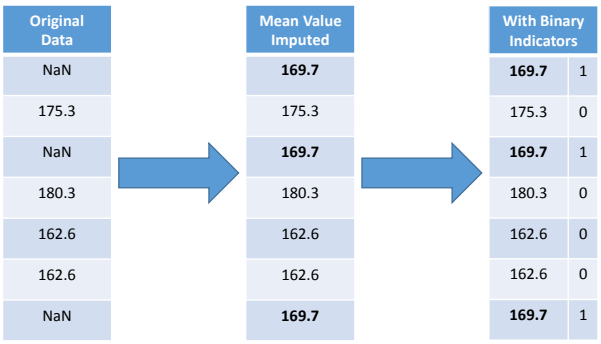

Figure 2.7: *Example of missing value handling as applied to the height feature. The original data appears on the left, from which the mean (169.7 cm) is calculated using all finite observations. This mean is imputed for missing values (middle), and an optional final step is a binary indicator variable marking observations for which the mean was imputed. This process would be repeated for all features when applied to a dataset.*

2.5 Data processing

This section describes the processing used to create a design matrix $\mathbf{X}$ used for model training and a design matrix $\mathbf{Z}$ used for model testing. Both design matrices will have the same columns (features), but the number of rows (patient observations) need not be the same.

2.5.1 Training set preprocessing

The initial data is represented by a set of time values t and a set of observation values v where both t and v contain M values. Consequently there are a pair of tuples t and v of the form:

$$\begin{aligned} v &= (v_1, v_2, ..., v_M) \\ t &= (t_1, t_2, ..., t_M) \end{aligned} \tag{2.20}$$

Optionally, BCOR preprocessing was performed on the time-series prior to feature extraction. This involves deleting the values in v which are deemed outliers. First, the vectors l and h are determined for each variable j using the data, as detailed in Algorithm 1.

Note that j indexes a variable, not a feature. That is, we select the value of l_j and

Algorithm 1 Selection of thresholds used to reject extreme values on training set

for $j = 1 \rightarrow D$ **do**
  $\lambda_1 = \text{argmax}_\lambda \frac{N}{2}\log(\tilde{\sigma}^2) + (\lambda - 1)\sum_{i=1}^{N}\log(\tilde{x}_i)$ {See Equation 2.2}
  $\mathbf{v}'_j = \frac{\mathbf{v}_j^{\lambda_1} - 1}{\lambda_1}$ {If $\lambda_1 = 0$, $\mathbf{v}'_j = \log(\mathbf{v}_j)$}
  $\alpha = 0.01$
  $\Phi \sim \mathcal{N}\left(\bar{\mathbf{v}}'_j, \sqrt{\frac{1}{M}\sum_i^M (v'_{ij} - \bar{\mathbf{v}}'_j)}\right)$
  $l_j = \frac{\frac{\alpha}{2}}{N}\Phi^{-1}\left(\frac{\alpha}{2}\right)$
  $h_j = \frac{1-\frac{\alpha}{2}}{N}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$
end for

h_j by aggregating all measurements of a single variable. For example, when j indexes the heart rate variable, we first collect all heart rate measurements for all patients in the dataset into the vector $\mathbf{v}_j$. This vector of data is used to select parameter λ_1, which is then used to transform $\mathbf{v}_j$ into $\mathbf{v}'_j$ and this vector is used to determine thresholds (l_j and h_j) for rejecting extreme values (outliers). Importantly in this example, *all* heart rate measurements are used to determine these thresholds. Pseudocode for the subsequent extreme value removal process is detailed in Algorithm 2.

Algorithm 2 Application of thresholds used to reject extreme values

for $j = 1 \rightarrow D$ **do**
  for $n = 1 \rightarrow N$ **do**
    for $m = 1 \rightarrow M$ **do**
      if $v_{j,n,m} < l_j$ OR $v_{j,n,m} > h_j$ **then**
        DELETE $v_{j,n,m}$, $t_{j,n,m}$
      end if
    end for
  end for
end for

Here $v_{j,n,m}$ is the m^{th} value for the j^{th} feature in the n^{th} record, l is a vector of D lower thresholds and h is a vector of D upper thresholds. For a single feature j, l_j and h_j were selected using values of v from all records as described in Section 2.2.

Features were then extracted as detailed in Section 2.1.2 and in Figure 2.3 resulting in a design matrix $\mathbf{X}$. The data was then optionally preprocessed a second time using BCOR. Given the design matrix $\mathbf{X}$, this involved application of a function $p(\cdot)$ to a single feature $\mathbf{x}_j$ of the following form:

$$p(\mathbf{x}_j) = \begin{cases} NaN, & \text{if } x_{ij} < l_j \ , \\ NaN, & \text{if } x_{ij} > h_j \ . \\ x_{ij}, & \text{otherwise.} \end{cases} \tag{2.21}$$

Here l_j is a lower threshold and h_j is an upper threshold, both of which have been learnt using the data in feature vector $\mathbf{x}_j$. Thus Equation 2.21 replaces extreme values in $\mathbf{x}_j$ with missing values, and this process is repeated for all features from $j = 1, ..., D$.

Note the difference between the time-series preprocessing and the design matrix preprocessing. In the earlier time-series preprocessing, *all* measurements for a *variable* are utilised to determine thresholds and remove values which are deemed too extreme. In contrast, the design matrix preprocessing uses N values for a single *feature* to determine the thresholds for that feature. Furthermore, time-series preprocessing has the capability of removing an extreme value without affecting subsequent features (e.g. removal of all heart rates of 0 causes the minimum heart rate feature to take on a value of, say, 45), in contrast to design matrix preprocessing which removes the value entirely (e.g. removal of a minimum heart rate value of 0 causes the value to be set to NaN).

The data was then normalised by subtracting the mean and dividing by the standard deviation, where both the mean and standard deviation were calculated after excluding missing values. The calculation for the mean and standard deviation for a single feature $\mathbf{x}_j$ is given in the following equations:

$$\mu(\mathbf{x}_j) = \frac{\sum_{i=1}^{N} x_{ij} \mathbf{1}_{\mathbb{R}}(x_{ij})}{\sum_{i=1}^{N} \mathbf{1}_{\mathbb{R}}(x_{ij})}, \tag{2.22}$$

and

$$\sigma(\mathbf{x}_j) = \sqrt{\frac{\sum_{i=1}^{N} ((x_{ij} - \bar{\mathbf{x}}_j)^2 \mathbf{1}_{\mathbb{R}}(x_{ij}))}{\sum_{i=1}^{N} \mathbf{1}_{\mathbb{R}}(x_{ij})}}, \tag{2.23}$$

where $\mathbf{1}_{\mathbb{R}}(x_{ij})$ is the indicator function, defined as:

$$\mathbf{1}_{\mathbb{R}}(x_{ij}) = \begin{cases} 0, & x_{ij} \notin \mathbb{R}, \\ 1, & x_{ij} \in \mathbb{R}. \end{cases} \tag{2.24}$$

Each vector $\mathbf{x}_j$ was then normalised as follows:

$$\mathbf{x}_j = \frac{\mathbf{x}_j - \mu(\mathbf{x}_j)}{\sigma(\mathbf{x}_j)}. \tag{2.25}$$

Elements in $\mathbf{X}$ with value NaN (i.e. missing values) were set to zero (equivalent to replacing these missing values by the mean of non-missing values) as follows:

$$x_{ij} = \begin{cases} 0, & x_{ij} \notin \mathbb{R}, \\ x_{ij}, & x_{ij} \in \mathbb{R}. \end{cases} \tag{2.26}$$

where $NaN \notin \mathbb{R}$.

Optionally, binary indicator features were concatenated to the design matrix (see Section 2.4.2). Note that binary and categorical variables do not have missing value indicators. This results in a final processed design matrix $\mathbf{X}$.

2.5.2 Test set preprocessing

$\mathbf{Z}$, which represents the test set design matrix, was extracted in the same fashion as $\mathbf{X}$ except no parameters were learned using $\mathbf{Z}$.

If time-series preprocessing is applied, then the initial time-series data was preprocessed using BCOR identically, except vectors l and h were already selected using $\mathbf{X}$. These vectors were used to outlier reject $\mathbf{Z}$ in the same format as before, that is, if time-series preprocessing is utilised the pseudocode in Algorithm 3 is applied where $t_{j,n,m}$ is the time stamp for value $z_{j,n,m}$. Note that the vectors l and h are learned prior during the preprocessing of $\mathbf{X}$.

Algorithm 3 Application of thresholds used to reject extreme values on test set

```
for j = 1 → D do
  for n = 1 → N do
    for m = 1 → M do
      if z_{j,n,m} < l_j OR z_{j,n,m} > h_j then
        DELETE z_{j,n,m}, t_{j,n,m}
      end if
    end for
  end for
end for
```

Features are then extracted from these time-series measurements, and optionally a

second stage of outlier rejection, $p(\cdot)$, is performed as follows:

$$p(\mathbf{z}_j) = \begin{cases} NaN, & \text{if } z_{ij} < l_j \ , \\ NaN, & \text{if } z_{ij} > h_j \ . \\ z_{ij}, & \text{otherwise.} \end{cases} \tag{2.27}$$

The (optionally) processed design matrix $\mathbf{Z}$ is then normalised as follows:

$$\mathbf{z}_j = \frac{\mathbf{z}_j - \mu(\mathbf{z}_j)}{\sigma(\mathbf{z}_j)} \tag{2.28}$$

with $\mu(\cdot)$ defined in Equation 2.22 and $\sigma(\cdot)$ defined in Equation 2.23. Missing values are then imputed as follows:

$$z_{ij} = \begin{cases} 0, & z_{ij} \notin \mathbb{R}, \\ z_{ij}, & z_{ij} \in \mathbb{R}. \end{cases} \tag{2.29}$$

where $NaN \notin \mathbb{R}$. A final, optional step concatenates missing value indicators to the design matrix $\mathbf{Z}$. The features which had missing value indicators created for $\mathbf{X}$ also had missing value indicators created in $\mathbf{Z}$. These steps result in a final design matrix $\mathbf{Z}$ which will be used for final model evaluation.

2.6 Model development

Once both design matrices $\mathbf{X}$ and $\mathbf{Z}$ have been prepared for model development the goal becomes prediction of hospital mortality $\mathbf{y}$, i.e. learn some functional mapping $f(\mathbf{X}; \Lambda, \beta, \theta)$ which minimises a measure of distance $D(\mathbf{y}, f(\mathbf{X}; \Lambda, \beta, \theta))$ or maximises a difference $D(f(\mathbf{X}; \mathbf{y} = 0, \Lambda, \beta, \theta), f(\mathbf{X}; \mathbf{y} = 1, \Lambda, \beta, \theta))$. Here $f(\cdot)$ represents a model which creates a vector of predictions $\hat{\mathbf{y}}$ when presented a design matrix $\mathbf{X}$ (or creates a single prediction $\hat{y}_i$ when given a row vector x_i). The vector θ represents the parameters of the model which are directly optimised during the training process (e.g. coefficients in a regression or support vectors in an SVM). The vector β represents two bias correction coefficients β_0 and β_1 which aim to adjust predictions $\hat{\mathbf{y}}$ on a held out data set to be

a closer estimate to the true outcome $\mathbf{y}$. When the model does not naturally output probabilistic predictions (e.g. SVM), this step has the added benefit of scaling the predictions to the range [0,1]. Finally, the hyperparameters Λ are tunable parameters for the specific model optimised, and these must be selected using an ideally unbiased estimate of model performance (in this work estimated using cross-validation).

To summarise, there are three stages of model development which select Λ, β and θ (in that order). These stages are hyperparameter selection, bias correction and final model development. These stages are now discussed in turn.

2.6.1 Hyperparameter selection

Hyperparameters of each model were determined using 4-fold cross-validation on $\mathbf{X}$. The goal of this process is to obtain held out performance estimates for the model across a set of potential hyperparameters. The best performing combination is then selected as the optimal set of hyperparameters to be used in the final model development. A matrix of potential hyperparameter values, $\mathbf{\Lambda}$, is set prior to model development. The matrix $\mathbf{\Lambda}$ defines the various combinations of hyperparameters which are tested, i.e. it represents grid of searchable values. The following shows an example of a hyperparameter matrix for a model requiring optimisation of L hyperparameters:

$$\mathbf{\Lambda} = \begin{bmatrix} \Lambda_{1,1} & ... & \Lambda_{1,L} \\ . & & . \\ . & & . \\ . & & . \\ \Lambda_{G,1} & ... & \Lambda_{G,L} \end{bmatrix},$$

where L is the number of hyperparameter sets searched in the grid. Here $\Lambda_i = [\Lambda_{i,1}, ..., \Lambda_{i,L}]$ and if Λ_i provided the best model performance (on the held out set) then the hyperparameters selected for the final model would be Λ_i.

Figure 2.8 shows the cross-validation approach used to determine the performance of a single set of hyperparameters. This process is repeated for all rows of the matrix $\mathbf{\Lambda}$, i.e. all combinations of hyperparameters in the grid search.

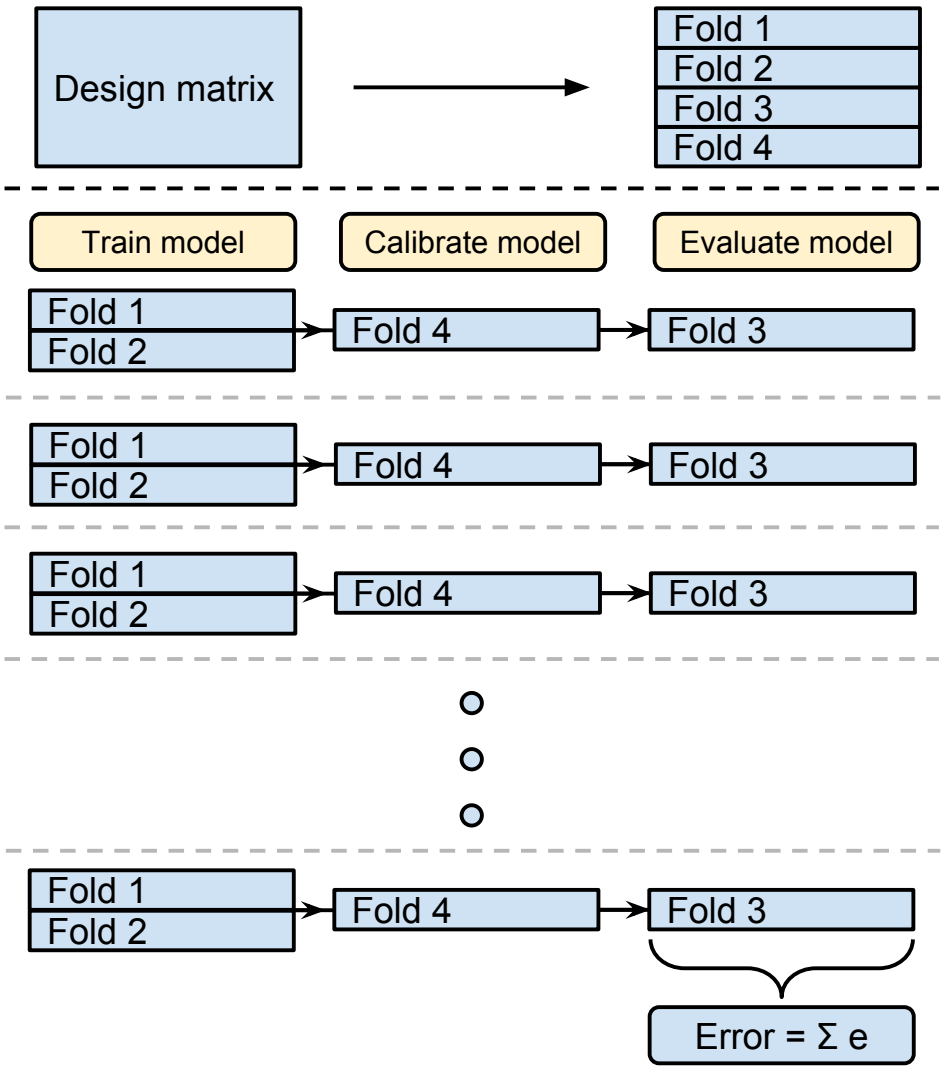

Figure 2.8: *Diagram of the procedure used to evaluate a set of hyperparameters. The data is first split into four folds. For one set of hyperparameters 12 repetitions of model development, calibration and evaluation are then performed using different combinations of folds. Finally, the error for the set of hyperparameter is calculated by summing across the folds (or equivalently averaging).*

Figure 2.8 displays a four fold segmentation and an internal model development with bias correction stage. The bias correction step is described in the following Section 2.6.2. The philosophy behind Figure 2.8 is to acquire the generalisation performance of all possible model combinations which: 1) use two of the four folds of data for model development and 2) a third fold for bias correction. There are 12 possible combinations of splitting folds into 2-1-1 (2 folds for model development, 1 for bias correction and 1 for model evaluation). As a a result, this process generates 12 performance estimates for a set of hyperparameters. These performances are averaged to obtain the overall performance for this set of hyperparameters, and the set with the higher performance are used for final model development.

2.6.2 Bias correction

Bias correction involves a regression of a set of values $\hat{\mathbf{y}}$ onto a set of outcomes $\mathbf{y}$ by maximising the following equation:

$$\beta = \underset{\beta}{\operatorname{argmax}} \ln(\mathcal{L}(\mathbf{y}; g(\hat{\mathbf{y}}))), \tag{2.30}$$

where β is a vector of two elements, $\mathcal{L}(\cdot)$ is the binomial likelihood function (or cross-entropy) and $g(\cdot)$ is defined as follows:

$$g(\hat{\mathbf{y}}) = \begin{cases} \frac{1}{1-e^{-\beta\hat{\mathbf{y}}}}, & \hat{\mathbf{y}} \in [0,1] \\ \hat{\mathbf{y}}, & \text{otherwise.} \end{cases} \tag{2.31}$$

Equation 2.31 scales the values of $\hat{\mathbf{y}}$ onto real values if they are probabilities and retains their values if they are not.

This process is equivalently considered as a logistic regression using a single covariate $g(\hat{\mathbf{y}})$ as the input (or independent variable) and $\mathbf{y}$ as the target (or dependent variable), i.e. we learn the following mapping:

$$\mathbf{y} = \beta_0 + \beta_1 g(\hat{\mathbf{y}}) + \epsilon \tag{2.32}$$

where ϵ is an error term. The result of this process is two coefficients, β_0 and β_1, which

act to recalibrate the predictions $\hat{\mathbf{y}}$ and adjust for any bias in the model development process.

2.6.3 Final model development

After selection of the optimal hyperparameters, Λ, using a grid search (Section 2.6.1) and selection of two bias correction coefficients β_0 and β_1 (Section 2.6.2) the final model is developed. This involves learning parameters θ in the function $\hat{\mathbf{y}} = f(\mathbf{X}; \Lambda, \beta, \theta)$. The exact nature of the parameters θ and their optimisation is model specific and covered in Section 2.6.

2.7 PN_{db} model development and evaluation

Data for PN_{a} was downloaded from PhysioNet[4] [107] as provided by the Physionet Challenge 2012 authors [108]. This resulted in a set of 4,000 records with associated outcomes for model development. Each record contained a set of tuples for clinically relevant variables, see Section 2.1.1 for further detail on the original data format. Furthermore, data from PN_{b} was downloaded resulting in an additional 4,000 records from distinct patients in the same format as PN_{a}. Data was preprocessed as described in Section 2.5 resulting in a design matrix $\mathbf{X}$ for PN_{a} and a design matrix $\mathbf{Z}$ for PN_{b}. Design matrix $\mathbf{X}$ was the training set and design matrix $\mathbf{Z}$ was the test set. All model performance is reported on $\mathbf{Z}$.

A single final model was developed predicting in-hospital mortality using all 4,000 data vectors available in $\mathbf{X}$ as described in Section 2.6. This model was evaluated on the 4,000 data vectors in $\mathbf{Z}$ (sourced from PN_{b}). Evaluation of model performance on PN_{b} was performed by the challenge authors [108] who subsequently provided the performance measures, i.e. all supervised model development is guaranteed to not use the test set $\mathbf{Z}$ as the outcomes are censored. The overall process of developing a model for PN_{db} is shown in Figure 2.9. Models were assessed using the $\mathcal{I}_{\mathcal{L}}$, AUROC, B, B_{adj}, SMR and $HL_{\hat{C}}$ (see Chapter 1 for details).

[4] http://www.physionet.org

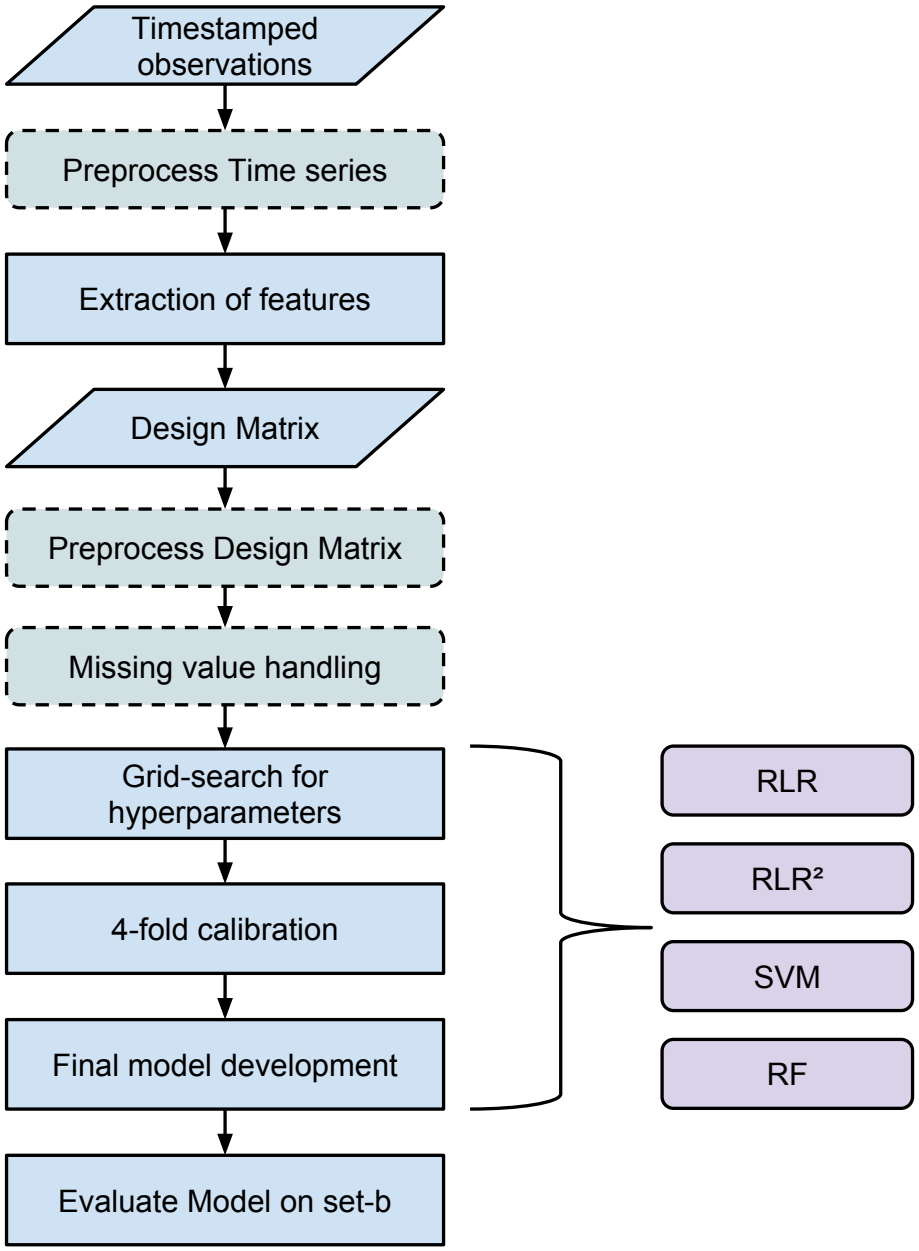

Figure 2.9: *Flowchart of model development process for the four available models. Dashed lines indicate optional processes in the development process which are compared. The four models compared are shown in the right.*

	RLR	RLR^2	SVM	RF
BCOR - timestamp	✓	✓	✓	×
BCOR - design matrix	✓	✓	✓	×
MVF	✓	✓	×	✓
AUROC	0.863	0.863	**0.865**	0.837
$\mathcal{I}_{\mathcal{L}}$	0.257	0.257	**0.294**	0.241
SMR	**1.006**	**1.006**	0.925	1.013
$HL_{\hat{C}}$	95.8	98.3	**14.6**	54.4
B	0.091	0.091	**0.088**	0.093
B_{adj}	0.406	0.406	**0.428**	0.391

Table 2.4: *Best configuration for each model and the associated performance statistics. The best values are emphasised.*

2.8 Results

Figure 2.10 compares the performance of the various model processing configurations as measured by the AUROC, and thus highlights the improvement in model discrimination. Figure 2.11 compares the performance of the various model processing configurations as measured by the $\mathcal{I}_{\mathcal{L}}$, and thus highlights the improvement in both model discrimination and calibration. In these figures, two sets of four statistics are plotted for each model: the lower set of statistics were derived when no preprocessing was applied at the time-series stage, whereas the upper set of statistics had this preprocessing applied. The SVM with both BCOR preprocessing steps applied and missing value flags had the highest AUROC of 0.8653, but the same model without the missing value flags had the best $\mathcal{I}_{\mathcal{L}}$ (0.294). The SVM also had a consistently higher $\mathcal{I}_{\mathcal{L}}$ than all other models, indicating better performance overall on the test set PN_b. Preprocessing at the time-series stage was consistently better than no preprocessing, except for the three models (RLR, RLR^2 and RF) where BCOR was also applied at the design matrix stage.

While the regression models also benefited from BCOR at the time-series stage, the AUROCs were indistinguishable if BCOR was also applied at the design matrix stage. The RF had the worst performance of all models.

The performance of the model configurations with the highest $\mathcal{I}_{\mathcal{L}}$ for the eight combinations of preprocessing and missing value handling is shown in Table 2.4.

Table 2.5 shows the odds ratios[5] for the features retained in the best performing RLR model as measured by the $\mathcal{I}_{\mathcal{L}}$. This model had missing value flags added to the

[5]The odds ratio is calculated as $r = e^{\beta_j}$, where β_j is the coefficient for feature j.

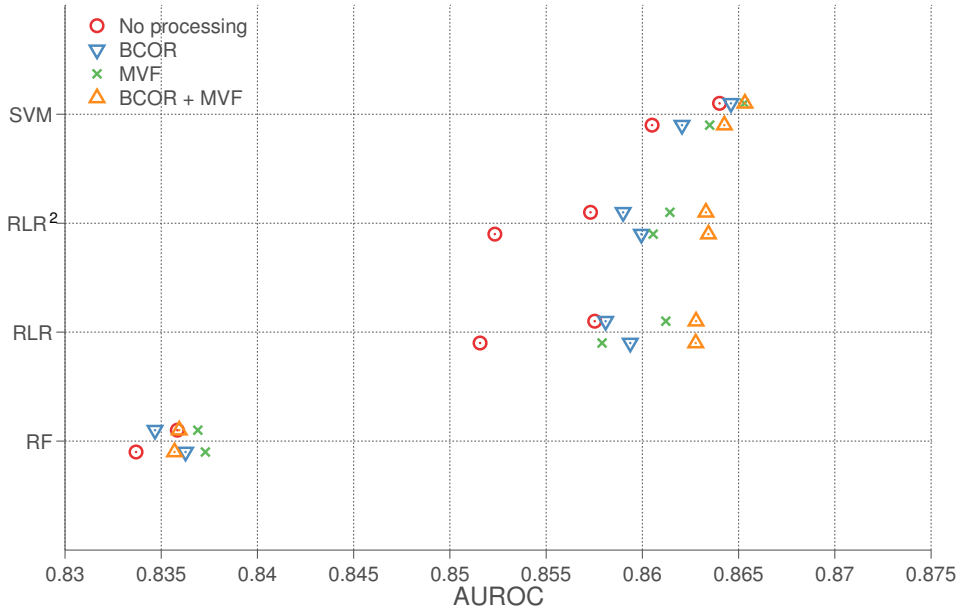

Figure 2.10: *Performance of the models developed on PN_a and evaluated on PN_b as measured the AUROC. For each model, two rows of performances are shown. The top row involves preprocessing the data at the time-series stage using BCOR, while the bottom row does not involve preprocessing the data before creation of the design matrix. The symbols correspond to different levels of preprocessing at the design matrix stage.*

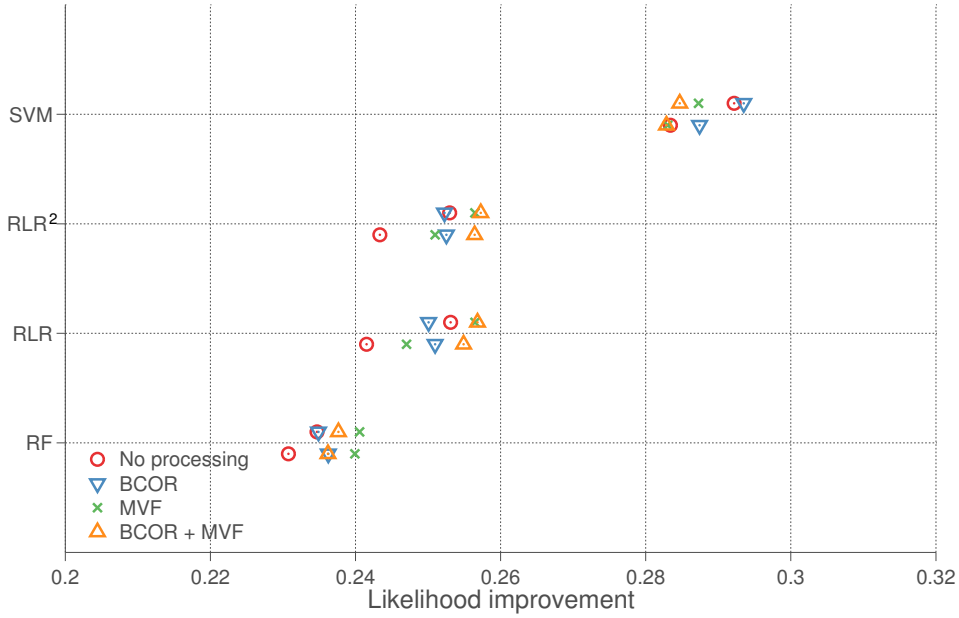

Figure 2.11: *Performance of the models developed on PN_a and evaluated on PN_b as measured the $\mathcal{I}_{\mathcal{L}}$. For each model, two rows of performances are shown. The top row involves preprocessing the data at the time-series stage using BCOR, while the bottom row does not involve preprocessing the data before creation of the design matrix. The symbols correspond to different levels of preprocessing at the design matrix stage.*

design matrix and had BCOR preprocessing applied at both the timestamp and preprocessing stages. Features with positive coefficients, and hence odds ratios greater than one, are correlated with increasing patient severity. Conversely, features with negative coefficients, and consequently odds ratios less than one, are negatively correlated with severity. These coefficients were derived from the best performing classifier for RLR for which all preprocessing methods were applied.

2.9 Discussion

Time-series based preprocessing was much more effective in improving predictive performance as compared to design matrix based preprocessing. This was hypothesised to be the case, as the removal of outliers from the time-series values allows for later feature extraction to retain a true value. This is not possible for design matrix based preprocessing, for which the outlier can only be replaced by a missing value (and in this work is subsequently replaced with the mean value for that feature).

In terms of preprocessing for the RF, one would expect that design matrix preprocessing would have very little effect. This is because the RF has sufficient flexibility to assign unique risks to both missing values (invalidating the use of binary indicator variables) and outliers, reducing the importance of outlier rejection. As discussed before, time-series preprocessing would still have the potential to ensure feature extraction acquires true measurements rather than artefacts, and thus would allow for an improved prediction model even for a flexible model such as the RF. For RF, it seems that design matrix preprocessing has little to no effect, while time-series preprocessing seems to have a moderately beneficial effect, improving $\mathcal{I}_{\mathcal{L}}$ in all but one pairwise comparison. It is worth noting that these differences are small, and there appears to be no obvious advantage to preprocessing the data when using a RF.

Of note is the markedly lower performance exhibited by the RF models across the many of the evaluation statistics. While RFs are extremely flexible and versatile classifiers, a drawback of this flexibility is the need for sufficient data. While the amount of "sufficient" data varies from dataset to dataset, it is empirically clear that for the given problem RFs would benefit from a larger sample size.

Variable	Feature	Odds Ratio	Variable	Feature	Odds Ratio
ALP	Min	1.1171	PaO2	Max	0.9344
ALT	Min △	0.8989	PaO2	First	0.9739
ALT	Max △	1.0000	RespRate	Median △	1.0000
ALT	First △	1.0000	RespRate	Last △	1.0000
AST	Min	1.0143	RespRate	Max △	1.0000
Age		1.4446	RespRate	First △	1.0041
Albumin	Max	0.9181	RespRate	Median	1.0159
Albumin	First	0.9957	RespRate	First	1.0470
BUN	Last	1.2983	RespRate	Min △	1.0957
Bilirubin	Min	1.0607	Temp	Median	0.8409
Bilirubin	Last	1.2060	Troponin	Measured	1.0263
CSRU		0.8961	Troponin-I	Min △	0.8374
DiasABP	Last	0.9943	Troponin-I	Max △	1.0000
FiO2	Min	1.0059	Troponin-I	First △	1.0000
FiO2	Last	1.0278	Troponin-I	Last △	1.0000
FiO2	Median	1.0358	Troponin-I	Median △	1.0000
GCS	Last	0.4959	Troponin-I	Min	1.0331
GCS	Max	0.9256	Urine	Max	0.9704
Glucose	Last	1.1129	Urine	Last	0.9966
HR	Median	1.0206	WBC	Min △	0.8809
HR	Last	1.0410	WBC	Max △	1.0000
HR	Max	1.0820	WBC	Last △	1.0000
Height	△	1.0143	WBC	First △	1.0000
Lactate	Median	1.0052	WBC	Last	1.0568
Lactate	Min	1.0725	Weight	Last	0.9944
Lactate	Last	1.1407	WeightInit	Min	0.9371
MechVent	StartTime	0.9811	WeightInit	First	0.9811
NIDiasABP	Median	0.9601	WeightInit	Last	1.0000
NIMAP	Last	0.9318	WeightInit	Max	1.0000
Na	First	0.9725	WeightInit	Median	1.0000
PaCO2	First	0.9929	pH	Max	1.0009
PaCO2	Max	0.9936			

Table 2.5: *Odds ratios for coefficients in the best performing L^1 regularised logistic regression model which used BCOR at the time-series and design matrix stage as well as missing value flags (indicated by the △). Each feature is listed with its extraction method. For example, "Max" indicates the feature was the maximum of that variable over the first 48 hours, where as "Max △" indicates it was a missing value flag for the maximum of that variable. Note that these missing value flags can vary from feature to feature even if it is the same variable due to the removal of values in the design matrix stage BCOR. Also note that the odds ratios are for normalised, transformed variables and represent the increase in risk for a standard deviation increase in the transformed variable.*

Both the RLR and RLR^2 models perform well, both achieving AUROC= 0.863 for the best model combination and $\mathcal{I}_{\mathcal{L}} = 0.257$. Both models also exhibit similar behaviour under different levels of preprocessing. With no preprocessing, the model performances are substantially lowered, likely due to extreme values shifting predictions either too high or too low. When preprocessing is only applied at the design matrix stage, both models improve equivalently. Overall, the BCOR preprocessing seems to have benefited all models except the RF.

One benefit of regression models, as highlighted earlier, is the ease of their interpretation. The odds ratios in Table 2.5 are congruent with clinical knowledge. It is worth noting that the odds ratios can only be compared relatively as they are derived from transformed, normalised variables. For example, if we consider $\mathbf{x}_j$ to be the age variable, then the odds ratio of 1.44 for age indicates that for every standard deviation increase in $\mathbf{x}_j^{\lambda_1}$ the risk of mortality increases by 1.44. While this makes direct comparison of the odds ratios with other studies difficult, the sign and relative ordering of the odds ratios remains interpretable.

GCS measures a patient's neurological status, with increasing GCS indicating better neurological function. GCS features consistently have odds ratios less than one, indicating that GCS values were correlated with better patient outcomes and further implying that better neurological function is an indicator of better patient outcome. Troponin measurements also follow clinical intuition, though the use of missing value flags slightly obscures this fact. Heart ischaemia can lead to cardiac cell death, a consequence of which is the release of troponin into the blood. Thus we would hypothesize that higher levels of troponin would be correlated with severity, and this is confirmed with the troponin-I (minimum) feature having an odds ratio = 1.0263. The troponin measured feature has an odds ratio of 1.06 indicating that measurement of the biomarker, and thus suspicion of cardiac cell death, is also an indicator of higher severity. Due to the data preprocessing, an additional feature troponin-I Min Δ (odds ratio = 0.837) also captures measurement of the troponin biomarker. As a result, this feature is capturing the same phenomenon: if the troponin-I measurement is missing, the patient has better prognosis.

The existence of almost collinear features, such as troponin-I min Δ and troponin

measured, is due to the automated preprocessing and is a limitation of the study. It is clear from Table 2.5 that there exist many spurious variables. Many missing value flags (e.g. ALT max Δ) have odds ratios = 1.0000 which indicates little clinical significance but, as these features were included in a model with regularisation, their inclusion impacts model performance. Note also that due to the data preprocessing of the design matrix, the missing value flag feature for the minimum respiration rate differs as compared to the missing value flag for the last respiration rate. As the model retained these features, they are considered useful for reducing the model's negative log likelihood, but an argument can be made that they lack generalisability. The principled approach to solving this issue would be removal of these features prior to model development, and this highlights that regularisation, while powerful, is not a sufficient replacement for manual review of the features.

Even though spurious features appear in the best performing RLR model, the final addition of missing value flags still improved model performance on a held out dataset. This indicates that it is advantageous for the models to learn a unique contribution for certain variables with missing values. The inclusion of the missing value features has three possible interpretations. The first implication is that the mean value for these features is not a sufficient approximation for the missing values. If this is the case, then we would hypothesize that both the feature value and the binary indicator flag for the missing values would be present in the model. An example of this situation occurs with respiration rate. As shown in Table 2.5, the first respiration value and a missing value flag are included in the model. In this situation, the missing values for respiration are given a much higher weight than would normally be assigned by mean imputation. The second is that the existence of an observation for the feature is predictively useful. As previously mentioned, this is highlighted by the positive odds ratio for the existence of a troponin measurement. A similar argument can be made for ALT and hepatic failure. The inclusion of whether the height feature is missing may be a surrogate for emergency admissions or a surrogate for a beneficial treatment which requires calculation of body surface area. Finally, the third advantage of missing value flags may be due to the data preprocessing. Recall that if BCOR removed extreme values from a feature of the

design matrix, these values were given a missing value indicator which = 1 if the value is missing and = 0 if it was present. Thus, it is possible that the BCOR removed extreme values from a feature which represented true extreme physiology. In this situation, the missing value indicator becomes a surrogate for extreme physiology in a small subset of patients, and allows for improved model fit in this subset. As the addition of missing value flags improves model performance regardless of BCOR use at the design matrix stage, this is a relevant issue but is likely not the dominant factor in model improvement due to missing value flags.

Overall, the most important variables appear to be the GCS, age and admission of the patient to a cardiac surgery recovery unit (CSRU). GCS is frequently the only predictor which evaluates neurological function in severity scores [7, 35, 54, 55]. The GCS has been shown to have a relatively high univariate AUROC of 0.723 and improve the AUROC of the APS III from 0.818 to 0.858 in non-trauma patients [121]. This is attributable to the poor prognosis of patients with severe neurological dysfunction, regardless of the pathological mechanism, further substantiated by the improvement in AUROC of the APS III in cardiac arrest patients when adding GCS as a covariate (0.804 to 0.869) [121]. Higher age is correlated with risk of mortality and it has been argued that it may act as a surrogate for physiologic reserve [122]. Finally, a large proportion of patients admitted to the CSRU will have received a cardiac artery bypass graft. This procedure is a common low risk surgery which has a mortality rate between 1-3%, an order of magnitude lower than those in the general ICU. Patients undergoing CABG are such a large and distinct cohort of the general ICU population that the APACHE system of equations models these patients with a distinct set of predictive equations [35].

The SVM model benefited from preprocessing primarily at the time-series stage and also to a lesser extent at the design matrix stage. It is interesting to note that the addition of missing value flags always deteriorated the model's performance. As the SVM has no inherent feature selection (though the capacity hyperparameter acts as an L^2 norm regularisation term), this indicates that the random errors introduced by the addition of the binary indicator features outweighed their prognostic value. An extension of the SVM which incorporates the L^1 norm to encourage sparsity in the feature space [123] has

been proposed, and the use of an L^1 regularised SVM would provide a fairer comparison to the RLR models which utilised L^1 regularisation.

The RLR model performs extremely well given the simple nature of logistic regression. While the superiority of the SVM indicates that there is some performance to be gained by utilising non-linear sophisticated machine learning methods, it may not be sufficiently large to justify the additional complexity. While logistic regression is widely used in clinical practice, it is important to note that the RLR model presented here has an additional critical component: regularisation. Regularisation allowed the reduction of the feature space from 228 features to 63, even though it could be reasonably assumed that all the features would be correlated with severity of illness. Regularisation is very rarely employed for regression models in the mortality prediction literature. The traditional approaches to reducing the feature space include clinical judgement [124] [7], step-wise backward selection [34] or step-wise forward selection [35]. Step-wise forward selection involves starting with the most predictive feature and adding a single feature if it is statistically significantly predictive of the outcome. The approach here is similar to forward selection, except instead of fully adding a single feature, it begins with the most predictive feature and sequentially adds a portion of a single variable which maximally improves the likelihood of the model. The results in this work have shown that proper use of L^1 regularisation in combination with a linear model on a reasonably sized dataset is sufficient for an excellent mortality prediction model. While non-linear methods do provide some benefit, it may not be worth the added complexity.

It is also clear from the results that there is an insufficient number of observations for the RF to learn a strong mapping between the features and mortality. It is likely that the use of more training data would improve the performance of the RF. The MIMIC II database [125], from which both $\mathrm{PN_a}$ and $\mathrm{PN_b}$ are derived, is continually updated with more patient data after a quality control and anonymisation process. The most recent release of MIMIC II, version 2.6, contains patient data for admissions up to and including 2008. An updated version containing patient data up to and including 2012 is scheduled for release (though the exact date is unknown). The addition of this data could be sufficient to allow for the development of better RF models, though it would need to

be empirically determined if the gain is sufficient for departure from linear models.

Overall, the RLR and RLR^2 models performed extremely well given their simplicity. The SVM with BCOR preprocessing applied both at the timestamp and the design matrix stage provided superior performance. The SVM with BCOR preprocessing at the time-series and design matrix stage was the best performing model as measured by the $\mathcal{I}_{\mathcal{L}}$. However, if simplicity is preferred, then the RLR model with preprocessing applied at both stages and missing value flags performed competitively.

Chapter 3

Preprocessing applied to a large multi-center database

quod gratis asseritur, gratis negatur.

What is asserted gratuitously may be denied gratuitously.

Anonymous

The goals of this chapter are: i) comparison the performance of the SVM, RLR, RLR^2 and RF models on a much larger multi-center database, ii) evaluation of the additional application of BCOR preprocessing to the design matrix, and iii) evaluation the addition of missing value flags. This results in four possible combinations of model development for each model evaluated, and a total of sixteen models are compared.

3.1 APACHE Outcomes (AO)

In the process of providing APACHE IV predictions for various hospitals, Cerner corporation (Kansas City, MO, United States) have collected a large multi-center database, hereafter referred to as the AO database [126]. An overview of the data storage and extraction process is provided in Figure 3.1. The data was collected from multiple institutions either via manual data entry on a patient by patient basis, an automatic "data crawler" which extracted data from a local electronic medical record system or

a batch upload of data from the electronic medical record system. The data was first uploaded to an intermediate database, and a subsequent extraction from this database was performed to acquire a subset of patients.

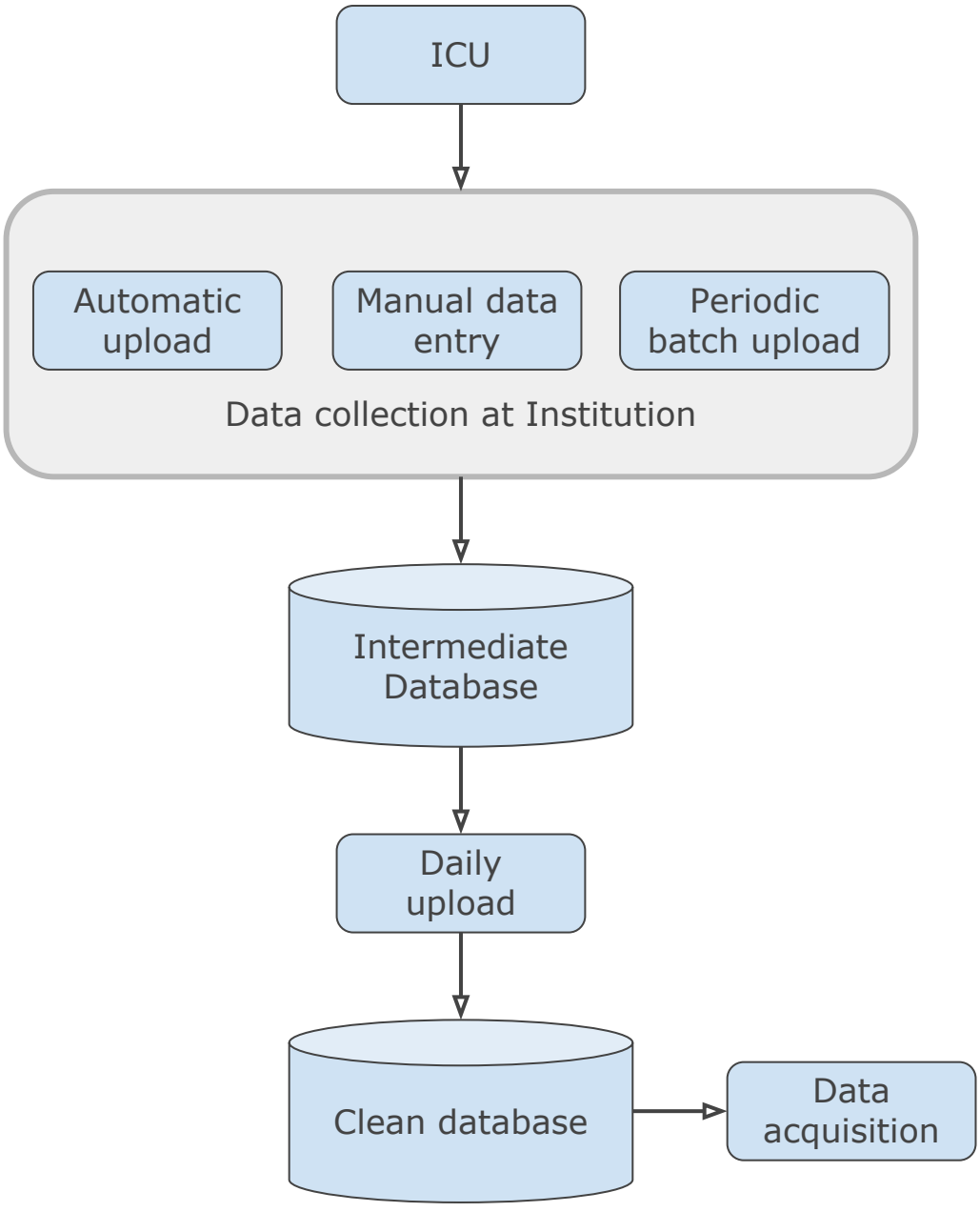

Figure 3.1: *Overview of the storage and extraction process for institutions with the APACHE system.*

Data quality assurance was performed before data acquisition. The data were checked for both physiologic and internal consistency at three stages in the data warehousing

and extraction process. First, the interface used by institutions for data entry contains physiologic limits on certain variables. These limits and the associated variables are shown in Table 3.1. Data which did not rest within the "critical" limits, defined in the software to represent extrema imposed by physiology, were discarded. These limits are shown in Table 3.1 for each variable. For "vital" measurements, denoted in Table 3.1 by an asterisk, data were reacquired from the host institution if missing. Consequently, the amount of missing data for these vital measurements is extremely low (but non-zero as for a very small number of patients the data could not be acquired). These variables were originally selected on the basis that all patients admitted to an ICU should have at least one measurement per day for these variables. Additional data preprocessing performed before acquisition included imputation of 385.7143 for missing values of the ratio of PaO_2:FiO_2 (fraction of arterial blood gas oxygen to inspired airflow) and a GCS of 15 if the patient was sedated. Note that a binary flag ("unable") was also included to indicate whether the GCS imputation had been performed.

Data from each hospital were stripped of patient identifiers in compliance with Health Insurance Portability and Accountability Act requirements. A subset of the AO dataset was used to develop the APACHE IV set of equations [35], and informed consent for the collection of this data was not required due to Institutional Review Board waivers [7].

A flat design matrix which contains a set number of features for each observation was then electronically transferred from Cerner corporation to Oxford. Each observation in the design matrix corresponded to a patient's first day in the ICU. Patient data was excluded if they were admitted with burns, < 16 years old, a post-transplant patient (except kidney and liver transplantation, which were retained in the data) or if the patient was an in-hospital readmission. These data were originally collected from 89 ICUs at 49 distinct hospitals currently utilising the commercial healthcare products provided by Cerner. There were 81,001 first day ICU admissions all occurring between 2007 and 2011, inclusively. Physiologically based features recorded for each observation correspond to the worst measurement across the first 24 hours in ICU, where worst is defined as the value in the first 24 hours which gave the highest component for the APS III. Other features include patient demographics, chronic health information and

Variable	Unit	Ranges				
		Normal Low	Normal High	Normal Value	Critical Low	Critical High
Temperature*	Celsius	32	40.5	38	25	44
S-BP*	mm Hg	40	300		1	400
D-BP*	mm Hg	20	180		1	300
MAP*	mm Hg	40	140	90	1	334
Respiratory Rate*	breaths/min	4	60	19	1	300
Ventilated for this RR?*		0	1	0	0	1
Heart Rate*	beats/min	20	250	75	1	400
Total Urine Output	mL/day	0	30000		0	
24Hr Urine Output	mL/day	0	30000		0	
GCS-Meds?*	0	0	1	0	0	1
GCS-Eyes*	0	1	4	4	1	4
GCS-Motor*	0	1	6	6	1	6
GCS-Verbal*	0	1	5	5	1	5
ABG-Intubated?	0	0	1	0	0	1
ABG-FiO2	%	21	100	21	21	100
ABG-PaO2	mm Hg	30	450	80	0	
ABG-PaCO2	mm Hg	10	150	40	0	
ABG-pH	0	6.8	7.7	7.4	0	
WBC	10^3/uL	0.1	100	11.5	0	
Hematocrit	%	11	54	45.5	0	
Serum Na+	mEq/L	90	160	145	0	
Serum BUN	mg/dL	1	160	0		
Serum Creatinine	mg/dL	0.1	20	1	0	
Serum Glucose	mg/dL	20	1500	130	0	
Serum Albumin	g/dL	1.5	4.5	3.5	0	
Serum Bilirubin	mg/dL	0.2	25		0	
Potassium	mEq/L	1.5	8	4.5	0	
Serum Bicarbonate	mEq/L	4	40	27	0	
Platelets	10^3/uL	5	1000		0	
INR	0	0.8	5		0	
Hemoglobin	g/dL	2	18		0	

Table 3.1: *Measurements which had physiologically based thresholds applied at the source of data acquisition prior to synthesis into a research database. Values outside the critical range were reviewed and this review process resulted in exclusion (resulting in a missing value) or reacquisition of the correct value by manual review of the patient's medical record.*

Type of measurement	Feature(s)
Demographic	Age, Gender
Race	African American, Caucasian, Latino
Comorbidity	AIDS, cirrhosis, multiple myeloma, immunosuppression, lymphoma, hepatic failure, tumour with metastasis, existence of any comorbidity
Admission type	Operative, elective, emergency
Other	Pre-ICU length of stay

Table 3.2: *Static features available in the AO data.*

admission diagnosis.

3.2 Data

The data were first separated into two sets: one for training and the other for model evaluation (i.e. a test set). Patient admissions between 2007 and 2009 (inclusive) were used for the training set, $\mathbf{X}$, and patient admissions between 2010 and 2011 (inclusive) were used for the test data set, $\mathbf{Z}$. This temporal splitting was chosen as it tests any predictive model developed on future data; the same data which the model would be applied to if it were used in practice. The resulting training dataset had 53,758 admissions (rows in $\mathbf{X}$) and the test dataset had 27,243 admissions (rows in $\mathbf{Z}$). Of these features, 17 were static features based upon administrative information (e.g. elective admission), comorbid status (e.g. cirrhosis) or demographics (e.g. age). These features are listed in Table 3.2. A further 24 features related to daily measurements, and the worst measurement across the day was stored for continuous measurements. Worst in this case is defined as the measurement which gives the highest value of the APS III. These features are listed in Table 3.3. Diagnosis was coded as 126 binary indicator variables and is listed in Table C.1 in the appendix (Section C.1). ICU type, body system, hospital type (e.g. community operated teaching hospital) and location prior to ICU admission were also coded using binary indicator variables resulting in an additional 30 features. These features are listed in Table 3.4. In total there were 197 features in both $\mathbf{X}$ and $\mathbf{Z}$.

Type of measurement	Feature(s)
Blood gas	PaO_2, $PaCO_2$
Oxygen delivery	FiO_2, PaO_2/FiO_2
Treatment	Thrombolytic therapy, mechanical ventilation, intubation, existence of any of these treatments
Chemistry	Albumin, bilirubin, blood urea nitrogen (BUN), creatinine, glucose, sodium
Haematology	Haematocrit, white blood cell count
Neurological status	Glasgow coma scale (GCS), unable to acquire GCS
Physiology	Heart rate, mean arterial pressure (MAP), pH, respiratory rate, temperature
Other	Worst respiratory rate occurred during mechanical ventilation

Table 3.3: *Features available in the AO data extracted from the first twenty four hours of a patient's stay in the ICU.*

Type of measurement	Acronym	Features
ICU type	ICU	Cardiothoracic, coronary, medical, mixed, surgical, trauma, unknown
Body system	System	CABG, Genitourinary, haematological, metabolic/endocrine, musculoskeletal/skin, respiratory, transplant, trauma, unknown
Hospital type	Hosp	Community operated teaching hospital, non-teaching hospital, small teaching hospital, unknown
Location prior to ICU admission	Source	Direct admission, floor, ICU transfer, operating room, other hospital, other, recovery room, step down unit, unspecified, unknown

Table 3.4: *Nominal features available in the AO data which were coded as binary indicator variables. The acronym listed here is used to identify features of this type in the presentation of the results.*

3.3 Model development

An overview of the model development methodology is shown in Figure 3.2. The model development follows that detailed in Chapter 2 Sections 2.5 and 2.6, except there is no possibility to preprocess the data at the time-series stage as this data was not available. When discussing results, the term "model configuration" will refer to the possible application of BCOR preprocessing and possible concatenation of missing value flags. For example, a RLR model with BCOR preprocessing and missing value flags added is one configuration of the RLR model, and a RLR model with no BCOR preprocessing and no missing value flags is another configuration.

The development of each of these models involved: i) a grid search for hyperparameters, ii) calculation of bias correction coefficients, iii) final model development on the training set $\mathbf{X}$, and iv) evaluation of the final model on the held out test set $\mathbf{Z}$. For details of the grid search, calibration and final model development process see Section 2.6.

Statistical significance was evaluated using the test of proportions for binary variables and Student's t-test (unpaired samples) for continuous variables. Three levels of significance were considered: $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***). Models were evaluated using the AUROC, $\mathcal{I}_{\mathcal{L}}$, SMR, $HL_{\hat{C}}$, B and B_{adj}. Statistical significance for paired AUROCs was calculated using the non-parametric method of DeLong and DeLong [127]. Statistical significance for other statistics was performed using bootstrap resampling with bias and acceleration correction [49]. Note that adjustment for multiple hypothesis testing was not performed. Also note that many binary comparisons which use the test of proportions are done on subsets of a feature type; i.e. we perform a comparison of a nominal variable (e.g. ICU type) by comparing each individual category (e.g. mixed ICU). An alternative test could have used the nominal counts, which could be done using a χ^2 test. However, this was not done in this work.

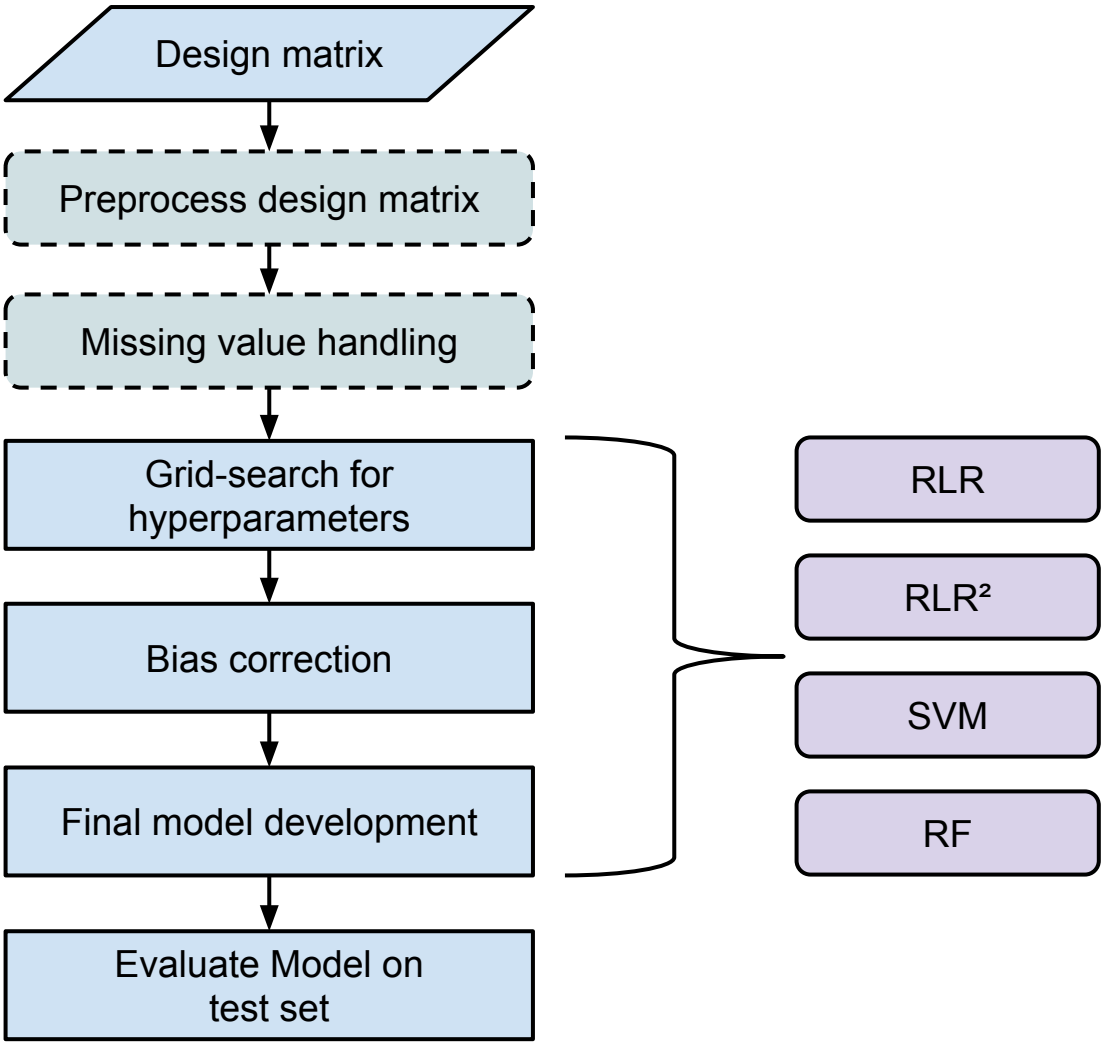

Figure 3.2: *Overview of the evaluation of four binary classification models on the AO dataset.*

3.4 Results

The demographics of the data set are shown in Table 3.5, which compares the training and test sets. There were 53,758 observations in the training set and 27,243 observations in the test set. The demographics after grouping patients further into those who survived past hospital discharge and those who did not are shown in Table 3.6. There were significant differences between the demographics of the training and test set. Patients in the test set had shorter stays, lower crude rates of mortality and were ventilated less frequently. In both the training and the test set, non-survivors were much more likely to be ventilated, older and have longer ICU/hospital stays. The frequency of ventilation was lower for both surviving and non-surviving subpopulations in the test set as compared to the training set.

	Training	Test	
Age	61.22 (17.78)	61.83 (17.63)	***
Ventilation status	19254 (35.82)	8941 (32.82)	***
ICU LOS	3.96 (6.52)	3.39 (5.19)	***
ICU mortality	3938 (7.33)	2043 (7.50)	
Hospital LOS	10.10 (19.06)	8.88 (10.10)	***
Hospital mortality	6264 (11.65)	3006 (11.03)	**
Gender: Female	24448 (45.48)	12573 (46.15)	*

Table 3.5: *Demographics in the training (2007-2009) and test (2010-2011) datasets. LOS - Length of stay.*

	Training			Test		
	Survivors	Non-survivors		Survivors	Non-survivors	
Age	61.22 (17.78)	67.37 (15.92)		61.83 (17.63)	68.95 (15.15)	
Ventilation status	15205 (32.01)	4049 (64.64)	***	7158 (29.53)	1783 (59.31)	***
ICU LOS	3.96 (6.52)	6.03 (10.34)		3.39 (5.19)	4.94 (7.11)	
ICU mortality	0 (0.00)	3938 (62.87)	***	0 (0.00)	2043 (67.96)	***
Hospital LOS	10.10 (19.06)	11.25 (40.19)		8.88 (10.10)	8.83 (11.77)	
Hospital mortality	0 (0.00)	6264 (100.00)	***	0 (0.00)	3006 (100.00)	***
Gender: Female	21585 (45.45)	2863 (45.71)	***	11164 (46.06)	1409 (46.87)	***

Table 3.6: *Demographics in the training (2007-2009) and test (2010-2011) datasets grouped by in hospital mortality. *** Significant at the 0.001 level. LOS - Length of stay.*

Details regarding ICU type, admission type, admission source, comorbidities and primary body system affected are provided in Table 3.7. A statistically significantly larger proportion of the test set were admitted to community operated teaching hospitals (62.21% versus 34.99%, $p < 0.001$). No patients were admitted to a trauma ICU or a neurologic ICU in the test set though these ICU types made up a small proportion of the training set (2.48% trauma and 4.43% neurologic). The largest ICU type represented was the mixed ICU, and a statistically significantly larger amount of patients were admitted to mixed ICUs in the test set (50.28% versus 36.36% in the training set, $p < 0.001$). Most patients were admitted from the emergency room in both the training (39.06%) and the test (43.11%) datasets. All comparisons in Table 3.7 were statistically significant at the 0.001 level except for the comparison of missing ICU types ($p < 0.01$).

Table 3.8 shows ICU type, admission type, admission source, comorbidities and primary body system affected after grouping patients into survivors and non-survivors. Survivors were more frequently admitted from the operating room, recovery room, emergency room and as a direct admission. Non-survivors were more frequently admitted

	Training	Test	
Prior Location			
Operating Room	6103 (11.35)	3787 (13.90)	***
Recovery Room	9425 (17.53)	3083 (11.32)	***
Emergency Room	20999 (39.06)	11744 (43.11)	***
Floor	6347 (11.81)	2576 (9.46)	***
ICU Transfer	636 (1.18)	217 (0.80)	***
Other Hospital	5253 (9.77)	2368 (8.69)	***
Direct Admission	699 (1.30)	0 (0.00)	***
SDU	1899 (3.53)	1477 (5.42)	***
Other	1550 (2.88)	1213 (4.45)	***
Missing	847 (1.58)	778 (2.86)	***
Hospital type			
COTH[1]	18811 (34.99)	16948 (62.21)	***
Small teaching hospital	20983 (39.03)	5316 (19.51)	***
Non-teaching hospital	12805 (23.82)	4979 (18.28)	***
Missing	1159 (2.16)	0 (0.00)	***
ICU Type			
Missing	1577 (2.93)	881 (3.23)	**
Coronary	3654 (6.80)	3696 (13.57)	***
Cardio-Thoracic	4978 (9.26)	1535 (5.63)	***
Medical	7021 (13.06)	3483 (12.78)	
Neurologic	2384 (4.43)	0 (0.00)	***
Surgical	13264 (24.67)	3949 (14.50)	***
Trauma	1333 (2.48)	0 (0.00)	***
Mixed	19547 (36.36)	13699 (50.28)	***
Admission type			
Elective surgery	12537 (23.32)	5180 (19.01)	***
Emergency surgery	3074 (5.72)	1765 (6.48)	***
Medical	38147 (70.96)	20298 (74.51)	***

Table 3.7: *Administrative information for patients in the training (2007-2009) and test (2010-2011) datasets.*
*** Significant at the 0.01 level.*
**** Significant at the 0.001 level.*
[1] *Community operated teaching hospital.*

	Training		Test	
	Survivors	Non-survivors	Survivors	Non-survivors
Prior Location				
Emergency Room	18622 (39.21)	2377 (37.95)	10547 (43.52)	1197 (39.82)
Recovery Room	8963 (18.87)	462 (7.38)	2949 (12.17)	134 (4.46)
Operating Room	5724 (12.05)	379 (6.05)	3541 (14.61)	246 (8.18)
Floor	5103 (10.74)	1244 (19.86)	2178 (8.99)	398 (13.24)
Other Hospital	4499 (9.47)	754 (12.04)	1997 (8.24)	371 (12.34)
SDU	1498 (3.15)	401 (6.40)	1172 (4.84)	305 (10.15)
Other	1235 (2.60)	315 (5.03)	989 (4.08)	224 (7.45)
Missing	740 (1.56)	107 (1.71)	692 (2.86)	86 (2.86)
Direct Admission	623 (1.31)	76 (1.21)	0 (0.00)	0 (0.00)
ICU Transfer	487 (1.03)	149 (2.38)	172 (0.71)	45 (1.50)
Hospital type				
Small teaching hospital	18896 (39.79)	2087 (33.32)	4826 (19.91)	490 (16.30)
COTH[1]	16206 (34.12)	2605 (41.59)	14893 (61.45)	2055 (68.36)
Non-teaching hospital	11351 (23.90)	1454 (23.21)	4518 (18.64)	461 (15.34)
Missing	1041 (2.19)	118 (1.88)	0 (0.00)	0 (0.00)
ICU Type				
Mixed	17138 (36.08)	2409 (38.46)	11983 (49.44)	1716 (57.09)
Surgical	12046 (25.36)	1218 (19.44)	3684 (15.20)	265 (8.82)
Medical	5734 (12.07)	1287 (20.55)	3013 (12.43)	470 (15.64)
Cardio-Thoracic	4629 (9.75)	349 (5.57)	1427 (5.89)	108 (3.59)
Coronary	3223 (6.79)	431 (6.88)	3338 (13.77)	358 (11.91)
Neurologic	2114 (4.45)	270 (4.31)	0 (0.00)	0 (0.00)
Missing	1420 (2.99)	157 (2.51)	792 (3.27)	89 (2.96)
Trauma	1190 (2.51)	143 (2.28)	0 (0.00)	0 (0.00)
Admission type				
Medical	32739 (68.93)	5408 (86.33)	17675 (72.93)	2623 (87.26)
Elective surgery	12065 (25.40)	472 (7.54)	5015 (20.69)	165 (5.49)
Emergency surgery	2690 (5.66)	384 (6.13)	1547 (6.38)	218 (7.25)

Table 3.8: *Administrative information for patients in the training (2007-2009) and test (2010-2011) datasets after grouping patients into survivors and non-survivors (at hospital discharge). All comparisons were statistically significant at the 0.001 level.*
[1] *Community Operated Teaching Hospital.*

from the floor, another ICU, another hospital or the SDU. Elective surgery patients made up a large proportion of survivors relative to the same admission type in non-survivors.

Table 3.9 compares the comorbidities and primary body system afflicted for patients in the training and test sets. There were fewer patients for each comorbidity in the test set, except for lymphoma. The differences were only statistically significant for AIDS, immunosuppression and hepatic failure. There was a statistically significantly higher proportion of patients whose primary body system afflicted was either cardiovascular, metabolic/endocrine and CABG. Statistically significantly lower proportions occurred for transplant, trauma and respiratory patients.

	Training	Test	
Comorbidities			
Immunosuppressed	3749 (6.97)	1457 (5.35)	***
Tumour with metastasis	2913 (5.42)	1422 (5.22)	
Cirrhosis	1647 (3.06)	787 (2.89)	
Hepatic failure	779 (1.45)	245 (0.90)	***
Multiple myeloma	642 (1.19)	303 (1.11)	
Lymphoma	384 (0.71)	201 (0.74)	
AIDS	244 (0.45)	55 (0.20)	***
Body System			
Cardiovascular	17425 (32.41)	9631 (35.35)	***
Neurologic	9671 (17.99)	4499 (16.51)	***
Respiratory	9210 (17.13)	4103 (15.06)	***
Gastrointestinal	7048 (13.11)	3628 (13.32)	
Trauma	4514 (8.40)	2080 (7.63)	***
Metabolic/Endocrine	1743 (3.24)	1078 (3.96)	***
Genitourinary	1369 (2.55)	643 (2.36)	
Musculoskeletal/Skin	1318 (2.45)	675 (2.48)	
CABG[1]	948 (1.76)	772 (2.83)	***
Transplant	380 (0.71)	59 (0.22)	***
Haematological	113 (0.21)	49 (0.18)	
Missing	19 (0.04)	26 (0.10)	***

Table 3.9: *Comparison of the diagnostic and comorbid groups in the AO dataset.*
**** Significant at the 0.001 level.*
[1] *Coronary artery bypass graft.*

	Training		Test	
	Survivors	Non-survivors	Survivors	Non-survivors
Comorbidities				
Immunosuppressed	2975 (6.26)	774 (12.36)	1171 (4.83)	286 (9.51)
Tumour with metastasis	2315 (4.87)	598 (9.55)	1091 (4.50)	331 (11.01)
Cirrhosis	1265 (2.66)	382 (6.10)	614 (2.53)	173 (5.76)
Hepatic failure	557 (1.17)	222 (3.54)	180 (0.74)	65 (2.16)
Multiple myeloma	447 (0.94)	195 (3.11)	227 (0.94)	76 (2.53)
Lymphoma	279 (0.59)	105 (1.68)	164 (0.68)	37 (1.23)
AIDS	190 (0.40)	54 (0.86)	46 (0.19)	9 (0.30)
Body System				
Cardiovascular	14927 (31.43)	2498 (39.88)	8182 (33.76)	1449 (48.20)
Neurologic	8789 (18.51)	882 (14.08)	4121 (17.00)	378 (12.57)
Respiratory	7841 (16.51)	1369 (21.86)	3575 (14.75)	528 (17.56)
Gastrointestinal	6247 (13.15)	801 (12.79)	3288 (13.57)	340 (11.31)
Trauma	4137 (8.71)	377 (6.02)	1915 (7.90)	165 (5.49)
Metabolic, Endocrine	1654 (3.48)	89 (1.42)	1047 (4.32)	31 (1.03)
Genitourinary	1264 (2.66)	105 (1.68)	591 (2.44)	52 (1.73)
Musculoskeletal, Skin	1219 (2.57)	99 (1.58)	639 (2.64)	36 (1.20)
CABG	935 (1.97)	13 (0.21)	759 (3.13)	13 (0.43)
Transplant	369 (0.78)	11 (0.18)	57 (0.24)	2 (0.07)
Haematological	95 (0.20)	18 (0.29)	41 (0.17)	8 (0.27)
Missing	17 (0.04)	2 (0.03)	22 (0.09)	4 (0.13)

Table 3.10: *Demographics of the dataset used for model training and test. All comparisons were significant at the 0.001 level.*
[1] *Community Operated Teaching Hospital.*

Table 3.10 shows the comorbidities and primary body system associated with the admissions of patients in the training and test datasets grouped by survival to hospital discharge. A higher proportion of non-survivors had a comorbidity for all comorbidities assessed. Patients in the test set had fewer comorbidities across surviving and non-surviving groups. The body systems associated with the primary diagnosis which were more common among non-survivors were the haematological, respiratory and cardiovascular systems. This pattern for both training and test datasets.

The performance of all models as measured by the AUROC is shown in Figure 3.3. The SVM model with BCOR preprocessing and missing value flags has the highest AUROC of 0.8917. The second best discriminating model was RLR[2] with BCOR preprocessing and missing value flags (AUROC = 0.8916), and the third best performing model was the RLR with BCOR preprocessing and missing value flags (AUROC = 0.8912).

The performance of the models as measured by the $\mathcal{I}_{\mathcal{L}}$ is shown in Figure 3.4. The

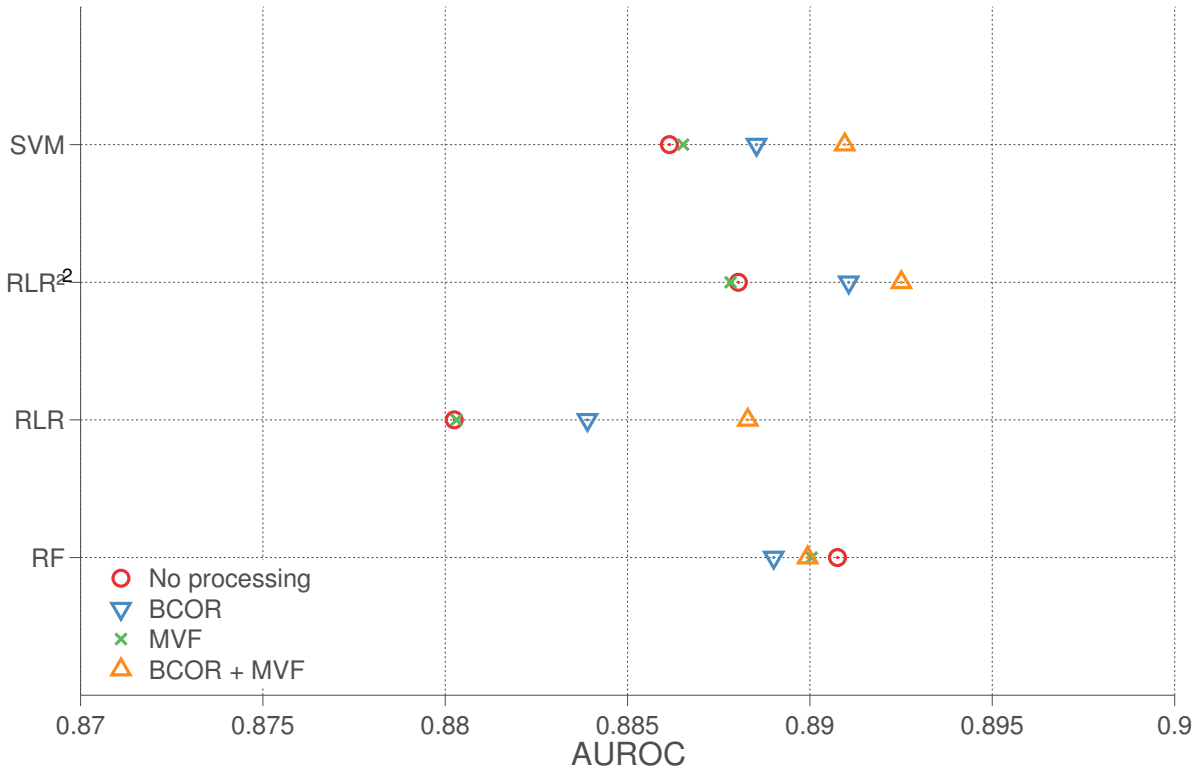

Figure 3.3: *Performance of the models developed on AO and evaluated on the test set as measured the AUROC. The symbols correspond to different levels of preprocessing at the design matrix stage.*

best performing model in terms of the $\mathcal{I}_{\mathcal{L}}$ was the RLR^2 model with BCOR preprocessing and missing value flags ($\mathcal{I}_{\mathcal{L}} = 0.361$). The next best model was a RF with no preprocessing or missing value flags ($\mathcal{I}_{\mathcal{L}} = 0.360$). The worst model was the RLR with no preprocessing or missing value flags ($\mathcal{I}_{\mathcal{L}} = 0.321$).

Statistics for all models evaluated are shown in Table 3.11. The best performing configurations for each model assessed are emphasised. While the best SVM and RF had excellent discrimination (AUROC = 0.891), the RLR^2 model had the highest (AUROC = 0.893). The RLR^2 also had an extremely competitive $\mathcal{I}_{\mathcal{L}} = 0.361$ which was slightly superior to the best $\mathcal{I}_{\mathcal{L}} = 0.360$ for the RF.

Calibration curves for the best configurations of each model evaluated are shown in Figure 3.5. The RLR overpredicted mortality in the lower ranges of risk but underpredicted mortality in the higher ranges of risk. The RLR^2 model overpredicted mortality consistently across all levels of risk. The RF overpredicted mortality in the lower ranges of risk, and underpredicted mortality in the higher ranges of risk. Finally, the SVM overpredicted mortality across all ranges of risk.

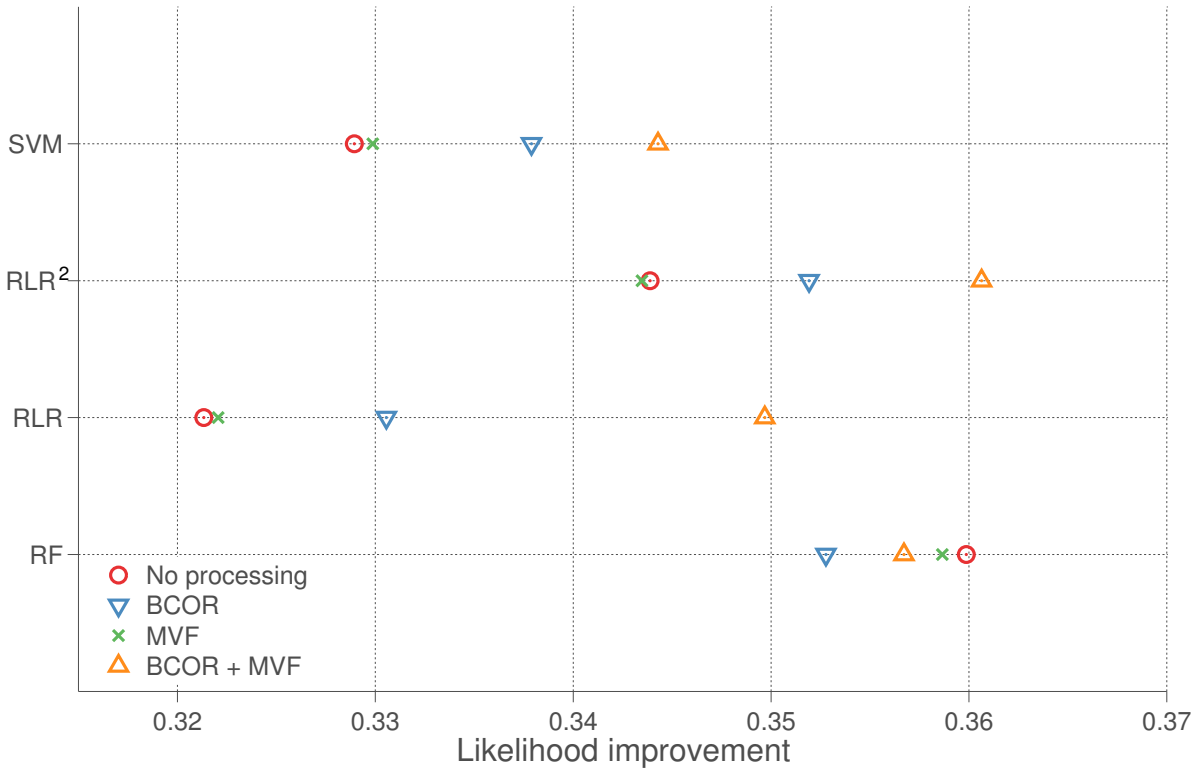

Figure 3.4: *Performance of the models developed on AO and evaluated on the test set as measured the $\mathcal{I}_{\mathcal{L}}$. The symbols correspond to different levels of preprocessing at the design matrix stage.*

The odds ratios for the RLR^2 model using BCOR and no missing value flags are shown in two tables. Note that these odds ratios are for the *second* best performing RLR^2 model as measured by the $\mathcal{I}_{\mathcal{L}}$. Table 3.12, shows the odds ratios for features directly measuring physiology. Table 3.13 shows odds ratios for the remaining features such as age, comorbidities and admission source. Note that these tables exclude odds ratios for diagnostic categories.

Figure 3.6 provides a graphical plot of the contributions to mortality risk from the original mean arterial pressure feature and the squared mean arterial pressure feature for the RLR^2 model with BCOR preprocessing and without missing value flags. The figure provides a visual interpretation of including both the original covariate and its square term in a regression model. The corresponding odds ratios are shown in Table 3.12.

	BCOR	MVF†	AUROC	$\mathcal{I}_{\mathcal{L}}$	SMR	$HL_{\hat{C}}$	B	B_{adj}
RF	×	×	0.891	0.360	0.96	50.1	**0.064**	**0.344**
RF	×	✓	0.890	0.359	0.97	41.0	0.065	0.341
RF	✓	×	0.889	0.353	0.97	98.3	0.065	0.336
RF	✓	✓	0.890	0.357	0.97	58.7	0.065	0.339
RLR	×	×	0.880	0.321	0.99	33.1	0.069	0.294
RLR	×	✓	0.880	0.322	**1.00**	**30.2**	0.069	0.297
RLR	✓	×	0.884	0.331	**1.00**	35.6	0.069	0.301
RLR	✓	✓	0.888	0.350	**1.00**	35.3	0.066	0.325
RLR^2	×	×	0.888	0.344	0.92	56.4	0.066	0.324
RLR^2	×	✓	0.888	0.343	0.92	53.8	0.066	0.325
RLR^2	✓	×	0.891	0.352	0.92	58.5	0.066	0.329
RLR^2	✓	✓	**0.893**	**0.361**	0.92	51.4	0.065	0.340
SVM	×	×	0.886	0.329	0.91	60.0	0.069	0.295
SVM	×	✓	0.887	0.330	0.91	62.0	0.069	0.296
SVM	✓	×	0.889	0.338	0.92	54.4	0.068	0.305
SVM	✓	✓	0.891	0.344	0.90	67.5	0.067	0.313

Table 3.11: *Performance of models evaluated on the test set of 27,243 observations. Each model was developed four times using the four combinations of no preprocessing, BCOR preprocessing, no missing value flags and missing value flags. The best combination is emphasised for each model, where best is defined by the $\mathcal{I}_{\mathcal{L}}$.*
†Use of an additional binary feature, per feature with missing data, which was only true for observations with a missing value.

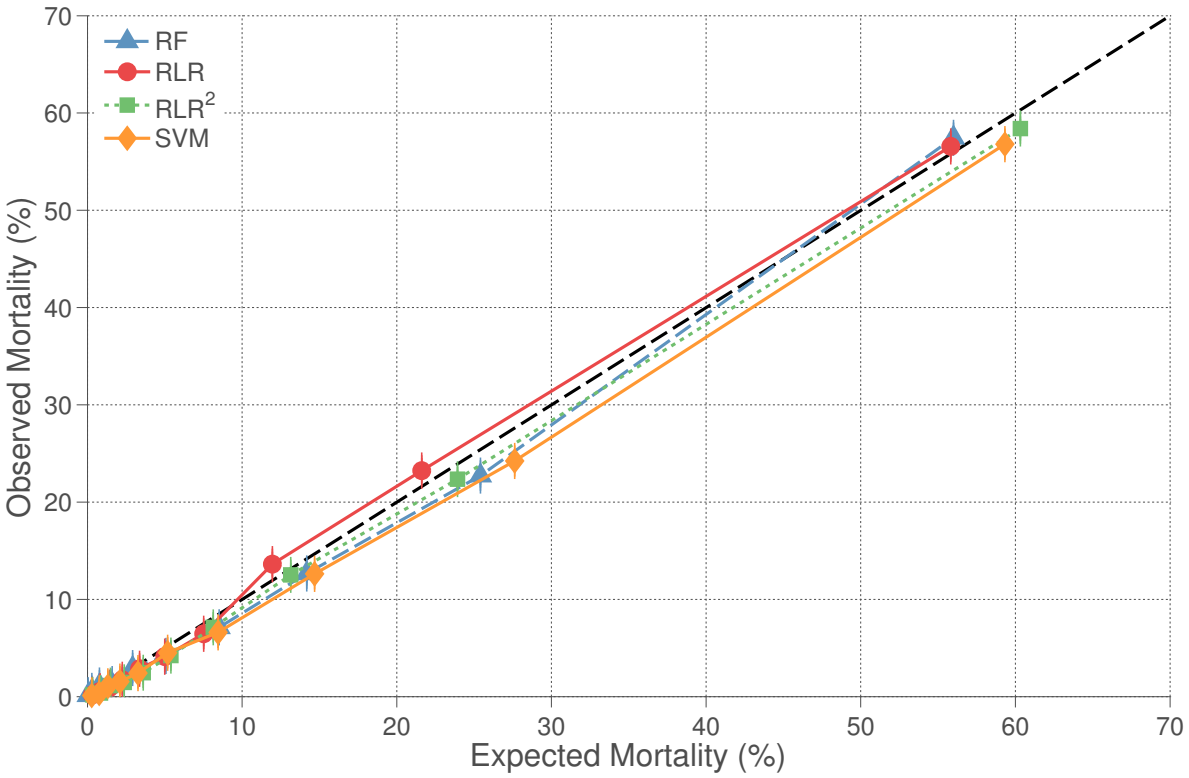

Figure 3.5: *Calibration curve of the best performing models developed on the AO dataset as measured by the $\mathcal{I}_{\mathcal{L}}$. Each colour/symbol combination represents a distinct model.*

Variable	Odds Ratio	Variable	Odds Ratio	Variable	Odds Ratio
Albumin	0.8373	Haematocrit^2	1.0377	Resp Rate^2	1.0662
Albumin^2	1.0539	Heart Rate	1.1455	Sodium	1.0036
Bilirubin	1.2555	Heart Rate^2	1.0926	Sodium^2	1.0899
Bilirubin^2	1.1163	MAP	0.8415	Temperature	0.9196
BUN	1.3262	MAP^2	1.2043	Temperature^2	1.1653
BUN^2	0.9744	$PaCO_2$	0.9155	Unable to acquire GCS	0.9890
Creatinine	0.9560	${PaCO_2}^2$	1.0777	Urine output ‡	0.7686
Creatinine^2	0.9764	PaO_2	1.2135	Urine output^2 ‡	1.1877
Emergency surgery	1.0616	PaO_2	0.9570	White blood cell count	1.0376
GCS	0.8106	PaO_2/FiO_2	0.9404	White blood cell count^2	1.0414
GCS^2	1.2898	$(PaO_2/FiO_2)^2$	1.0502	pH	0.8970
Glucose	1.0110	${PaO_2}^2$	1.0448	pH^2	1.1080
Glucose^2	1.0246	${PaO_2}^2$	1.0364		
Haematocrit	0.9855	Resp Rate	1.2420		

Table 3.12: *Odds ratios for covariates based upon physiology in the RLR^2 model with BCOR and no missing value flags. Odds ratios are for standardised, transformed variables and represent the increased odds of mortality for a one standard deviation increase in the covariate. For binary covariates, odds ratios represent the increase in risk if the observed value is true.*
BUN - Blood urea nitrogen. GCS - Glasgow coma scale. MAP - Mean arterial pressure. Resp Rate - Respiration rate.
△ *Missing value indicator for the stated covariate.*
^{2} *Represents the normalised variable squared.*
‡*Total over the first 24 hours.*

Variable	Odds Ratio	Variable	Odds Ratio	Variable	Odds Ratio
AIDS	1.6321	ICU - Trauma	0.8403	Source - Other Hospital	1.1944
Age	1.6143	Immunosuppressed	1.3146	Source - Recovery Room	0.7995
Age^2	1.0692	Intubated	0.9142	Source - SDU	1.1481
Any comorbidity	1.0193	Lymphoma	1.7744	System - CABG	0.2735
Cirrhosis	1.4316	Multiple myeloma	1.9059	System - Genitourinary	0.6312
Elective surgery	0.6911	Pre-ICU length of stay	1.1810	System - Hematological	1.1626
Gender: Female	0.9188	Pre-ICU length of stay^2	1.1496	System - Metabolic/Endocrine	0.8263
Hepatic failure	1.4981	Race: Black	1.0846	System - Musculoskeletal/Skin	1.0238
Hosp - COTH[1]	1.1224	Race: Caucasian	1.1327	System - Respiratory	1.0447
Hosp - Unknown	0.9411	Race: Latino	1.0179	System - Transplant	0.3266
Hosp - Small teaching hospital	0.8331	Source - Direct Admission	0.9506	System - Trauma	1.0047
ICU - Cardio-Thoracic	0.8081	Source - Floor	1.0802	System - Unknown	0.0705
ICU - Coronary	1.1473	Source - ICU Transfer	1.0220	Thrombolytic therapy	0.5518
ICU - Medical	1.1501	Source - Missing	0.9740	Tumour with metastasis	2.4462
ICU - Missing	0.8908	Source - Missing	9.8003	Ventilation status	1.3110
ICU - Mixed	1.0365	Source - Operating Room	0.6940	Worst RR during ventilation	1.1280
ICU - Surgical	0.8496	Source - Other	1.3631	Any active treatment	1.4799

Table 3.13: *Odds ratios for covariates not directly based on physiology in the RLR^2 model with BCOR and no missing value flags. Odds ratios are for standardised, transformed variables and represent the increased odds of mortality for a one standard deviation increase in the covariate. For binary covariates, odds ratios represent the increase in risk if the observed value is true.*
△ Missing value indicator for the stated covariate.
‡Community operated teaching hospital.
[2] Represents the normalised variable squared.

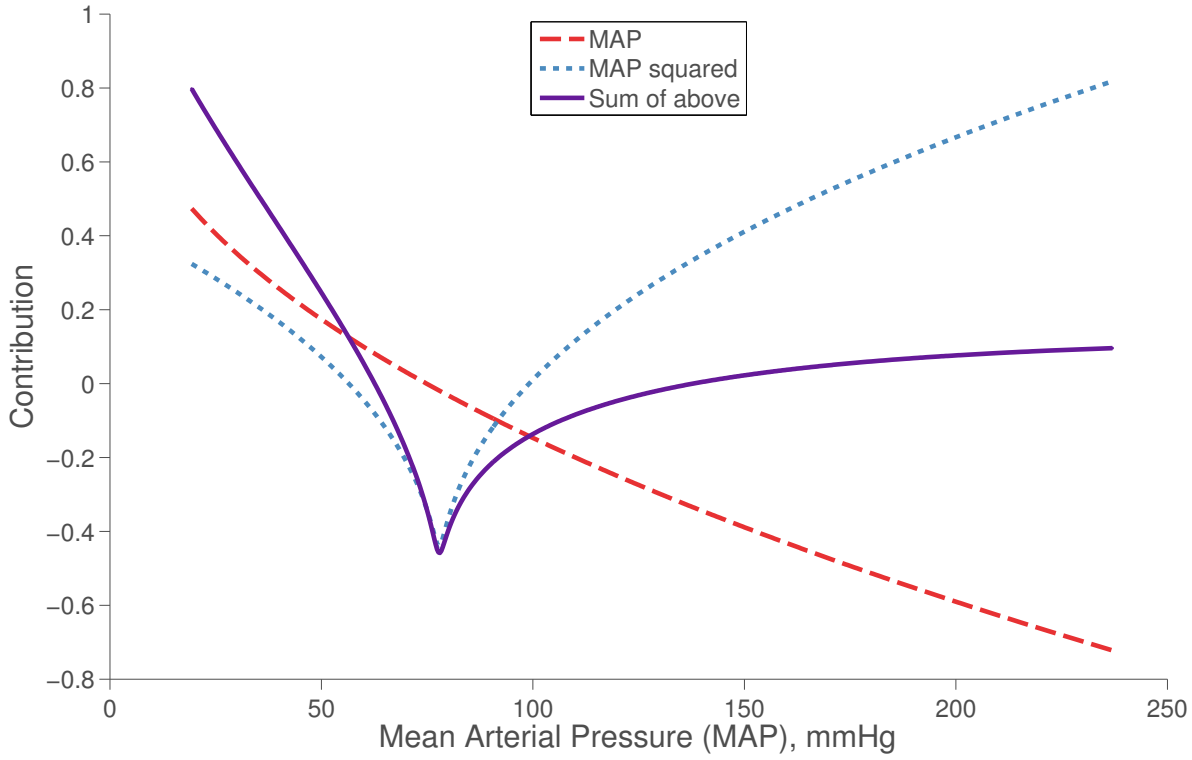

Figure 3.6: *Contribution of the original mean arterial pressure variable and its squared covariate in the RLR^2 model with BCOR and without missing value flags.*

3.5 Discussion

In the previous chapter, the SVM had the highest performance as measured by the AUROC and $\mathcal{I}_{\mathcal{L}}$ though clinically it was difficult to ascertain if this was sufficient increase in performance to warrant the transition to a non-linear "black box" model. When evaluated on a large multi-center database with over 50,000 observations, the performance difference between linear to non-linear models is not substantial. The RLR^2 with both BCOR preprocessing and the addition of missing value flags had the highest discrimination of all models (AUROC = 0.893), though it was only 0.002 higher than the best performing SVM and RF. In the previous chapter, the SVM instead had an AUROC that was 0.002 higher than the equivalently processed RLR^2. When ranking the models based upon the $\mathcal{I}_{\mathcal{L}}$, a measure of both the calibration and discrimination of the model, the RLR^2 had the highest $\mathcal{I}_{\mathcal{L}} = 0.361$ and consequently the highest ranking. The best performance of the remaining models was $\mathcal{I}_{\mathcal{L}} = 0.360$ for the RF, $\mathcal{I}_{\mathcal{L}} = 0.350$ for the RLR and $\mathcal{I}_{\mathcal{L}} = 0.344$ for the SVM. The conclusion from this empirical study is that the simplest regression model, with or without the addition of square terms, performed competitively with the other non-linear classifiers.

Figure 3.6 shows a very interesting relationship which has been modelled between mean arterial pressure (MAP) and mortality. The model contains both the MAP values and the square of the MAP values (after mean subtraction). As can be seen, low MAP values are considered high risk for both the linear MAP term and the quadratic MAP term. As MAP increases, the quadratic term risk reduces to its lowest value which occurs at the mean of the distribution (recall that calculation of the square term first involves mean subtraction). As MAP increases past its mean value, the quadratic MAP term begins to increase as well. Meanwhile, the linear MAP term decreases across the entire range of MAP. The final relationship modelled assigned a very high risk to low MAP values and a reduced (but non zero) risk to high MAP values. The addition of the square terms has allowed for a non-linear modelling of the relationship between MAP and risk of mortality. Furthermore, this relationship correlates with clinical knowledge as low MAP measurements are the hallmark of septic shock, a severe acute illness [128]. Conversely, high MAP carries a non-zero risk, but this risk is lower than that of hypotension. Note

also that the transformation of the data during BCOR has caused the contribution of risk from the MAP feature to be proportional to a monotonic power transformation of MAP (the decaying red curved line in Figure 3.6) instead of proportional to the MAP value itself (which would result in a straight diagonal line with a fixed slope).

A limitation to the analysis is the lack of regularisation or feature selection for the models aside from the logistic regressors. Inherent to the RLR is the learning of a regularisation term, which controls the number of covariates to be included in the model. There is no direct equivalent of this parameter in either the RF or the SVM models (though the capacity parameter C can be considered as an L^2 regularisation term). As such, the noise in irrelevant features may not be handled as well by the non-linear models as compared to the RLR. This is evident across the RF models: the addition of missing value flags (which do not add any flexibility in a tree based model) lowered the performance of the model both with and without preprocessing. In Chapter 2, the use of regularisation for the linear models had a much larger impact as the data size was much smaller relative to the number of covariates (4,000 observations to 198 features). Here, the ratio of data to covariates is much higher (53,758 observations to 197 features), and so L^1 regularisation to control model complexity plays less of a role. It remains a worthwhile investigation to determine if the addition of regularisation penalties (or feature selection) to either the SVM or the RF would improve performance further.

Another factor is the ease of hyperparameter optimisation for the RLR and RLR^2 models. For the RLR^2 and RLR model, a single hyperparameter is optimised during the training phase, and fast algorithms to estimate all possible values of this hyperparameter exist (i.e. least angle regression or LARS [129]). Conversely, the SVM has two hyperparameters and the RF has two hyperparameters, and the existence of twice as many hyperparameters squares the size of the search space (though an algorithm similar to LARS has been proposed for the SVM [130], and this algorithm would ease estimation of the capacity hyperparameter C).

A second explanation for the lackluster performance improvement in the more complex models is a lack of information in the non-linear interactions of the features. While it is readily obvious that the interaction between blood pressure and heart rate is a

Negative coefficients		Positive coefficients	
Feature	Odds ratio	Feature	Odds ratio
Intubated △	0.0005	Glasgow Coma Scale[2] △	1.8539
Glucose △	0.2329	pH △	2.8941
Albumin △	0.3915	Heart Rate △	2.9311
Haematocrit △	0.3984	MAP △	11.4731
Creatinine[2] △	0.5668	PaO_2 △	39.6127

Table 3.14: *The five highest and lowest odds ratios in the best RLR^2 model developed on the AO dataset.*
△ *Missing value flag*
[2] *Covariate was squared before input into the model.*

key component of assessing a healthy state, this does not necessarily carry over to the interaction of daily summary values for these parameters. In fact, the low sampling rate of the data (once per day), combined with the asynchronous measurement of these values (e.g. the worst heart rate may not be measured at the same time as the worst blood pressure), may contribute to removing any potential information in the feature interactions. In the development of the MPM-II model, Lemeshow *et al.* found no significant interactions warranting inclusion in the final model [11]. However, Harrison *et al.* did discover beneficial interactions in the development of the ICNARC model, primarily between diagnosis and an aggregate of physiology [34]. This may be the explanation for the gain of performance in the RF model as compared to the RLR model, though it is worth noting that this was not a physiologic interaction and may have simply aided the regression adjust for deranged physiology which is normal for a given condition (e.g. high respiratory rates for patients with acute respiratory distress syndrome).

Table 3.14 shows the most extreme odds ratios for covariates in the best performing RLR^2 model with BCOR and missing value flags. The majority of the features are missing value flags, some of which have extremely high or low odds ratios. This is likely due to the method of preprocessing removing extreme values which were in fact not outliers, but rather true extreme physiology. After their removal, the missing value flag now represents extreme severity and consequently has a very high odds ratio. This further explains why missing value flags on their own do not improve model performance substantially, but after preprocessing provide significant improvement in model fit. For this reason, Tables 3.12 and 3.13 present more meaningful odds ratios associated with

the best performing RLR^2 model which did *not* contain missing value flags.

A key observation is the marked change in performance which occurred when preprocessing the dataset to remove artefacts. As extremely low or high values will be linearly combined with other covariates in a regression model their effects on model performance can be substantial. The use of the Box-Cox transformation and normalisation procedures likely reduced the potential issue caused by large outliers. The removal of these outliers if they were statistically significantly deviating from the parametrised normal distribution of each feature appears to be very beneficial. For the RLR the $\mathcal{I}_{\mathcal{L}}$ increased from $\mathcal{I}_{\mathcal{L}} = 0.321$ to $\mathcal{I}_{\mathcal{L}} = 0.331$ and resulting in a model superior to an SVM without preprocessing ($\mathcal{I}_{\mathcal{L}} = 0.329$). The SVM also benefited from the preprocessing which is not entirely unexpected as the model is sensitive to scaling of the input variables. The RF was actually worsened by the preprocessing of the BCOR. It is likely that the flexibility of the tree based RF allowed it to better handle outliers, which is not the case with the other models. The empirical conclusion is that while linear models can have competitive performance with more complicated methods, preprocessing of the data and regularisation are imperative.

The study here has limitations. First, it is focused on data collected at hospitals with the APACHE system installed, and is thus a non-random selection of ICUs in the United States. Secondly, the features used in the models are snapshots of a patient state. More advanced machine learning techniques are likely limited by the features available, rather than being inappropriate for mortality prediction in general. Future work should focus on: i) extracting features from time series and demonstrating that they provide additional information over the benchmark models presented here, and/or ii) application of non-feed forward approaches for mortality prediction. Additionally, a set of CABG patients remained in the data which accounted for less than 2% of patients (see Tables 3.9 and 3.10). While a small fraction, these patients should not have been included in the analysis.

As simpler models have been demonstrated to perform equivalently to more complex machine learning methods, it becomes worthwhile to consider the complexity versus performance trade off inherent to these models. In particular, while electronic data

management systems are on the rise, there still exist a number of institutions which do not have the facility to automatically calculate patient risks. This additionally holds true for clinical trials which run auxiliary to but in partnership with these institutions. Severity scores used for these purposes are vastly outdated, with the most commonly used models (APACHE II and SAPS II) being developed in 1985 [53] and 1993 [9]. Thus there is an opportunity for developing a simple model which has competitive performance to even the most sophisticated of machine learning techniques. The subsequent chapters detail both the development of a score of this form and empirical evaluation of its performance.

Chapter 4

OASIS: Development of a parsimonious severity score

It is pointless to do with more what can be done with fewer.

William of Occam

This chapter overviews the development of a parsimonious severity score for patients in the ICU. The derivation data set, collected from multiple institutions across the continental United States, is the same as the AO dataset used in Chapter 4. A Genetic Algorithm (GA) is described and used to select a subset of the available features. A customised Particle Swarm Optimization (PSO) method is used to develop the score itself. The techniques are applied using ICU mortality as the predicted outcome. The newly developed score is assessed univariately as the only predictor in a regression model and in a larger regression model similar to APACHE IV. The score is compared with the APS III directly and with APACHE IV. The utility and potential of the new score is then discussed.

4.1 Genetic algorithm

The theoretical groundwork of GAs was developed in the late 1950s [131], and then used to solve a variety of practical engineering problems in the early 1970s [132]. GAs are

based upon genetic recombination, the process by which the information contained in two parent genomes is exchanged. In genetic recombination, strands of DNA are paired based upon location in the genome and portions of the two parent DNA strands are exchanged, resulting in two hybrid genomes. This process is commonly referred to as crossover, as portions of one parent's DNA are crossing over to the other's. Occasionally, an error in the duplication process occurs, and the information content is changed. These small changes in the new genome are referred to as mutations. Organisms which utilise genetic recombination increase the heterogeneity of their genomes, and increase the likelihood of receiving a beneficial trait coded for by the new genome. However, they also increase the likelihood of receiving a detrimental trait. Over generations of these organisms, the ones with detrimental traits are less likely to reproduce, whereas the ones with beneficial traits are more likely. One can consider the overall balance of beneficial and detrimental traits to constitute an organism's fitness. Crossover, mutation and selection based upon fitness constitute a naturally occurring process which has generated organisms with remarkable fitness, and it is this process which the GA aims to replicate.

The GA in this report is used solely for feature selection, and the final output of the algorithm is a subset of features which, when used as inputs to a model, provide equivalent prediction efficacy compared to models which use the full feature set. Mathematically, if we have a matrix of data $\mathbf{X}$ and target vector $\mathbf{y}$, we wish to maximises $f(\mathbf{y}, g(\mathbf{X}))$ given a design matrix $\mathbf{X}$ which uses some subset of features $j \in 1, ..., D$. Here $f(\cdot, \cdot)$ is our chosen fitness function and $g(\cdot)$ is a learnt mapping from the data to the target of interest (e.g. a logistic regression).

A single subset of features is represented as a binary vector and referred to as a "gene". That is, each gene is a vector v of length D where $v_j = 1$ indicates inclusion of the j^{th} feature and $v_j = 0$ indicates exclusion of the j^{th} feature.

The algorithm begins by randomly initialising a set of genes, referred to as a population. After generating the initial population, each gene is then evaluated by a predefined fitness function. The genes are sorted according to their fitness, and a percentage of the poorest performing genes are removed from the population. The remaining genes are then modified using crossover followed by mutation to produce a new population of the

same size.

The crossover operation involves splitting the top performing genes into segments randomly, and combining segments from distinct genes to create new genes. That is, given two genes v_i and v_k, we randomly select a split point $j \sim \mathcal{U}\{1, D\}, j \in \mathbb{Z}$ where $\mathcal{U}\{1, D\}$ represents a discrete uniform distribution spanning the range $[1, D]$ inclusive. After selection of the crossover point j, two new genes are generated as follows:

$$\begin{aligned} \tilde{v}_i &= \{v_{i,1}, v_{i,2}, ..., v_{i,j}, v_{k,j+1}, ..., v_{k,D}\} \\ \tilde{v}_k &= \{v_{k,1}, v_{k,2}, ..., v_{k,j-1}, v_{i,j}, ..., v_{i,D}\} \end{aligned} \tag{4.1}$$

where $\tilde{v}_i$ and $\tilde{v}_k$ are the "children" of parent genomes v_i and v_k. A visual example of single point crossover is shown in Figure 4.1.

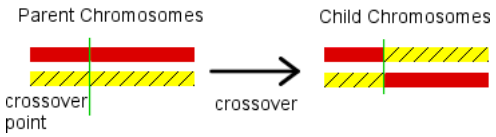

Figure 4.1: *Visualization of single point crossover operation. Selected genes swap any number of continuous segments depending on the number of crossover points.*

Mutation consists of switching the state of a small proportion of the genes from true to false or from false to true in order to maintain diversity and increase exploration of the solution space [133]. That is, for each child gene $\tilde{v}_i$, the elements are updated as:

$$\tilde{v}_{i,j} = \begin{cases} 1 - \tilde{v}_{i,j} & , m < M \\ \tilde{v}_{i,j} & , \text{otherwise,} \end{cases} \tag{4.2}$$

where m is a randomly generated for each $\tilde{v}_{i,j}$ as $m \sim \mathcal{U}(0, 1)$ and M is a parameter defining the mutation rate.

This ranking, crossover and mutation process is repeated until a termination criterion is met. Termination criteria include an error gradient below a predefined threshold or a maximum number of iterations. In an elitist GA framework, as used in this work, a percentage of the best performing genes are not modified between consecutive iterations. The gene with the best fitness after the final iteration contains the algorithm's selected

features. The GA optimisation is often repeated to ensure the best gene is consistent.

4.2 Particle swarm optimization

Particle swarm optimisation (PSO) is an interesting development in the field of optimisation based on observation of birds flocking [134]. The general methodology involves iteratively encouraging a set of vector positions to migrate toward their respective prior bests as well as the global set best.

Formally, the algorithm begins with a random initialisation of two matrices: one set of position vectors, hereafter referred to as particles or $\mathbf{X}$, and another set of iterative update vectors, hereafter referred to as velocities or $\mathbf{V}$.

$$\mathbf{X} = \begin{vmatrix} x_{1,1} & ... & x_{1,D} \\ . & & . \\ . & & . \\ x_{N,1} & ... & x_{N,D} \end{vmatrix} \quad \mathbf{V} = \begin{vmatrix} v_{1,1} & ... & v_{1,D} \\ . & & . \\ . & & . \\ v_{N,1} & ... & v_{N,D} \end{vmatrix}$$

Here there are N particles and velocities, and each matrix has D dimensions. N is a user defined parameter, while D is partially dependent on the input data dimensionality and the problem formulation. For example, if two particle elements are used for each feature, then D is twice the input data dimensionality. The performance of the particles is measured by a fitness function, $f(\cdot)$.

A new matrix $\mathbf{P}$ stores the prior best position of each respective particle, i.e. the best performance across past iterations. If the current iteration of the algorithm is L, then we define the matrix $\mathbf{P}$ as:

$$\mathbf{P} = \begin{vmatrix} \{x_1^{(k)} \mid f(x_1^{(k)}) \geq f(x_1^{(l)}) \; \forall \; l = 1, ..., L\} \\ \\ \\ \{x_N^{(k)} \mid f(x_N^{(k)}) \geq f(x_N^{(l)}) \; \forall \; l = 1, ..., L\} \end{vmatrix},$$

where we can interpret the i^{th} row of $\mathbf{P}$ equalling the vector position $x_i^{(k)}$ given that $x_i^{(k)}$ had the highest fitness of all prior iterations.

Finally, the matrix G contains the vector $\mathbf{P}_i$ with the best performance as measured by the fitness function $f(\cdot)$ in all rows, i.e. all rows are identical and equal to the best particle position. The repetition in the matrix G is purely for notational convenience.

The algorithm begins by randomly initialising $\mathbf{X}$ and $\mathbf{V}$ as $\mathbf{X} \sim \mathcal{U}(a_1, b_1)$ and $\mathbf{V} \sim \mathcal{U}(a_2, b_2)$, where a and b are user defined boundaries for the the particles' position and velocity. $\mathbf{P}$ is then set equal to $\mathbf{X}$ matrix. The $\mathbf{G}$ matrix is updated to contain the best performing particle in matrix $\mathbf{P}$ repeated across all rows. Thus at the start of the first iteration, the historical best particle position and the current particle position are identical for each particle. Each subsequent iteration involves updating $\mathbf{V}$ as follows:

$$\mathbf{V}^{(k+1)} = \mathbf{V}^{(k)} + \mathbf{B}^{(k)} \circ (\mathbf{P}^{(k)} - \mathbf{X}^{(k)}) + \mathbf{C}^{(k)} \circ (\mathbf{G}^{(k)} - \mathbf{X}^{(k)}), \tag{4.3}$$

and then updating $\mathbf{X}$ as:

$$\mathbf{X}^{(\mathbf{k+1})} = \mathbf{X}^{(k)} + \mathbf{V}^{(k)} \tag{4.4}$$

Here $\mathbf{B}^{(k)}$ and $\mathbf{C}^{(k)}$ are matrices with each element newly generated at each iteration according to the following uniform distributions: $\mathbf{B}^{(k)} \sim \mathcal{U}(0, \alpha)$ and $\mathbf{C}^{(k)} \sim \mathcal{U}(0, \eta)$ where α and η are hyperparameters which act to control the degree of local versus global optimisation. The matrices are set to uniform random variables instead of constants as it has been shown to improve convergence speed [135]. In this work the ranges of the constants' distribution are identical in order to weight global and local optimisation equally in expectation, i.e. $\alpha = \eta$. The particle velocities in Equation 4.3 are calculated independently for each particle from $i = 1, 2, ..., N$. Thus, each iteration encourages particles to drift toward respective prior best positions (P_i) and the single global best position (G_i). At the end of an iteration, if any individual particle position has a better fitness than the currently stored prior particle best (P_i), the prior particle best is updated to the better position. This update can be stated for the i^{th} particle on the k^{th} iteration as follows:

$$p_i^{(k)} = \begin{cases} x_i^{(k)} & \text{if } f(x_i^{(k)}) \geq G_i^k \\ \\ p_i^{(k)} & \text{otherwise.} \end{cases} \tag{4.5}$$

This process repeats until a termination criterion is reached; either a desired fitness or a maximum number of iterations. Pseudocode of the particle swarm algorithm is provided in the appendix Section B.3.

4.2.1 Particle mapping and fitness

The PSO implemented in this work was very similar to the general method presented, and the primary customisation is in the mapping between a particle vector X_i and the outcome to be predicted y_i. Each particle element represented the value assigned to data residing between two quantiles for single input feature. In this way, each particle assigns a score for the data after it has been binned into a set of fixed quantiles.

The severity score calculation for all observations in design matrix $\mathbf{X}$ can be represented as:

$$\mathbf{s} = m(\mathbf{X}_1) + m(\mathbf{X}_2) + ... + m(\mathbf{X}_D), \tag{4.6}$$

where D represents the total number of features, $\mathbf{X}_j$ is the j^{th} feature in the design matrix, $m(\mathbf{x})$ is a function that maps the feature j to a score representing severity and $\mathbf{s}$ is a vector containing the final score for each row in $\mathbf{X}$. An example of the particle mapping $m(\mathbf{x})$ for a single feature, heart rate, is shown in Equation 4.7.

$$\begin{aligned} m_j(\mathbf{x}) \quad &= 3 \times H[\mathbf{x}] - 3 \times H[\mathbf{x} - 60] \\ &+ \ 0 \times H[\mathbf{x} - 60] - 3 \times H[\mathbf{x} - 100] \\ &+ \ 2 \times H[\mathbf{x} - 100] - 3 \times H[\mathbf{x} - 120] \\ &+ \ 5 \times H[\mathbf{x} - 120]. \end{aligned} \tag{4.7}$$

Here $H[x - n]$ represents the Heaviside function, defined as follows:

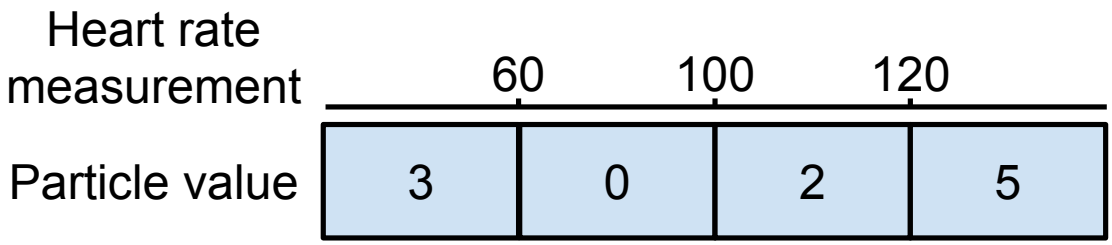

Figure 4.2: *An example of the mapping between the heart rate feature and the associated elements in a particle. Heart rate values between 0 and 60 receive a value of 3, values between 60 and 100 receive a value of 0, and so on. The lower values are sourced from the optimised particle positions.*

$$H[x-n] = \begin{array}{ll} 0, & x \leq n, \\ 1, & x > n \end{array} \tag{4.8}$$

Thus if observation X_i had a heart rate = 30 then $m_j(X_i) = 3$, where we have assumed the j^{th} feature is heart rate. One functional mapping, $m_j(x)$, exists for each feature $j = 1, ..., D$. Figure 4.2 provides an illustrative example of Equation 4.7.

The particle mapping described here is intentionally designed to be very similar to prior severity scores such as the APS III [7], and allows for direct optimisation of a similarly structured severity score. Once the severity score $\mathbf{s}$ has been calculated the AUROC of $\mathbf{s}$ with respect to ICU mortality is output as the particle fitness.

4.3 APACHE Outcomes dataset preparation

The data used to develop the score is a subset of the AO dataset described in Section 3.1. An overview of the data collection is provided in Figure 3.1. Section 3.1 details the data quality assurance performed prior to acquisition. Exclusion criteria consisted of patients with burns, ICU stay lasting <4 hours and patients aged <16 years. Post-transplant patients were excluded except in the cases of hepatic or renal transplantation. Observations where the patient had been readmitted to the ICU were removed in the creation of the clean data repository in Figure 3.1. These exclusion criteria were applied before data acquisition.

Similar to Chapter 3, the data were split temporally. All observations occurring between 2010-2011 were assigned to the test set[1], while all observations occurring between 2007-2009 were placed in the development dataset. This segmentation allows for evaluation of how the model would perform on "future" data, a more rigorous evaluation of performance [136] due to significant model drift observed in prior ICU severity scoring systems [42]. Patients in this subset were admitted to 86 ICUs at 49 hospitals across the United States, all of which had the APACHE IV system (Cerner Corporation, Kansas City, MO). There were 81,087 admissions in this subset, and this was the cohort used for development of the severity score. Preprocessing after data acquisition as described in Chapter 2 (such as with BCOR) was not performed.

A total of 37 features of interest were selected from the AO dataset. These features were chosen as they: i) are routinely collected in the ICU and ii) do not use information that is not feasibly utilised for prognostication on admission (e.g., length of ICU stay). Furthermore, diagnosis is not included as it would require any future user of the severity score to code all patients according to the diagnostic categories. This is a burdensome task and has been indicated as a possible barrier to APACHE IV's uptake in the general community [137]. A list of the features included in the severity score optimisation is shown in Table 4.1.

Observations with missing data for daily urine output, GCS, respiration rate, mean arterial pressure and heart rate were deleted. Missing values for other parameters were replaced with mean imputation using the training dataset to calculate the respective means. The feature PaO_2/FiO_2 was imputed with a fixed value of 385.7143, rather than the mean of the training data.

4.4 Score development

A GA as described in Section 4.1 was used in conjunction with the PSO method described in Section 4.2 to develop a severity of illness score. The development dataset was segmented into a training set containing 70% of the data and a validation set containing the remaining 30%. Each iteration of the GA proceeded as follows. First, a new score

[1]The test set is never used for model development.

Variable	Format
Gender	Male (reference), Female
Race	White (reference), Black, Hispanic, Other
Age	Continuous measure
Physiologically derived variables	The most abnormal value on ICU day 1 for the following variables: pulse rate, mean blood pressure, temperature, respiratory rate, PaO2:FiO2 ratio (or A-aD02 for intubated patients with FIO2 $>$ 0.5), PaO2, PaCO2, haematocrit, white blood cell count, creatinine, urine output, blood urea nitrogen, sodium, albumin, bilirubin, glucose, pH and neurological abnormalities based on Glasgow Coma Scale.
Chronic health items	AIDS, cirrhosis, hepatic failure, immunosupression, lymphoma, leukemia or myeloma, metastatic tumor. Not used for elective surgery patients.
Length of stay before ICU Admission	Square root of time from hospital admission to ICU admission (in fractional days)
Patient type	Medical, elective surgery, emergency surgery
Admitted to ICU from general floor unit	Binary
Ventilated on day 1	Binary
Unable to Assess Glasgow coma score due to sedation/paralysis on day 1	Binary
Patient received thrombolytic therapy for an acute myocardial infarction	Binary
Mortality before hospital discharge	Binary
Mortality before first ICU discharge	Binary
ICU length of stay, first admission	Continuous measured, truncated at 30.0 days.

Table 4.1: *Information captured for model development and evaluation.*

determined by the PSO was developed using features selected by a single gene in the GA population. This was repeated for all genes, resulting in an optimised score for each gene in the population which utilised only the features selected by that individual gene. The quantiles used in the PSO were determined before optimisation by calculating the 2.5^{th}, 25^{th}, 50^{th}, 75^{th} and the 97.5^{th} percentile of the input data for each continuous feature. Binary features were split into two quantiles; true and false. Consequently there were six score values optimised for each continuous feature and two values optimised for each binary feature. The size of the population was set to twenty four particles, though this has been shown to have a small impact on overall performance regardless of the number of input dimensions [135].

The PSO used the training set and optimised the AUROC for predicting ICU mortality. PSO termination occurred after a maximum of 500 iterations or if the AUROC did not improve by 0.0005 across 250 iterations. The final score optimised at the end of the PSO was evaluated on the validation set, and this provided the evaluation of fitness for a single gene. The PSO was applied to each gene in this way, resulting in single fitness value for each genere (the AUROC of the optimised severity score on the validation set).

The selection, crossover and mutations operations were then applied to the GA population. This resulted in a new population, and the GA continued to generate new populations using the PSO to estimate a gene's fitness. This process was repeated until the AUROC of the GA did not improve by at least 0.005 across 10 generations or the maximum of 30 generations had been reached. The best performing severity score across the entire GA population, as measured on the validation set, was then extracted and evaluated on the test set. This methodology was used to create the Oxford Acute Severity of Illness Score (OASIS) [138]. The overall process is depicted in Figure 4.3. Hyperparameters for the GA are provided in Table 4.2 and hyperparameters for the PSO are provided in Table 4.3.

OASIS was pruned after development. Pruning involved collapsing multiple consecutive categories which had the same score value. For example, if two ranges of temperature corresponding to [34.5,35.0) and [35.0,35.5) were both assigned a score value of 3, then these two ranges were collapsed into a single one spanning [34.5,35.5). This pruning

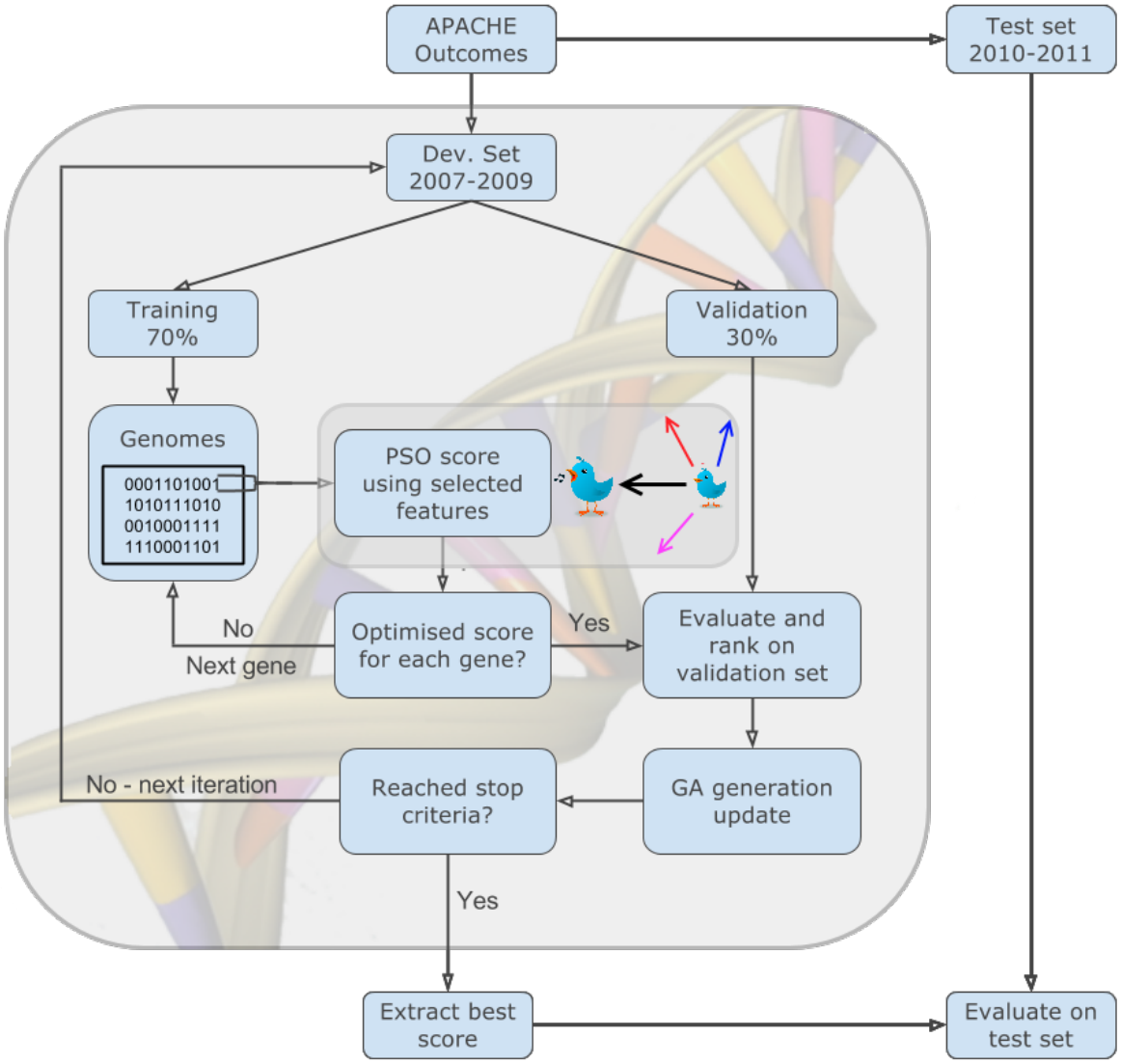

Figure 4.3: *Development process of the severity score. The grey box represents actions taken within the genetic algorithm (GA). The GA generation update includes the processes of selection, crossover and mutation. "Dev. Set" represents the development dataset.*

Parameter	Specification
Fitness Function	PSO
Cost function	AUROC
Population size	32
Population dimensions	37
Elitism proportion	Top 10%
Crossover proportion	Top 45%
Mutated population proportion	20%
Mutation rate	6%
Maximum number of generations	30
Minimum Error Gradient	0.005
Error Gradient	10 generations

Table 4.2: *Summary of the hyperparameters for the GA.*

Parameter	Specification
Maximum Iterations	500
Population Size	24
α	2.05, i.e. $\mathbf{B} \sim \mathcal{U}(0, 2.05)$
η	2.05, i.e. $\mathbf{C} \sim \mathcal{U}(0, 2.05)$
Inertia decay weight[1]	$0.9 \rightarrow 0.4$
Inertia weight iterations	400
Velocity decay weight[1]	$1 \rightarrow 0.2$
Velocity weight iterations	250
Minimum Error Gradient	5e-4
Error Gradient Iterations	250 iterations
Position threshold	[0,10]
Velocity threshold	[-2,2]

Table 4.3: *Summary of the hyperparameters for the PSO.*
[1] Value decrements linearly from its highest to lowest value across the specified number of iterations.

process does not modify the actual discrimination of the score or the calculated score values: it merely simplifies the representation of the score.

To provide an overview of how frequently a set of features were selected, the entire GA was repeated 100 times, and the average proportion of each feature retained in the final genome was calculated.

Logistic regression was used to map each observations' integer score into a probability of both ICU and hospital mortality. Two logistic regressions using OASIS as a covariate were developed for each outcome, i.e. two regressions for hospital mortality and two regressions for ICU mortality. The first was a regression with OASIS as the only independent variable and either hospital or ICU mortality as the dependent variable.

Independent variables	Dependent variable(s)	Number of covariates
Hospital mortality and ICU mortality	OASIS	1
Hospital mortality and ICU mortality	APS III	1
Hospital mortality and ICU mortality	OASIS, age, ventilation status, pre-ICU LOS, GCS, unable to acquire GCS flag, admission type binary indicators, comorbidity binary indicators, diagnostic binary indicators	134
Hospital mortality and ICU mortality	APS III, age, ventilation status, pre-ICU LOS, GCS, unable to acquire GCS flag, admission type binary indicators, comorbidity binary indicators, diagnostic binary indicators	134

Table 4.4: *Table listing the recalibrations performed using the OASIS and the APS III. Each row represents a configuration, and two recalibrations are performed: once for hospital mortality and once for ICU mortality. All recalibrations were performed using the entire development dataset.*

The second model was a recalibration of the APACHE IV model [35], except the APS III was substituted with OASIS. The covariates in this model (excluding the physiology score) were age, pre-ICU length of stay, PaO_2/FiO_2, binary indicator variables for comorbidities[2], admission from the floor, admission from another hospital, emergency surgery, thrombolytic therapy in the first 24 hours, a binary variable indicating inability to acquire the GCS, ventilation in the first 24 hours and primary admission diagnosis coded as binary variables.

To provide a comparison with standard clinical practice, additional logistic regression models were built for the APS III. The first set of logistic regression models utilised the APS III as the only independent variable and either hospital mortality or ICU mortality as the dependent variable. The second set utilised all the covariates listed prior, and was equivalent to a recalibration of APACHE IV on the development dataset. An overview of the models is shown in Table 4.4.

The single feature regressions use either OASIS or the APS III as the *only* input.

[2]The comorbidities were as in APACHE IV and included: HIV/AIDS, hepatic failure, lymphoma, tumour with proven metastases, multiple myeloma, immunosuppression and cirrhosis.

The multivariate models correspond to those using either OASIS or APS III (but not both) as a single covariate in a model with many more covariates. When the APS III is used as the covariate, this model can be thought of as a re-calibration of APACHE IV, as all the covariates are identical.

4.5 Results

4.5.1 Data demographics

Daily urine output was missing for 6,903 admissions, GCS was missing for 1,700 admissions, respiration rate was missing for 5 admissions, mean arterial pressure was missing for 4 admissions and heart rate was missing for a single admission. These admissions (8,613) were removed from the dataset. This resulted in a cohort of 72,474 admissions: 48,856 in the development dataset and 23,618 for the test dataset. The demographics of the development and test set are presented with statistical comparison of their similarity in Table 4.5. Patients in the test had shorter ICU stays, lower hospital mortality, less instances of elective surgery and were more frequently sedated or paralysed which prohibited the acquisition of the GCS.

	Development	**test**	**p-values**
Age	61.2 ± 17.8	61.8 ± 17.6	< 0.001
APS III	40.0 ± 25.4	39.8 ± 24.2	0.161
Female	45.5	46.2	0.069
White	74.6	83.6	< 0.001
Presence of Chronic Health Condition†	29.6	40.2	< 0.001
Pre-ICU length of stay	67.6	63.3	< 0.001
Admission from floor	11.8	9.46	< 0.001
Ventilated	35.8	32.8	< 0.001
Unable to acquire GCS	7.17	11.9	< 0.001
Emergency Surgery	5.72	6.48	< 0.001
Elective Surgery	23.3	19.0	< 0.001
ICU length of stay	3.96 ± 6.52	3.39 ± 5.19	< 0.001
Died in ICU	7.32	7.50	0.372
Died in hospital	11.7	11.0	0.009

Table 4.5: *The demographics of the data set. Continuous features are shown ± one standard deviation. Binary features are shown as a percentage. The p-values in the final column correspond to the two sample unpaired t-test between the development and test sets. For binary variables the test of proportions is used instead.*
†Chronic health conditions as defined in APACHE IV [35].

Variable	Entire cohort	Development	Test
PaO_2/FiO_2	41012 (56.59%)	27655 (56.61%)	13357 (56.55%)
PaO2	40968 (56.53%)	27622 (56.54%)	13346 (56.51%)
PaCO2	40968 (56.53%)	27622 (56.54%)	13346 (56.51%)
pH	40968 (56.53%)	27622 (56.54%)	13346 (56.51%)
White Blood Cell count	17616 (24.31%)	11710 (23.97%)	5906 (25.01%)
Haematocrit	15221 (21.00%)	9938 (20.34%)	5283 (22.37%)
Blood Urea Nitrogen	15067 (20.79%)	9837 (20.13%)	5230 (22.14%)
Creatinine	14715 (20.30%)	9542 (19.53%)	5173 (21.90%)
Glucose	9454 (13.04%)	6552 (13.41%)	2902 (12.29%)

Table 4.6: *List of variables which were missing data in the development and test datasets. Variables which are not listed did not have any values missing.*

Mean value imputation was used for remaining features which had missing data. The percentage of missing values for the development and test sets is provided in Table 4.6. Features which contained missing data were primarily blood gases and laboratory measurements.

4.5.2 PSO convergence

The convergence of the PSO (as measured by the AUROC) is shown in Figure 4.4. The figure shows the median AUROC across all particles, plotted against the iteration in the swarm optimisation, for the final set of features selected by the GA. The median particle performance converges to the optimal particle performance by the 300^{th} iteration.

4.5.3 GA feature selection

The features selected across 100 repetitions of the GA trained using the PSO as a fitness function are shown in Figure 4.5. GCS is the most frequently selected variable, appearing in the best selected feature set across all 100 repetitions of the GA. Haematocrit and a binary indicator flag for gender were never selected in any of the best genomes. A list of the features selected with the corresponding frequency of selection in the final GA generation is shown in Table 4.7.

The convergence of the AUROC across generations of the GA is shown in Figure 4.6. The figure shows the best genome in the GA reaches an optimal value (though not necessarily globally so) by at least the 10^{th} generation.

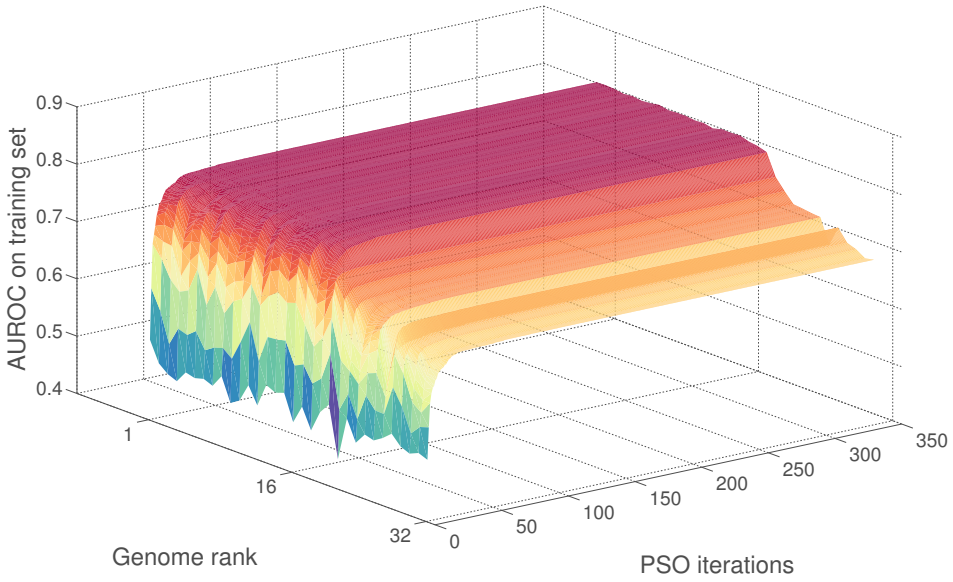

Figure 4.4: *AUROC of the median particle in the population across 32 genomes in the final generation of the GA as calculated on the training set.*

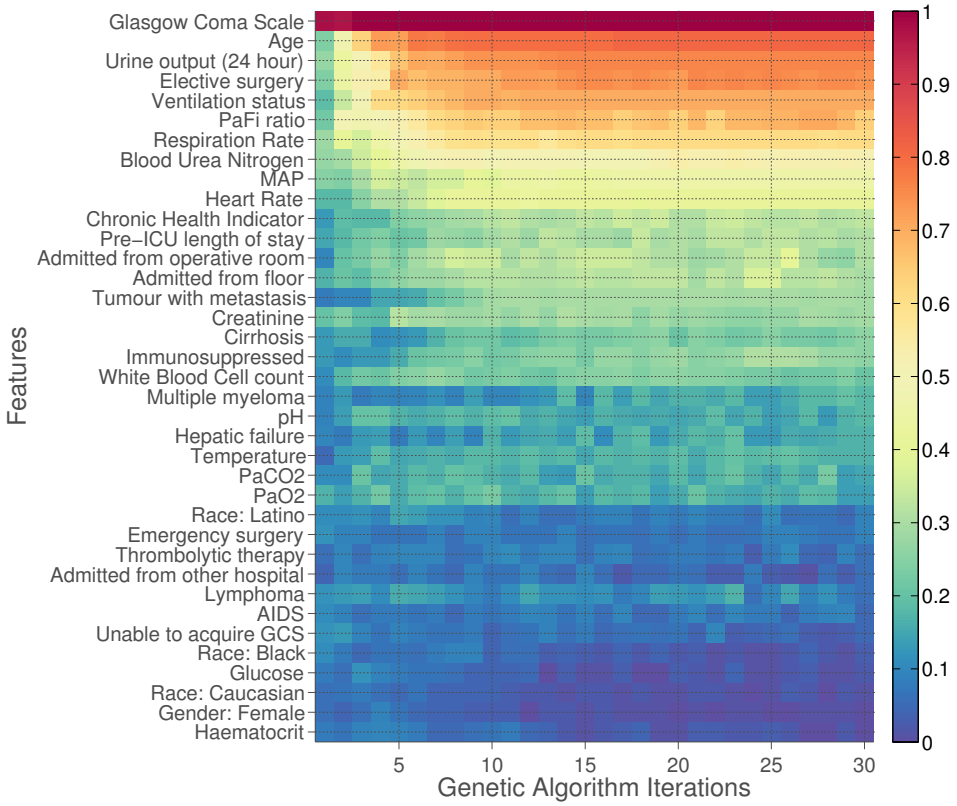

Figure 4.5: *Genetic algorithm feature selection using the AUROC of a PSO severity score as a fitness function for the 29 specified features. Darker values indicate that the feature was selected more frequently.*

Name	Inclusion Frequency	Name	Inclusion Frequency
Glasgow Coma Scale	**100.0**	White Blood Cell count	21.0
Age	**81.0**	Multiple myeloma	18.0
Urine output, 24 hour total	**76.0**	pH	18.0
Elective surgery	**74.0**	Hepatic failure	17.0
Ventilation status	**70.0**	**Temperature**	**17.0**
PaFi ratio	64.0	PaCO2	15.0
Respiration Rate	**61.0**	PaO2	14.0
Blood Urea Nitrogen	52.0	Race: Latino	9.0
MAP	**46.0**	Emergency surgery	9.0
Heart Rate	**42.0**	Thrombolytic therapy	6.0
Any comorbidity (CHI)	33.0	Other hospital admission	6.0
Pre-ICU length of stay	**31.0**	Lymphoma	6.0
Operative room admission	31.0	AIDS	5.0
		Unable to acquire GCS	5.0
Floor admission	30.0	Race: Black	3.0
Metastatic tumour	29.0	Glucose	2.0
Creatinine	29.0	Race: Caucasian	1.0
Cirrhosis	26.0	Gender: Female	0.0
Immunosuppressed	26.0	Haematocrit	0.0

Table 4.7: *A list of features used in the GA and the proportion of optimised feature sets which included them. The optimised feature set was that which had the best performance in the final generation of the GA. The proportion was calculated by averaging across 100 repetitions of the GA. Variables in OASIS are emphasised.*

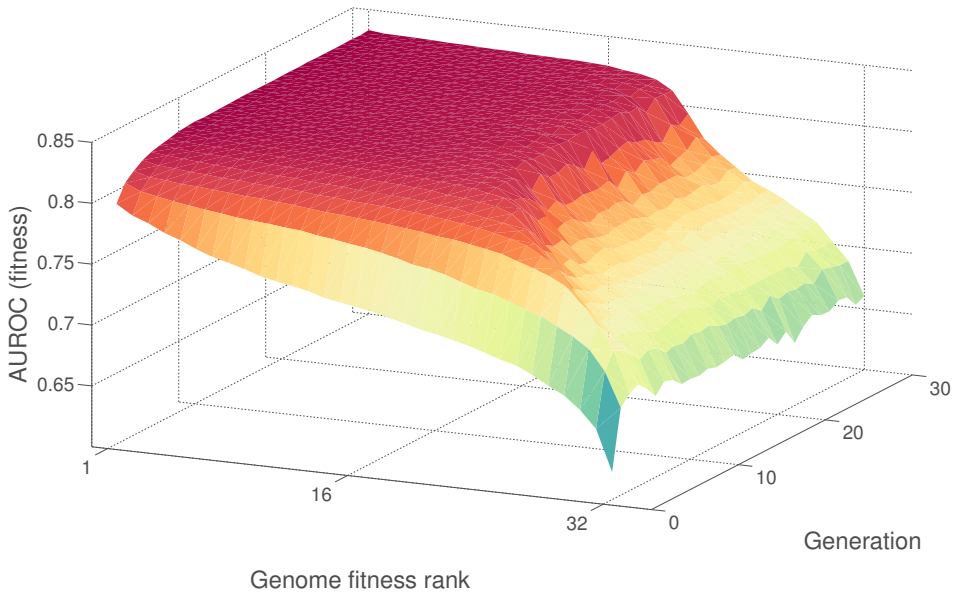

Figure 4.6: *Convergence of the AUROC across generations and iterations of the GA. The median performance across 100 repetitions of the GA is shown.*

4.6 Oxford Acute Severity of Illness Score (OASIS)

The best performing score, as measured by the AUROC, was extracted and designated the Oxford Acute Severity of Illness Score (OASIS) [138]. A score table for OASIS is shown in Figure 4.7.

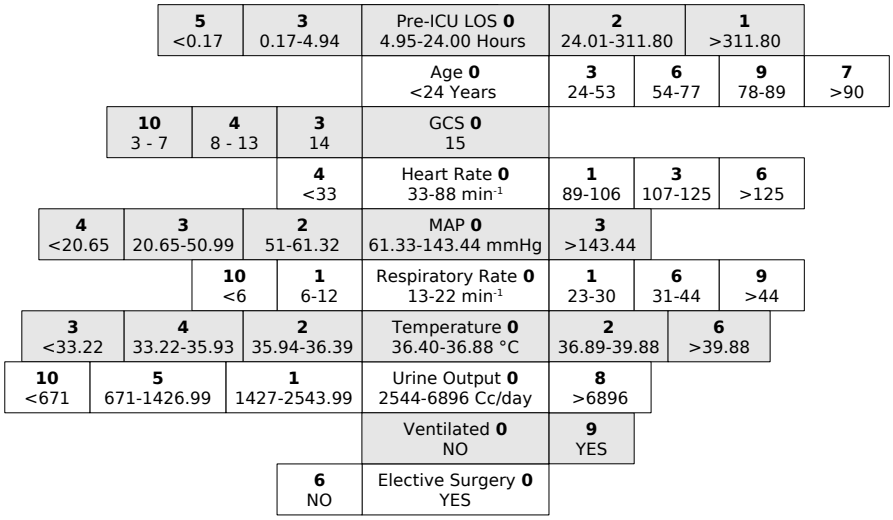

Figure 4.7: *Component weights and bins for the Oxford Acute Severity of Illness Score (OASIS) determined by the PSO methodology using 10 features as selected by the GA. The bolded values are the individual scores assigned to an associated range of measured values. For each variable, the worst score across the first day should be used to tabulate OASIS. The final OASIS score is the sum of all the component weights.*

The distribution of OASIS across the test set is shown in Figure 4.8, and the average value of OASIS in equally sized deciles of risk is shown in Figure 4.9. Both hospital and ICU mortality are seen to increase as the value of OASIS increases.

The performance of the logistic regression models developed for hospital mortality are shown in Table 4.8. In the univariate case with the severity scores as the only independent variables OASIS has a higher AUROC than the APS III. The SMR of the APS III was closer to 1, though neither included 1 in the 95% confidence interval. In the case using multiple predictors, the recalibrated APACHE IV model had a higher AUROC than a similar model with OASIS instead of the APS III. In all models the $HL_{\hat{C}}$ indicated a significant deviation from perfect calibration.

These models were also developed using in ICU mortality as the dependent variable, and the performance of the models using ICU mortality are shown in Table 4.9. Com-

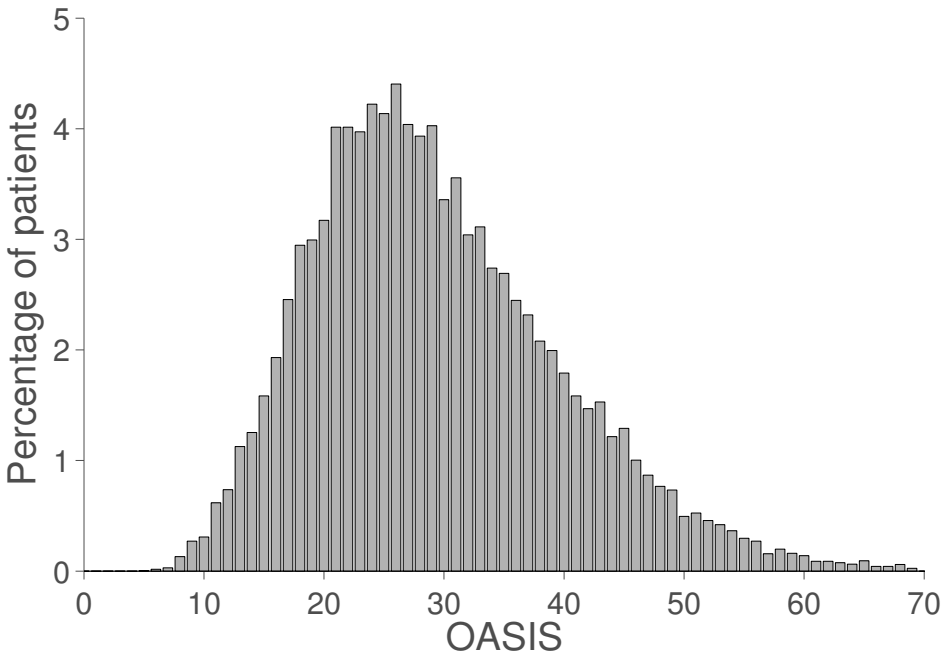

Figure 4.8: *Distribution of patients with a given OASIS value.*

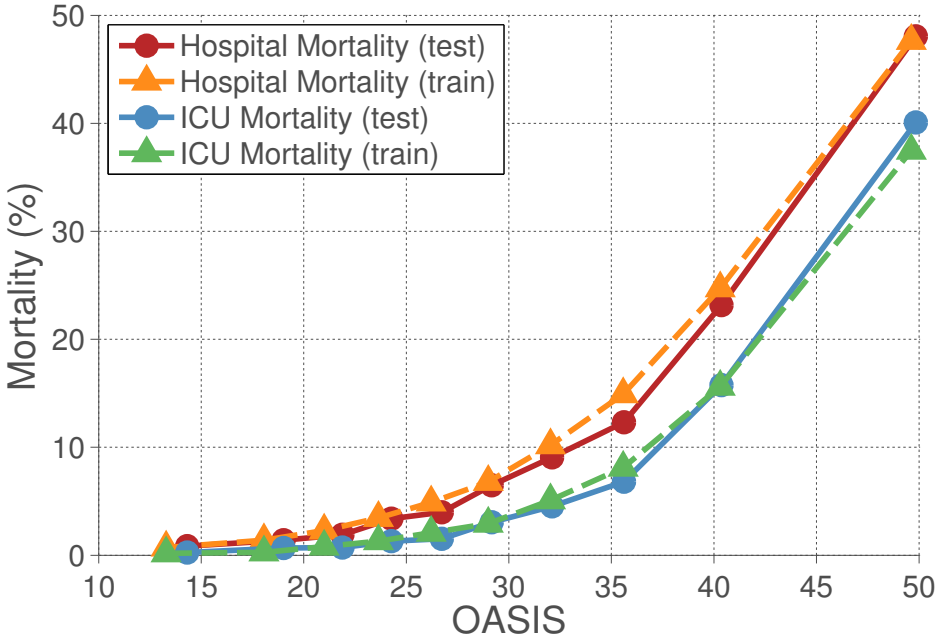

Figure 4.9: *Plot of OASIS against mean observed mortality in ten equally sized deciles of risk. Results on the training set shown as dashed lines with triangles and results on the test set shown as solid lines with circles.*

	Single predictor		Multiple predictors	
	OASIS	APS III	OASIS	APS III ‡
AUROC[1]	0.837	0.822	0.868	0.881
SMR	0.91	0.95	0.93	0.96
(95% CI)	(0.88, 0.94)	(0.92, 0.98)	(0.91, 0.96)	(0.93, 0.99)
B	0.075	0.074	0.069	0.068
$HL_{\hat{C}}$ [2,3]	43.8	62.4	33.7	29.2

Table 4.8: *Comparison of predictive accuracy for the following models of* ***hospital*** *mortality. The multiple predictor models use all covariates present in the APACHE IV model, substituting OASIS for the APS III if specified. The single predictor models use only the specified physiology score as a predictor.*
‡This model is equivalent to a recalibration of APACHE IV.
[1]All standard errors for area under the ROC curve were approximately 0.0009, and all 95% confidence intervals were less than 0.0003.
[2]Based on deciles of predicted mortality.
[3]All p-values were significant at $p < .05$

	Single predictor		Multiple predictors	
	OASIS	APS III	OASIS	APS III ‡
AUROC[1]	0.876	0.864	0.902	0.902
SMR	0.97	1.02	1.01	1.04
(95% CI)	(0.92, 1.01)	(0.98, 1.06)	(0.97, 1.04)	(1.00, 1.08)
B	0.051	0.049	0.048	0.047
$HL_{\hat{C}}$ [2,3]	22.0	71.8	19.6	22.3

Table 4.9: *Comparison of predictive accuracy for various models of* ***ICU*** *mortality on the test set. The multiple predictor models use all covariates present in the APACHE IV model, substituting OASIS for the APS III if specified. The single predictor models use only the specified physiology score as a predictor.*
‡This model is equivalent to a recalibration of APACHE IV.
[1]All standard errors for area under the ROC curve were approximately 0.0009, and all 95% confidence intervals were less than 0.0003.
[2]Based on deciles of predicted mortality.
[3]All p-values were significant at $p < .05$

pared to the APS III, OASIS had a higher AUROC in the single predictor case. When compared to APACHE IV, the use of either OASIS or the APS III as the acuity predictor resulted in the same AUROC = 0.902. All models included 1 in the SMR. The $HL_{\hat{C}}$ indicated a lack of fit for all models at the 0.05 significance level. The B was lowest for the recalibrated APACHE IV and highest for the single predictor model using OASIS.

4.6.1 Calibration of OASIS

The equation to calculate the risk of mortality given a severity score is as follows:

Dependent variable	Severity Score	β_0	β_1
Hospital mortality	OASIS	-6.1746	0.12750
	APS III	-4.4360	0.04726
ICU mortality	OASIS	-7.4225	0.14340
	APS III	-5.3566	0.05116

Table 4.10: *Calibration coefficients of the OASIS and APS III severity scores using the entire development cohort.*

$$\hat{y} = \frac{1}{1 + e^{\beta_0 + \beta_1 s}} \tag{4.9}$$

Here s is the severity score, β_0 is the intercept term of the calibration coefficients and β_1 is the slope term of the calibration coefficients. The actual calibration coefficients for the models with OASIS and the APS III as the only independent variables are provided in Table 4.10.

4.7 Discussion

4.7.1 Concurrent feature development and optimisation

The synthesis of feature selection and score optimisation in the context of severity score development is novel. Widely accepted severity scores, including all generations of the SAPS [8,9,54,55], APACHE [5,7,35,53] and MPM [10,11,41] models performed univariate feature selection prior to multivariate model development. Such an approach may lead to higher dimensional models than is necessary due to the strong correlations among the features commonly collected in the ICU. In this work, the optimisation of the score in tandem with the feature selection allowed for the combination of features which had independent (or complimentary) predictive information. However, a limitation of this approach is the lack of a constraint on the complexity of the model. Spurious features which are not helpful nor harmful were included in many of the best genomes since there was not a consistently applied penalty for their inclusion. One approach to mend this flaw is the modification of the fitness function to penalise complex models. This has been successfully applied in the context of sepsis prediction to obtain a parsimonious

regression model through use of automatic relevance determination [139, 140]. Unfortunately this method depends on a probabilistic prediction, whereas the score in this work outputs a continuous value. Another approach would be the addition of a hyperparameter and a dimensionality term to the fitness function, so that $f_{new}(x) = f(x) + \psi D$, where ψ is the hyperparameter value and D is the dimensionality. Selection of this hyperparameter would be non-trivial and this is an avenue for future work.

Diagnosis was intentionally excluded from the feature selection search as it requires classifying the entire patient set using a specific non-generic ontology. Since coding an entire patient database using a new ontology is a laborious task, this would heavily deter external use of the score. Furthermore, since the primary diagnosis can often be ambiguous, this may cause further complication to a model whose aim is simplicity. Notably, this argument could be similarly made for the comorbidities which were features selectable by the GA. However, the coding of patients into up to seven comorbidities requires substantially less effort than the coding of patients into one of 115 diagnoses. As such it was decided to include these comorbidities in the feature selection search.

4.7.2 OASIS

When evaluated independently with in-hospital mortality as the outcome of interest, OASIS had an AUROC = 0.837 which was higher than the widely clinically utilised severity score APS III (AUROC = 0.822). When used as a replacement covariate for the APS III in the APACHE IV system, OASIS had lower discrimination (0.868 versus 0.881).

A comparison of the features in OASIS with the most frequently selected features by the GA (Figure 4.5) shows that while PaO_2/FiO_2 was very frequently selected, it was not a covariate in the OASIS model. The severity score selected from the GA and PSO optimisation process was stochastic, thus while it is more likely to contain features occurring more frequently in the final genome in Figure 4.5, these features were not guaranteed to be present. A second important aspect of the visualisation in Figure 4.5 is its invariance to shared information between features. For example, if two features contained the exact same information, one could conceive the GA alternating between

including one of the two features (or both) across repetitions while always containing at least one feature. At worst, this could lead to two highly predictive (and correlated) features being selected in only 50% of the final genomes, and as a result these features would not appear as "important" as others in 4.5 and Table 4.7. Either of these effects could contribute to the exclusion of certain features (such as PaO_2/FiO_2) and inclusion of others (such as temperature) in OASIS. Note that a GA is not the only feature selection technique available, and there may be other approaches (e.g. greedy selection) which could have performed as well in the task of feature selection given the relatively small number of features (37 in total).

Regardless of the exact features, it is clear that OASIS performs as well as could be expected, as shown in Figure 4.6. The best performing gene in the optimisation process achieved an AUROC of approximately 0.83, while OASIS achieved an AUROC of 0.837 on the test set.

The variables included in OASIS are routinely monitored and rarely contain missing data. Heart rate, blood pressure, temperature, respiratory rate, urine output and GCS are variables which are charted for almost every ICU patient admitted. In the original exclusions urine output was missing for 6,903 patients (8.51%) and GCS was missing for 1,700 patients (2.10%). The remaining variables were present for all but 10 admissions. Furthermore, information such as age, pre-ICU length of stay and elective surgery are feasibly simple to extract from an ICU database. The only potentially difficult variable to measure retrospectively is ventilation status, as ventilators are commonly separated from ICU data management systems and their use is sometimes approximated with surrogate variables. Nevertheless, it is a straightforward variable to capture in the context of a prospective trial which requires risk adjustment.

An obvious and interesting question is why OASIS performs as well or better than previous ICU scoring systems (namely the APS III) when it includes fewer covariates. For ICU mortality as the dependent outcome, OASIS had higher discrimination as compared to the APS III and equivalent discrimination when both were used as components of a larger model. As OASIS was developed using ICU mortality, it is expected that it would perform better than the APS III which was developed using hospital mortality.

For hospital mortality as the dependent outcome, OASIS discriminated better independently than the APS III but had lower discrimination when used in a larger model (i.e. APACHE IV). This again may be partly attributable to the use of ICU mortality as the dependent outcome in the development of OASIS.

Another reason for OASIS' high discrimination given its reduced feature set compared to other severity scores is the method of feature selection. Severity scores such as SAPS, APS and MPM used step-wise feature selection techniques, usually setting a threshold of statistical significance and rejecting variables which tested above that threshold. Conversely, the feature selection approach presented here is done simultaneously for all features in the model. This allows for the case where, for example, both ventilation status and PaO_2/FiO_2 ratio are significant, but the information contained in both overlaps, and there is no independently useful information in PaO_2/FiO_2. It is worth noting that in the recalibrated APACHE IV model provided higher discrimination the equivalent model with OASIS instead of the APS III. This may indicate that OASIS is capturing information not present in the APS III but which is present in the APACHE IV, and after addition of this information some component of the APS III (either the higher dimensionality or the interaction terms present) allows for better performance. This hypothesis is partly supported by the inclusion of age in OASIS but not in the APS III.

This chapter detailed the development of a parsimonious severity score on a large multi-center database collected in the United States. The evaluation of this score was performed on held out data collected after the development data, providing a good estimate of generalisation performance. To further validate the score, it is necessary to evaluate it at institutions distinct from the development dataset. The next chapter provides a comparison of OASIS with other severity scores currently used in clinical practice on a fully external dataset extracted from an American hospital.

Chapter 5

Evaluation of OASIS in the US

We are trying to prove ourselves wrong as quickly as possible, because only in that way can we find progress.

Richard Feynman

5.1 The MIMIC II database

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II database [107] is a publicly available ICU database sourced from the Beth Israel Deaconess medical center in Boston, Massachusetts. The data collected pertains to all aspects of a given patient's ICU stay, including medication, laboratory tests, bedside monitoring, chart recordings, clinical notes, discharge summaries and patient outcomes. In compliance with the United States Health Insurance Portability and Accountability Act of 1996 (HIPAA), all Private Health Information (PHI) was removed from the dataset before release using bespoke open source software [141]. The relevant section, 45 CFR 164.514b(2), provides a list of PHI which must be removed from the data prior to its public release. Only the guardians of the MIMIC II database possess the ability to map each individual back to the original patient identifiable information.

The MIMIC II clinical database has undergone many iterative updates, and the version used in this work is MIMIC II v2.6. The data is stored in a relational database whose schema is available from PhysioNet, the primary host of the MIMIC II database

[107].

As MIMIC II provides access to a large, anonymised and publicly available ICU dataset it provides an excellent opportunity to benchmark various severity scores on a common dataset. Unfortunately, while there is an enormous amount of information in MIMIC II, not all covariates for all severity scores or predictive models are available. For example, as SAPS III requires information regarding treatment prior to the ICU admission, it is not possible to tabulate the score. MPM_0-III has similar requirements for data collected outside of the ICU. Similarly, while the covariates for the APACHE III and APACHE IV models are available, both require patients to be classified into diagnostic categories. MIMIC II does contain a classification of patients into diagnosis reason groups (DRGs). However this mapping is used primarily for billing purposes and does not have an obvious mapping to the APACHE diagnostic groups. Additionally, the APACHE diagnoses are determined on patient admission, whereas DRGs are primarily retrospectively determined. Nevertheless APACHE IV has been applied to MIMIC II in the past [139]. This was achieved by selecting a subpopulation (with existence of septicaemia) and required manual review of discharge summaries by an intensivist. As there are over 20,000 first day adult patient ICU admissions in the MIMIC II database, this would be a prohibitively time consuming task for all admissions.

Severity scores that were already available in MIMIC II include SOFA and SAPS. The formulae for these scores have already been provided by the maintainers of the MIMIC II database. Furthermore, it was possible to calculate the APS III, SAPS II and OASIS given the data available in MIMIC II. This chapter details the extraction of the data and compares the performance of these severity scores. This acts as an external evaluation of the various severity scores as the data in MIMIC II is sourced from a hospital external to their respective development datasets.

5.2 Severity score comparison

First, the data pertaining to the variables used in the severity scores were extracted. The severity scores were compared after application of exclusion critera detailed in Section 5.2.2. For SAPS and SOFA, only the discrimination could be compared as the original

articles did not provide coefficients to map the score to a probability of mortality [8,124]. Both discrimination and calibration were compared for the remaining severity scores.

Table 5.1 provides a full list of all the variables used in the severity scores compared in this chapter. The table lists the parameters required as inputs to the OASIS, APS III, SAPS, SAPS II and SOFA.

5.2.1 Extraction of data

A subset of tables were the primary sources of data for the analysis of severity scores in the MIMIC II database: the *icustay_detail* table, the *chartevents* table, the *labevents* table and the *ioevents* table. The latter three tables are each associated with a definitions table: *d_chartitems*, *d_labitems* and *d_ioitems* respectively. The *icustay_detail* table contains demographic information regarding each patient including outcomes, age, gender and a set of unique identifiers. Each patient has three unique identifiers (IDs) associated with their stay: an Intensive care unit Identifier (IID), a Hospital Identifier (HID) and a Subject Identifier (SID). The IID is a unique identifier for a single ICU admission, and this is the primary ID used to extract patient data. Each patient's distinct hospital stays are given a unique HID. Thus if a patient was discharged from the ICU, remained in the hospital, and was then later readmitted to the ICU they would have the same HID for both ICU stays but two unique IIDs. If a patient is discharged from the hospital and later readmitted to the hospital and to the ICU they would receive both a new HID and a new IID. The SID is unique to each patient. Data were extracted by grouping observations based upon the IID.

The severity scores tabulated were the APS III, SAPS II, OASIS, SAPS and SOFA. Both SAPS and SOFA are available in the database as they were previously calculated by the coordinators of the MIMIC II database [107]. The variables for the APS III, SAPS II and OASIS were manually searched for and extracted. The general procedure for acquiring data was to first search for text labels similar to the desired variable in one of the definitions tables (prefixed by *d_*). For example, if the desired variable is heart rate, text searches in the *d_chartitems* table would include "heart rate", "heart" and "hr". The result would be a set of unique identifiers, called Item Identifiers (*itemid*s),

Variables	OASIS	APS III	SAPS	SAPS II	SOFA
A-aDO_2		✓			
Albumin		✓			
Age	✓		✓	✓	
Bilirubin		✓		✓	✓
Bicarbonate			✓	✓	
Blood pressure - mean	✓	✓			✓
Blood pressure - systolic			✓	✓	
Blood urea nitrogen		✓	✓	✓	
Chronic dialysis		✓			
Creatinine		✓			✓
Elective surgery	✓			✓	
Emergency surgery				✓	
FiO_2		✓			✓
Glasgow coma scale	✓	✓	✓	✓	✓
Glucose		✓	✓		
Haematocrit		✓	✓		
Heart rate	✓	✓	✓	✓	✓
PaO_2		✓		✓	✓
$PaCO_2$		✓		✓	✓
pH		✓			
Respiratory rate	✓	✓	✓		✓
Platelets					✓
Potassium			✓	✓	
Pre-ICU length of stay	✓				
Pressors†					✓
Sodium		✓	✓	✓	
Temperature	✓	✓	✓	✓	
Urine output (daily)	✓	✓	✓	✓	
Ventilation	✓	✓	✓‡	✓‡	
White blood cell count		✓	✓	✓	
Comorbidities					
AIDS				✓	
Metastatic cancer				✓	
Leukemia, Lymphoma or Immunosuppression				✓	

Table 5.1: *Variables required for each severity score. '✓' indicates the variable is used by the model.*
†Treatment with pressors including dobutamine, dopamine, adrenaline and noradrenaline.
‡Also includes continuous positive airway pressure.

ITEMID	LABEL	CATEGORY	DESCRIPTION
184	Eye Opening	(null)	(null)
198	GCS Total	(null)	(null)
454	Motor Response	(null)	(null)
723	Verbal Response	(null)	(null)

ICUSTAY_ID	ITEMID	CHARTTIME	VALUE1	VALUE1NUM
4	184	08-SEP-82 08.00.00.000000000 AM US/EASTERN	1 No Response	(null)
4	723	08-SEP-82 08.00.00.000000000 AM US/EASTERN	1.0 ET/Trach	(null)
4	454	08-SEP-82 08.00.00.000000000 AM US/EASTERN	1 No Response	(null)
4	198	08-SEP-82 08.00.00.000000000 AM US/EASTERN	3	3
4	184	08-SEP-82 12.00.00.000000000 PM US/EASTERN	3 To speech	(null)
4	723	08-SEP-82 12.00.00.000000000 PM US/EASTERN	1.0 ET/Trach	(null)
4	454	08-SEP-82 12.00.00.000000000 PM US/EASTERN	(null)	(null)

Figure 5.1: *An example of the data format in MIMIC II. First, the* itemid *associated with a variable is identified in the definitions table, in this case the total GCS in* d_chartitems. *Data is then extracted from the associated data table, in this case* chartevents, *by searching for rows with the matching* itemid.

which are each associated with a text label. These *itemid*s are used to map values in the *chartevents* table to their meaning. An example is shown in Figure 5.1.

For each of these variables, the highest and lowest value over the first 24 hours were extracted. From the highest and the lowest value, the "worst" value is defined as the one which assigns a higher score for that feature. This is performed independently for each severity score. Thus while the highest heart rate may provide the worst score for the APS III, the lowest heart rate may provide the worst score for the OASIS. APS III would then use the highest heart rate for the final score tabulation, while OASIS would use the lowest heart rate. Certain variables only required either the highest or the lowest: only the lowest GCS is required as its maximum value is normal and only the highest value was required for systolic blood pressure for SAPS II. Other variables which were not extracted in this way include urine output (sum over the last 24 hours), age and pre-ICU length of stay.

The bulk of the variables presented in Table 5.1 were extracted from the *chartevents* table. This process involved finding the minimum and the maximum of the variable across the first day (excluding values of 0 which can occur as an artefact in the data). Urine output was stored in a similar table to *chartevents* labelled *ioevents*. Urine output was extracted from the *ioevents* table by taking the cumulative sum over the first 24 hours across a set of 38 distinct *itemid*s (provided in the appendix Section C.2). A list of the *itemid*s extracted for all variables is also provided in the appendix Section C.2.

Motor	Verbal	Eyes
1 No Response	1 No Response	1 No Response
	1.0 ET/Trach	
2 Abnorm extens	2 Incomp sounds	2 To pain
3 Abnorm flexion	3 Inapprop words	3 To speech
4 Confused	4 Flex-withdraws	4 Spontaneously
5 Localizes Pain	5 Oriented	
6 Obeys Commands		

Table 5.2: *Possible values for the various GCS components in the MIMIC II database.*

A subset of variables included in the severity scores required special extraction or could not be perfectly extracted. This subset includes urine output (already detailed), GCS, ventilation status, pre-ICU length of stay, comorbidities and admission urgency.

5.2.1.1 Glasgow coma scale

GCS represents the level of conciousness of a patient, and as such it is influenced by their level of sedation. For the APACHE IV model, an additional covariate was included to indicate that the patient was sedated when the corresponding GCS value was recorded. In the APS III, OASIS and APACHE IV system values of GCS were set to 15 (normal) if the care provider felt the GCS was not a true reflection of the patient's neurological status. This would occur if the patient was sedated or if a tracheostomy prevented a verbal response. As such, GCS values in the MIMIC II database which are recorded under similar conditions must be replaced by a normal value. Unfortunately, determining sedation status is a difficult task.

The Richmond Agitation Sedation Scale (RASS) [142] is commonly used to assess levels of sedation. However, while the progression from light sedation to heavy sedation correlates well with decreasing RASS values, a low RASS value does not necessarily imply sedation. Thus while scores between -2 to -4 could be caused by sedation, they could equally well be caused by trauma or another neurologically impairing injury. This precludes the use of RASS for determining if a GCS value should be replaced by a normal value due to patient sedation.

The GCS values in MIMIC II can take on up to a maximum of six unique values. These values are shown in Table 5.2. The values of "1.0 ET/Trachy" indicate that the

verbal score for the GCS could not be acquired due to the presence of an endotracheal tube/tracheostomy. As this GCS is not a true reflection of the patient's neurological status, the set of observations for which the verbal score was "1.0 ET/Trachy" were set to normal. Normal in this context is a score of 5 for verbal, a score of 6 for motor and a score of 4 for eyes (totalling 15). This emulates the data extraction which was performed for the APS III and OASIS [35, 143].

5.2.1.2 Ventilation status

Ventilation status was determined using a distinct table available in the MIMIC II development server named *ventilation*. This table contains IIDs with corresponding begin and end times for ventilation. If the begin time in the *ventilation* table occured during the first 24 hours of a patient's stay, that patient was considered ventilated. If the patient had a value of "1.0 ET/Trachy" for any of their GCS verbal scores, then the patient was also considered to be ventilated. While this is not strictly the case, and patients will spontaneously breathe with a tracheostomy, the hypothesis is that the proportion of patients with a tracheostomy on their first day in the ICU who are mechanically ventilated far outweights those who have a tracheostomy but are spontaneously breathing. If the patient does not have a tracheostomy and does not have start and stop times of ventilation in the *ventilation* table, the patient was assumed to not be ventilated.

5.2.1.3 Pre-ICU length of stay

The MIMIC II *icustay_detail* table contains a list of IIDs with corresponding hospital admission dates and ICU admission dates. While the ICU admission dates also include the time during the day at which the patient was admitted to the ICU, the hospital admission dates do not. As such, an accurate calculation of the pre-ICU length of stay could not be determined in the MIMIC II database. A proxy for the pre-ICU length of stay was determined by treating the hospital admission as occuring at midnight of that day. This will bias the value of pre-ICU length of stay higher than its true value, but is unavoidable as the information regarding the time of the hospital admission is unavailable.

5.2.1.4 Comorbidities

Comorbidities were determined using the *comorbidity_scores* table. This table was originally created to facilitate the calculation of the Elixhauser score, an aggregated measure of comorbid status for a patient [144]. There were three comorbidities required for the analysis, each being a component of SAPS II. The *METASTATIC_CANCER* column was used to determine the existence of metastatic cancer. The *AIDS* column was used to determine the existence of AIDS. Finally, the *LYMPHOMA* table was used to determine the existence of a haemotological malignancy. Note that as there are haemotological malignancies other than lymphoma, the comorbid flag used for haemotological malignancy in this work is incomplete and this may affect the predictions of SAPS II.

5.2.1.5 Admission urgency

The last variable which required bespoke extraction was surgery urgency. The *demographic_detail* table contains information pertaining to the patient's admission urgency. This urgency field can take the following values: elective, emergency, urgent or missing. Secondly, the *icustay_detail* table contains information regarding the patient's first service, which corresponds to the ICU which initially cared for the patient. Elective surgery patients were defined as those whose admission urgency was "elective" and whose first service was either the "CSRU" (cardiac surgery recovery unit) or the "SICU" (surgical intensive care unit). Emergency surgery patients were similarly determined as those whose admission urgency was "emergency" or "urgent" and whose first service was either "CSRU" or "SICU".

5.2.2 Exclusion criteria

Several of the severity scores compared had distinct exclusion criteria, which complicates their comparison. The APS III and OASIS development datasets excluded patients who were < 16 years of age, remained in the ICU for < 4 hours, were undergoing a transplant operation (except cases of hepatic or renal transplantation), suffered major burns, undergoing coronary artery bypass graft surgery, in-hospital readmissions, transfers from another ICU and patients in the hospital for over 1 year. The publication for SAPS un-

fortunately does not detail any exclusion criteria [8]. SAPS II excluded burns patients, coronary patients, cardiac surgery patients, patients staying in the ICU longer than 5 months, in-hospital readmissions and patients < 18 years of age. Finally, SOFA was designed by the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine [57]. As the scores were determined using clinical judgement, no development data exists and thus no exclusion criteria exist. It should be noted that the original aim of the SOFA score was to describe the complications afflicting patients with multiple organ failure associated with sepsis. Nevertheless, the authors proposed use of the SOFA score in a general ICU population.

Overall, the reasoning behind the varying exclusion criteria were consistently based upon expert intuition, data availability and empirical observation. Young patients were excluded as pediatric and neonatal physiology differs from that of adult ICU patients. Patients staying in the ICU for less than 4 hours were likely to be undergoing surgical preparation, and it was believed this subpopulation should not be considered part of the general ICU population [35]. In-hospital readmissions and admissions from other ICUs were excluded as the treatment provided in the previous ICU would bias patient physiology away from the true admission physiology of ICU patients [35]. Patients initially admitted to the ICU are expected to be physiologically deranged, and as such patients who were transferred after already receiving ICU treatment would appear more normal than their physiology truly is. Patients receiving a coronary artery bypass graft were excluded from model development as these patients comprise a large group with a very low rate of mortality as compared to other ICU patients. In fact, the APACHE IV system of equations built specific models of mortality for patients undergoing coronary artery bypass surgery [35].

In this work the following exclusion criteria were applied. Patients who were < 18 years of age, in-hospital readmissions and patients who were transferred from another ICU were excluded. Furthermore, patients undergoing coronary artery bypass graft surgery or a transplant surgery (except renal or hepatic) were excluded. These patients were identified using DRG codes and were the primary patient groups excluded in the APS III, OASIS and SAPS II development populations. The use of DRG codes to

estimate admission diagnosis is unideal and elaborated upon in the discussion. Note that there were no burns patients in the MIMIC II database as identified by DRG codes, but these patients would have been excluded had they been present.

5.2.3 Model calibration and evaluation

It is of interest to evaluate the scores both for their discrimination capability and for their calibration. Model discrimination can be assessed for each score by evaluating the AUROC across the entire dataset. As the AUROC is based upon relative rankings of the scores, the metric is usable even if a mapping from an integer scale to a probability does not exist, such as in the case of SOFA and SAPS.

Model calibration was more difficult to assess for two reasons: model drift causes model calibration to fade over time [145], and because two of the scores are not probabilistic estimates of mortality. For models which are convertible into a risk of mortality the B, SMR, $HL_{\hat{C}}$ and $\mathcal{I}_{\mathcal{L}}$ were calculated to evaluate the model calibration. Publicly available equations exist to convert OASIS [138] and SAPS II [9] directly into probabilities. The model to convert the APS III into a probability involves other covariates (including diagnosis) and is referred to as APACHE III. This model was not made publicly available. While the newer APACHE IV also uses the APS III and is publicly available, the large number of diagnostic covariates make it difficult to evaluate on MIMIC II. However, in the process of developing OASIS (see Chapter 4), the APS III was calibrated to a similar dataset to that which was used to develop APACHE IV. The calibration coefficients are listed in Table 4.10 and used to map the APS III to a probabilistic estimate of patient mortality. SAPS and SOFA were not published with calibration coefficients. As such, the calibration of these scores was not directly assessed.

Confidence intervals for the AUROC were calculated using a non-parametric technique which estimated the variance associated with each score and covariance amongst the scores as described by DeLong and DeLong [127]. Statistical differences were quantified using normal theory and the estimates variances and covariances, as detailed in [127]. Bootstrap resampling was used to provide 95% confidence intervals for the calibration statistics [49]. Differences in bootstrap estimates were tested for statistical significance

		Percentile	
	Mean or %	25th	75th
Age	64.76	51.07	77.99
Gender	54.73%		
Mechanical ventilation	45.21%		
Admission type			
Elective	11.12%		
Emergency	31.51%		
Medical	57.38%		
Outcomes			
ICU Length of Stay	4.626	1.276	4.763
Hospital Length of Stay	10.909	4	13
ICU Mortality	8.04%		
Hospital Mortality	12.09%		

Table 5.3: *Demographics of the cohort extracted from MIMIC II after all exclusions.*

using the percentile corresponding to a critical value of 0.05. Confidence intervals were calculated using the bias and acceleration corrected bootstrap percentile method [49]. Probability density functions for each of the scores were calculated using kernel density estimation [146] and plotted for visual inspection (see Appendix B.1 for detail on the kernel density estimation method). Calibration curves and ROC curves were also plotted for the severity scores and calibration coefficients were made publicly available.

5.3 Results

A total of 25,492 first day ICU admissions were extracted from the MIMIC II database. Of these, 1,159 (4.55%) were in hospital readmissions and 560 (2.20%) did not have a value for SAPS or SOFA available. After excluding these two subsets (1,689 patients, 6.63%), a total of 23,803 observations remained. Of these patients, 115 underwent transplants which were not coded as hepatic or renal, and thus were removed (0.48%). Additionally, 2,272 (9.55%) patients who received coronary artery bypass grafts were removed from the dataset. The final cohort analysed had 21,416 patients. Demographics for this cohort are shown in Table 5.3. The top 10 DRGs are shown in Table 5.4.

The AUROC of the SAPS, SOFA, SAPS II, APS III and OASIS are shown in Table 5.5 and the result of multiple statistical comparisons are provided in Table 5.6. When comparing AUROCs for hospital mortality most severity scores were statistically sig-

DRG code	Number of patients (%)	DRG description
105	929 (4.36%)	CARDIAC VALVE & OTH MAJOR CARDIOTHORACIC
416	707 (3.31%)	SEPTICEMIA AGE >17
475	659 (3.09%)	RESPIRATORY SYSTEM DIAGNOSIS WITH VENTIL
14	581 (2.72%)	INTRACRANIAL HEMORRHAGE & STROKE WITH IN
174	575 (2.70%)	G.I. HEMORRHAGE WITH CC
110	541 (2.54%)	MAJOR CARDIOVASCULAR PROCEDURES W CC
1	482 (2.26%)	CRANIOTOMY AGE >17 W CC
483	404 (1.89%)	TRACHEOSTOMY EXCEPT FOR FACE, MOUTH & NE
104	347 (1.63%)	CARDIAC VALVE & OTH MAJOR CARDIOTHORACIC
516	344 (1.61%)	PERCUTANEOUS CARDIOVASC PROC W AMI

Table 5.4: *Top 10 DRGs for the cohort of 21,416 patients after exclusions.*

	AUROC (95% CI)	
	Hospital mortality	ICU mortality
SAPS II	0.802 [0.794, 0.811]	0.828 [0.818, 0.838]
OASIS	0.790 [0.781, 0.799]	0.822 [0.813, 0.832]
APS III	0.785 [0.775, 0.794]	0.812 [0.801, 0.823]
SAPS	0.764 [0.755, 0.774]	0.809 [0.799, 0.819]
SOFA	0.748 [0.738, 0.757]	0.795 [0.785, 0.806]

Table 5.5: *AUROC of the severity scores as evaluated on 21,416 first day admissions in the MIMIC II database after exclusions. The severity scores are sorted in order of descreasing AUROC. 95% confidence intervals are provided as calculated using the non-parametric technique described by DeLong and DeLong [127].*

nificantly different at $p < 0.001$. OASIS and SAPS II were also significantly different ($p = 0.002$) and only the APS III and OASIS were insignificantly different ($p = 0.232$). For ICU mortality, most severity scores were statistically significantly different at the 0.001 level. Exceptions included the APS III and SAPS ($p = 0.581$) and the SAPS II with OASIS ($p = 0.169$). Comparisons which were statistically significant at a level higher than $p = 0.001$ included the SAPS vs SOFA ($p = 0.003$), SAPS vs OASIS ($p = 0.005$) and the APS III vs OASIS ($p = 0.041$). No adjustments for multiple hypothesis testing were applied to these p values. A visualisation of the differences in the AUROCs for the severity scores is shown in Figure 5.2. ROC curves are provided in the appendix Section D.2.1.

The distribution of each score calculated using kernel density estimation is shown in Figure 5.3 and these distributions are overlain on histograms binned at every unit. The

Hospital mortality

	SAPS	SAPS II	APS III	OASIS
SOFA	< 0.001	< 0.001	< 0.001	< 0.001
SAPS		< 0.001	< 0.001	< 0.001
SAPS II			< 0.001	0.002
APS III				0.232

ICU mortality

	SAPS	SAPS II	APS III	OASIS
SOFA	0.003	< 0.001	0.001	< 0.001
SAPS		< 0.001	0.581	0.005
SAPS II			0.001	0.169
APS III				0.041

Table 5.6: *Statistical significance of pair-wise comparisons of the AUROC between severity scores using the non-parametric technique proposed by DeLong and DeLong [127] on 21,416 first day admissions in the MIMIC II database. No adjustment for multiple hypothesis testing has been applied.*

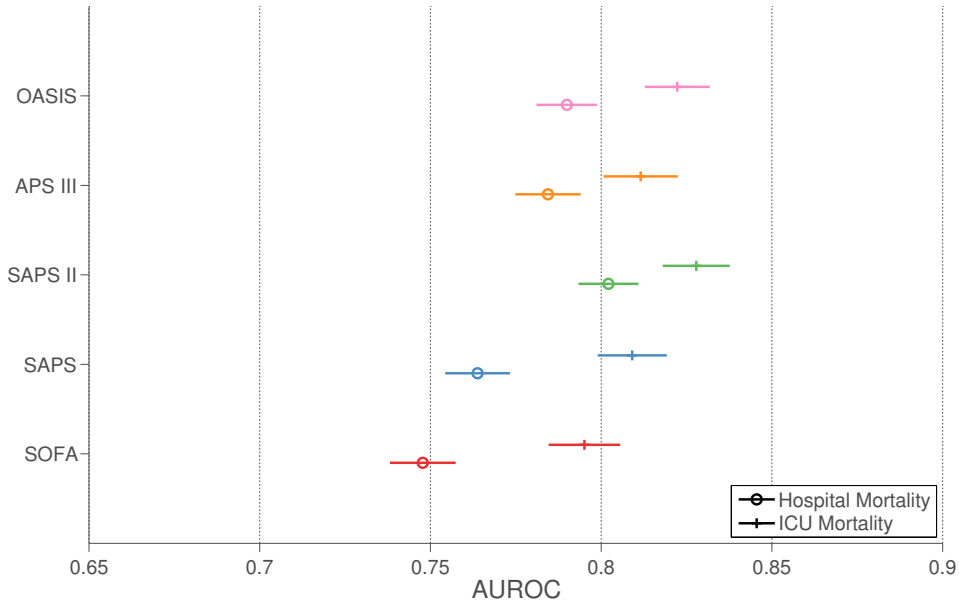

Figure 5.2: *AUROCs for the various severity scores plotted with 95% confidence intervals as calculated using the method of DeLong and DeLong [127]. Circles indicate hospital mortality is the outcome used for the calculation of the AUROC, whereas crosses indicate ICU mortality was used.*

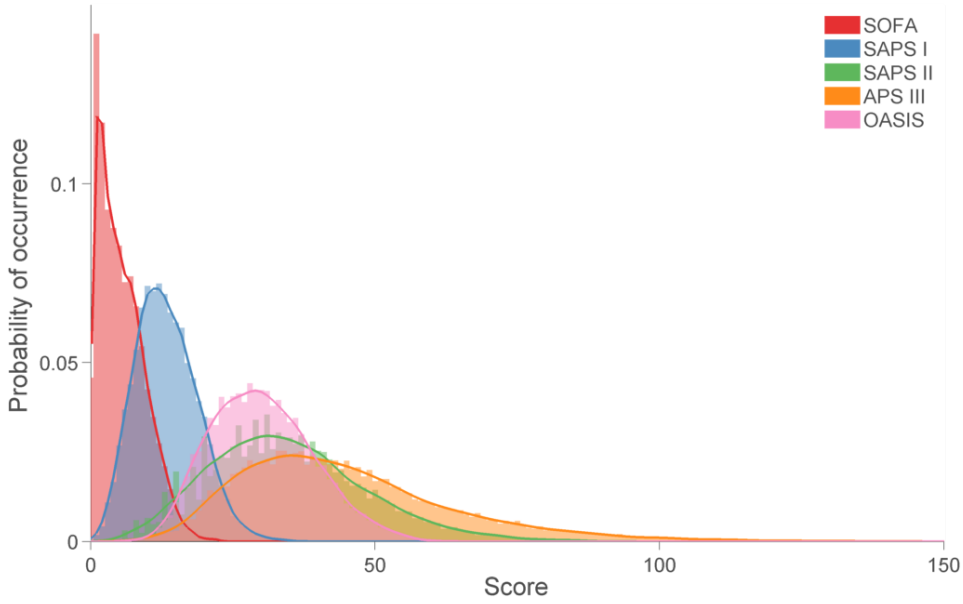

Figure 5.3: *Distribution of severity scores for 21,416 first day admissions in the MIMIC II database. Each line represents the estimated probability destination function using kernel density estimation. Histograms normalised to have unit area are provided below the estimated probability density function. The severity scores from left to right are: SOFA, SAPS, OASIS, SAPS II and APS III.*

distributions of the severity scores appear to smoothly vary from their lower to higher ranges for all severity scores except the SOFA where a large spike appears at a value of one. The distribution of the APS III appears to have a heavy tail, whereas the remaining scores are only slightly skewed towards higher values.

Statistics which primarily evaluate the calibration of the severity scores are shown in Table 5.7. For hospital mortality only the APS III had an SMR whose confidence interval included the ideal SMR = 1, while OASIS nearly had a perfect SMR for ICU mortality and was the only severity score which included the ideal SMR = 1 in its confidence interval. SAPS II had extremely poor calibration and greatly overpredicted hospital mortality. The $HL_{\hat{C}}$ was lower for OASIS in both hospital mortality and ICU mortality comparisons, though it indicating a significant lack of fit for all models ($p > 0.05$, 10 degrees of freedom). Statistics which evaluated both model discrimination and calibration favoured the APS III over OASIS for hospital mortality though these differences were statistically insignificant ($p > 0.05$). Conversely, the $\mathcal{I}_{\mathcal{L}}$ and B_{adj} favoured OASIS over the APS III when predicting ICU mortality though these differences were again statistically insignificant ($p > 0.05$).

Finally, calibration curves are shown in Figure 5.4. Visual inspection confirms that

	Hospital mortality		
	APS III	SAPS II	OASIS
SMR	1.017 [0.984, 1.050]	0.559 [0.540, 0.578]	0.921 [0.890, 0.952]
$HL_{\hat{C}}$	54.2 [38.3, 93.0]	1627.5 [1499.2, 1783.9]	42.8 [25.7, 71.3]
$\mathcal{I}_{\mathcal{L}}$	0.170 [0.158, 0.181]	0.080 [0.061, 0.098]	0.168 [0.156, 0.180]
B	0.089 [0.086, 0.092]	0.103 [0.101, 0.106]	0.091 [0.088, 0.094]
B_{adj}	0.160 [0.147, 0.172]	0.029 [0.005, 0.054]	0.144 [0.130, 0.157]

	ICU mortality		
	APS III	SAPS II	OASIS
SMR	1.157 [1.108, 1.206]	-	1.002 [0.959, 1.045]
$HL_{\hat{C}}$	89.3 [62.8, 142.9]	-	24.4 [14.5, 52.0]
$\mathcal{I}_{\mathcal{L}}$	0.190 [0.175, 0.204]	-	0.196 [0.181, 0.210]
B	0.063 [0.060, 0.065]	-	0.064 [0.061, 0.066]
B_{adj}	0.152 [0.137, 0.166]	-	0.137 [0.122, 0.151]

Table 5.7: *Performance of the severity scores as evaluated on 21,416 first day admissions in the MIMIC II database after exclusions. As all of these metrics require probabilistic estimates, they are unavailable for the SOFA and SAPS. 95% confidence intervals calculated using the bootstrap percentile method [49] and 5,000 bootstrap samples are provided.*

SAPS II consistently overpredicts mortality across all ranges of risk. OASIS appears to gradually overpredict mortality as the decile of risk increases, while the APS III appears to gradually underpredict mortality as the decile of risk increases.

5.4 Discussion

For hospital mortality, SOFA was the worst model as measured by the AUROC, while SAPS II was the best. For ICU mortality, this trend held. The probability density function of SOFA appears to have a large spike at one of the lower values. The number of patients with a total SOFA of zero through four are: 985, 3050, 2518, 1991 and 1879 respectively. SOFA is the only severity score to display such a large spike and also has the smallest operating range of all the severity scores. The large spike is likely an artefact of quantisation noise due to the use of integer scores which is not exhibited in other severity scores due to their larger operating range. The use of integer scores is itself an artefact of the need for care providers to have the ability to tabulate a severity score manually.

The AUROCs of each model compared to those presented in the original article are shown in Table 5.8. It is interesting to note that all severity scores experienced a degra-

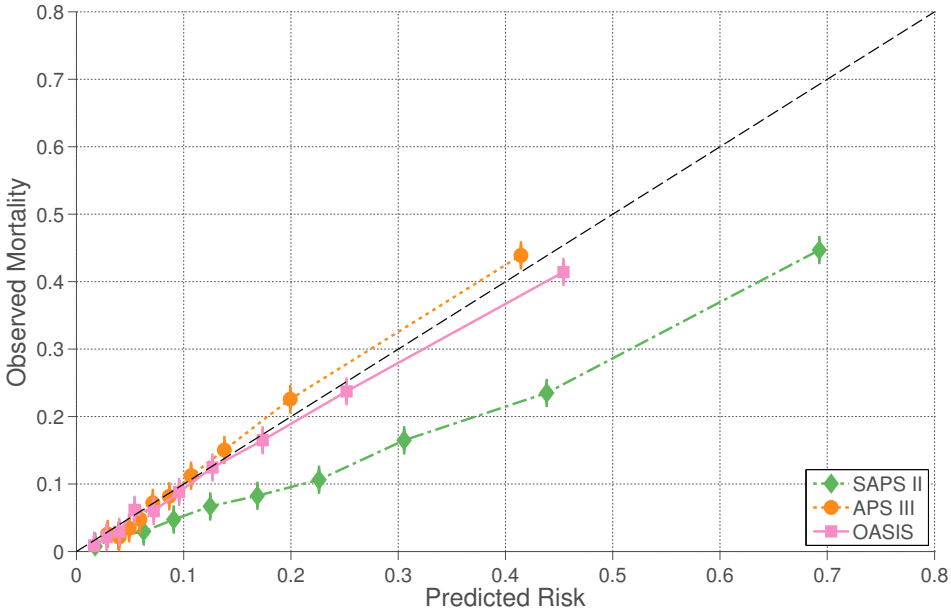

Figure 5.4: *Calibration curves for the three severity scores for which calibration coefficients were available. The dashed line represents the ideal performance of a model where the expected mortality is always equal to the observed mortality. Scores above this line underpredict mortality, while scores below this line overpredict mortality. The calibration curves are calculated across the entire MIMIC II cohort of 21,416 patients.*

	AUROC in publishing article	AUROC in MIMIC	Change in performance
SOFA	-	0.747	-
SAPS	0.770	0.764	-0.006
SAPS II	0.860	0.802	-0.058
APS III	0.822[†]	0.785	-0.037
OASIS	0.837	0.790	-0.047

Table 5.8: *Performance of each severity score for predicting hospital mortality as measured by the AUROC in MIMIC and the original published article.*
[†]*Performance as evaluated on the AO external validation data, not the original article's validation dataset, as this was the data used for calibrating the score.*

dation in performance when evaluated on the MIMIC data. This may be explained by the format of data collection in MIMIC II versus in the publishing articles for the severity scores. The collection of data for the development of the severity scores involved humans manually transcribing data. Special procedures were put in place to train personnel on the proper collection of data for the study. Conversely, the MIMIC II data is collected automatically, and though the values have been electronically charted by a nurse, this nurse was neither cognizant of the data collection rules for the various severity scores nor were they collecting data specially for the study of severity scores. This is highlighted by the complicated process of extracting GCS, which was coded differently in the APS III, OASIS and SAPS II studies (sedated patients being given a GCS of 15) as compared to the MIMIC II data (sedated patients given a GCS of 3). Nevertheless, the fact that the severity scores can be acquired on MIMIC II and still have good performance is extremely promising. The process for collecting the variables for the severity score, while initially laborious, can be run automatically in the future. This vastly reduces the overhead required to acquire severity scores for a cohort of patients in the MIMIC II database, and reduces the difficulty of performing studies on this database. The necessary code for the extraction and calculation of these severity scores will be made openly available to researchers to facilitate this process.

OASIS and APS III were very well calibrated. As the calibration coefficients for both OASIS and APS III were acquired from a database collected between 2007-2009 [138], the models were much less likely to be affected by model drift on the MIMIC II data (collected between 2001-2008). Both OASIS and APS III can be applied as is to the MIMIC dataset using the calibration coefficients provided in Chapter 4. Furthermore, OASIS and APS III may be suitably calibrated to other hospitals and ICUs in the United States, though the scores should always be empirically evaluated before use.

In terms of calibration, it is evident from Figure 5.4 that SAPS II vastly underpredicts mortality. SAPS II overpredicting mortality is perfectly in line with expectation, as this phenomenon of model drift has been observed when recalibrating older models [42]. As care practices improve, mortality rates for equally severe patients are reduced, and severity of illness scores begin to overpredict mortality. Due to its poor calibration,

SAPS II should not be applied for either risk-adjustment or benchmarking without recalibration. While the evidence here only supports the argument that recalibration of SAPS II is necessary on the MIMIC II database, it is not unreasonable to extrapolate this finding to other databases sourced from the US. Indeed there have been many publications which aimed to recalibrate SAPS II to a local population [60, 87]. A fairer comparison for SAPS II would involve a recalibration of the score to the AO dataset (as was done for APS III in Chapter 3). Unfortunately, the necessary covariates for the calculation of SAPS II are not available in the AO dataset prohibiting this exercise.

In terms of model discrimination, SAPS II has the highest performance for both in-hospital (AUROC = 0.802 versus AUROC = 0.790 for OASIS) and ICU mortality (AUROC = 0.828 versus AUROC = 0.822 for OASIS). Variables uniquely used by SAPS II include emergency surgery and existence of comorbidities. The use of additional variables is a plausible reason for the improved discrimination of SAPS II compared to other severity scores.

When comparing OASIS and the APS III, there is little difference between the model fit. In terms of the $\mathcal{I}_{\mathcal{L}}$, the models are almost equivalent ($\mathcal{I}_{\mathcal{L}} = 0.170$ for APS III versus $\mathcal{I}_{\mathcal{L}} = 0.168$ for OASIS when predicting hospital mortality). The major difference stems from the complexity of the severity scores: OASIS requires 10 variables whereas the APS III requires 17. The 17 variables required by the APS III include chemistry and haemotology measurements using blood samples whereas these measurements are not utilised by OASIS, increasing the ease of OASIS' implementation. APS III also requires incorporation of interactions based upon chronic dialysis, existence of acute renal failure and fraction of inspired oxygen, and these interactions increase the likelihood of errors in the calculation of the score. In this comparison, and indeed overall, OASIS appears to be a simpler alternative to other severity scores which has competitive performance.

This study does have limitations. First, it compares the calibration of models developed between 2007-2009 (APS III and OASIS) with the calibration of models developed in 1991-1992 (SAPS II) or earlier. As such this is an intrinsically biased comparison in favour of APS III and OASIS as they were more recently developed and have had a much more recent recalibration. While many recalibrations have been performed for

SAPS II [87, 95], the goal of this work was to evaluate the models as they are most likely to be applied: using the calibration coefficients in the publishing article. It is also worth noting that while special care was taken to ensure the GCS values for the SAPS II, OASIS and APS III severity scores were correct these same steps were not applied for the SOFA and SAPS models. This was done for the same reasoning as the utilisation of the original SAPS II coefficients: it allows for a better comparison of the performance reported here to other studies which have previously compared with these severity scores on the MIMIC II database [147, 148]. Nevertheless, it would be worthwhile to evaluate these scores after correction of the GCS values.

A second limitation of the analysis is the format of the data collection. MIMIC II provides a vast resource of ICU data, but this data can be difficult to extract due to the nature of the data warehousing. The database was not rigorously structured, which forces the researcher to acquire the same information from multiple disparate locations. This is a difficult task and even if care is taken, some information may be lost. In particular, ventilation status is difficult to determine as there were no mechanisms for patient intubation or extubation to be recorded in the database. As such, all estimates of mechanical ventilation are surrogates and may not accurately reflect the patient's ventilation status. This issue is slightly alleviated by the large 24 hour window for severity scoring, during which it is likely that a measurement implicating mechanical ventilation has been recorded. Nevertheless this remains a source of error which cannot be fully mitigated.

The use of DRGs to determine diagnosis is also not without flaws. DRGs are primarily used for determining the reimbursement to a hospital for care given. As such, they are not necessarily a reflection of the primary diagnosis for a given patient. As patients in the ICU are severely ill, they often have multiple diagnoses, and the selection of a single one of these for classification using DRGs can be influenced by the amount of reimbursement for the DRGs. As CABG is a common surgery with a relatively low frequency of mortality, it is likely that most patients who underwent this procedure were properly classified. Still, for the patients who developed complications, it is not unreasonable to imagine a different DRG code was assigned to these patients. Though

this DRG code would not reflect the reason for admission to the ICU, it may reflect the later complications which afflicted the patient, and furthermore may provide a larger reimbursement for the institution. This effect may bias the CABG population towards those who did not suffer further complications, and patients who underwent CABG but later deteriorated were not necessarily excluded from the cohort. Furthermore, four of the top ten DRG codes would be excluded from the analysis if the full SAPS II exclusion criteria were applied. As such, the case-mix of the population differs from that used to develop the SAPS II, and this may have unfavourably affected its predictions.

Finally, as the Beth Israel Institute is located in the United States, it could be argued that severity scores developed in the United States exclusively (APS III and OASIS) are at an advantage compared to those developed elsewhere (SAPS, SAPS II, SOFA). This issue will be addressed in Chapter 6, which benchmarks these models in an institution in the United Kingdom.

Overall, five severity scores were compared on a dataset collected externally to their respective development sets. As the dataset was collected automatically, it has undergone less quality assurance as compared to datasets used for the development of a given severity score. All severity scores performed worse than their respective development sets. Both OASIS and the APS III were shown to calibrate well on this external dataset. SAPS II had the highest discrimination for hospital mortality (AUROC = 0.802) but calibrated poorly ($HL_{\hat{C}} = 3264.9$). OASIS had similar discrimination (AUROC = 0.790) and calibration was far superior to that of SAPS II ($HL_{\hat{C}} = 24.4$). The APS III was similarly shown to have excellent calibration on the MIMIC II database and similar discrimination to OASIS and SAPS II. SAPS and SOFA both underperformed compared to the other severity scores, and have been supplanted by the more recently developed severity scores.

OASIS requires a reduced feature set compared to other severity scores, does not have any interaction terms and does not require potentially ambiguous variables (such as comorbidities). Though the results here provide a narrower comparison compared to the previous chapter, as the dataset pertains to a single institution, it nevertheless demonstrates that OASIS has good discrimination and calibration in an external insti-

tution. Furthermore it demonstrates that OASIS is usable in the setting where data is collected automatically, rather than by specially trained personnel.

Chapter 6

Evaluation of OASIS in the UK

Insanity: doing the same thing over and over again and expecting different results.

Albert Einstein, (attributed)

OASIS was developed on a large multi-center population of ICU patients admitted to hospitals in the United States. While it was demonstrated to calibrate well to an external hospital in the United States in Chapter 5, this does not guarantee the severity score would also generalise to ICUs external to the United States. The APACHE III model has been intensely scrutinised for being only well calibrated on a highly specific subset of the American population who are admitted to hospitals with the APACHE medical system installed [149]. In order to ascertain whether OASIS is well calibrated in critical care environments outside of the United States, a validation study was undertaken at a large tertiary teaching hospital in Oxford, England.

6.1 John Radcliffe database

Data from 3,577 patients admitted to the John Radcliffe hospital, a member of the Oxford University Hospitals NHS Trust in the United Kingdom, were extracted from a clinical data warehouse. This data warehouse was created as a part of the Post-

Intensive Care Risk-adjusted and Monitoring (PICRAM) trial[1]. The data encompassed patients admitted between 2007-2011. Though the study itself focused on post-intensive care, a key component of the study was risk-adjusting patients upon discharge from the ICU. The development of this institution specific risk-adjustment model required the extraction of an ICU database for patients previously admitted to critical care at the John Radcliffe. A secondary aim of the study was the creation of an anonymised dataset available for qualified researchers who submit a research proposal. A research proposal for the study of severity score performance was accepted and the study coordinators provided access to the data. This dataset is referred to as the John Radcliffe Database (JR_{DB}).

As the care structure in the UK is very different from that in the US, the JR_{DB} provides an extremely useful validation cohort for the generalisation performance of OASIS. Evaluating a model in an institution that is both temporally and geographically distinct from the development cohort is highlighted as the strongest evaluation of a model's performance [150].

6.2 Methodology

The evaluation of severity scores in the JR_{DB} proceeded much as the evaluation of severity scores in the MIMIC database in Chapter 5. First, JR_{DB} was assessed to determine which severity scores could feasibly be computed. This assessment isolated the SAPS II [9], APS III [7], OASIS [138] and the ICNARC physiology score [60] as feasible candidates. Note that the ICNARC physiology score is a subcomponent of the ICNARC risk prediction model, much as the APS III is a subcomponent of APACHE IV. In order to emphasise this distinction, the ICNARC physiology score is referred to here as the IPS. Note that the IPS was not applied in previous chapters as it requires covariates unavailable both in AO and in MIMIC II.

After extracting the various severity scores, calibration coefficients from the original article (if available) were applied to convert each score into a risk of mortality. Calibra-

[1]Research Ethics Committee Reference: 11\SC\0440 (Phase 1), 12\SC\0357 (Phase 2). International Standard Randomised Controlled Trial Number: 32008295.

tion coefficients were available from the original publishing article for SAPS II [9] and OASIS [138]. Calibration coefficients for the APS III were derived from a later recalibration [138] which was described in Chapter 4 (see Section 4.6.1 for coefficients). There are no calibration coefficients available for the IPS and as such its calibration was not evaluated. Thus, the analysis of the severity scores proceeded in two stages. In the first stage, the severity scores were calculated for all patients in the dataset and the AUROCs were compared. Statistical significance and confidence intervals were calculated using the non-parametric method of DeLong and DeLong [127]. The second analysis only evaluated the SAPS II, OASIS and APS III and focused on model calibration. Calibration curves were plotted for the three severity scores and the B, B_{adj}, SMR, $HL_{\hat{C}}$ and $\mathcal{I}_{\mathcal{L}}$ were calculated. For each metric, the confidence intervals were calculated as the 5^{th} and 95^{th} percentiles of the distribution generated from 5,000 bootstrap samples of the prediction target pairs [49]. In both analyses the risks of mortality were evaluated across the entire JR_{DB} and no recalibration is performed. The probability density function for the severity scores was calculated using kernel density estimation (see Appendix B.1) and plotted against histograms of the same data to aid in visualising the distributions.

6.2.1 Data extraction

Data was extracted from two sources. The first was a database copy of the electronic data management system at the John Radcliffe hospital. Custom queries were created to extract variables required for each severity score. All physiologic parameters, such as heart rate or blood pressure, were extracted from the database copy. While there are multiple ICUs at the John Radcliffe hospital, only data from the adult general ICU were extracted. The second source was from the case-mix programme (CMP) [151], a national audit for UK institutions of which the John Radcliffe is a member. This programme requires hospitals to manually enter a specified number of fields primarily for the purpose of risk adjustment by the ICNARC model, and this data is stored locally at the John Radcliffe. This data source was used to acquire the comorbidity information for the calculation of SAPS II. Information regarding in hospital mortality was also acquired from the CMP data. A comparison of the variables used for each severity score is shown

in Table 6.1. See Section 1.3 for a description of each variable.

6.2.2 Exclusion criteria

Each severity score was originally developed after excluding certain patients from the data due to a variety of reasons. One such example is the exclusion of paediatric patients due to their differing physiology as compared to adult patients. For the purposes of this comparison, the most stringent exclusion criteria used for developing any of the severity scores were applied. These exclusions were: patients $<$ 16 years of age, transplant patients except for liver or kidney transplants, patients staying in the ICU for less than four hours, patients admitted for coronary artery bypass grafts and patients admitted with severe burns.

6.3 Results

6.3.1 Demographics

Patients were excluded based upon the primary admission diagnosis as coded by the ICNARC coding method [152]. Of the 3,577 patients, 136 patients (3.80%) were excluded as they were undergoing a transplant surgery that did not involve a kidney or a liver. One patient (0.03%) was excluded for being admitted with burns and 12 (0.33%) were excluded for undergoing coronary artery bypass graft surgeries. As all data collected were from an adult medical ICU, no patients were excluded based upon age. There were 64 patients (1.79%) missing information regarding hospital mortality, and these patients were removed from the dataset. A total of 211 patients (5.90%) were removed. This resulted in a final dataset of 3,366 patients. The demographics of the dataset are shown in Table 6.2.

A smoothed estimate of the probability density functions, calculated using kernel density estimation [146] (as detailed in the Appendix B.1), is shown in Figure 6.1.

	OASIS	IPS	SAPS II	APS III
A-aDO_2				✓
Age	✓		✓	
Albumin				✓
Bilirubin			✓	✓
Bicarbonate			✓	
Blood pressure - mean	✓			✓
Blood pressure - systolic		✓	✓	
Blood urea nitrogen		✓	✓	✓
Chronic dialysis				✓
Creatinine		✓		✓
Elective surgery	✓		✓	
Emergency surgery			✓	
FiO_2		✓		✓
Glasgow coma scale	✓	✓	✓	✓
Glucose				✓
Haematocrit				✓
Heart rate	✓	✓	✓	✓
PaO_2		✓	✓	✓
$PaCO_2$				✓
pH		✓		✓
Platelets				
Potassium			✓	
Pre-ICU length of stay	✓			
Respiratory rate	✓	✓		✓
Sedated or paralysed		✓		
Sodium		✓	✓	✓
Temperature	✓	✓	✓	✓
Urine output (daily)	✓	✓	✓	✓
Ventilation	✓	✓	✓‡	✓
White blood cell count		✓	✓	✓
Comorbidities				
AIDS			✓	
Metastatic cancer			✓	
Leukemia, Lymphoma, and/or Immunosuppression			✓	
Chronic dialysis				✓

Table 6.1: *Variables used in the OASIS, SAPS II, IPS and APS III.*
‡Also includes continuous positive airway pressure.

Demographic	Mean or %	[25th, 75th] percentile
OASIS	33	[25, 41]
APS III	50	[36, 72]
SAPS II	38	[27, 52]
ICNARC	18	[12, 26]
Age	59.6	[47.0, 74.1]
Gender (male)	60.3%	
Ventilated‡	60.1%	
Admission type		
Medical	53.3	
Elective surgery	24.4	
Emergency surgery	22.3	
Outcomes		
ICU length of stay	5.13	[1.0, 4.9]
ICU Mortality	14.4%	
Hospital length of stay	22.4	[6.0, 26.0]
Hospital Mortality	21.9%	

Table 6.2: *Summary of severity scores and demographics of the patient cohort. ‡Presence of mechanical ventilation during the first day of their ICU admission.*

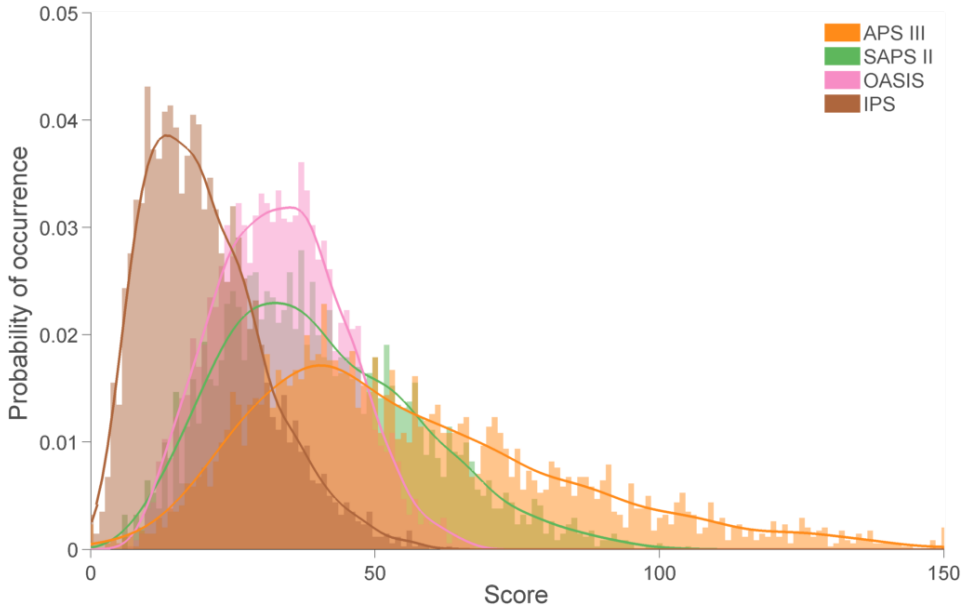

Figure 6.1: *Distribution of severity scores for first day admissions in the JR_{DB} after exclusions (3,366 patients). Each line represents the estimated probability destination function using kernel density estimation. Histograms normalised to have unit area are provided below the estimated probability density function. The severity scores as ordered by the height of their mode from left to right are: ICNARC, OASIS, SAPS II and APS III.*

	Hospital mortality	ICU mortality
IPS	0.782 [0.763, 0.801]	0.832 [0.813, 0.851]
OASIS	0.776 [0.758, 0.795]	0.805 [0.785, 0.826]
SAPS II	0.767 [0.748, 0.786]	0.797 [0.777, 0.818]
APS III	0.751 [0.731, 0.771]	0.801 [0.780, 0.821]

Table 6.3: *AUROCs of the severity scores evaluated with confidence intervals calculated using the non-parametric technique described by DeLong and DeLong [127].*

	Hospital mortality			ICU mortality		
	IPS	SAPS II	OASIS	IPS	SAPS II	OASIS
APS III	<0.001	0.028	0.007	<0.001	0.687	0.626
IPS		0.021	0.452		<0.001	0.001
SAPS II			0.196			0.320

Table 6.4: *Statistical significance of pair-wise comparisons of the AUROC between severity scores using the non-parametric technique proposed by DeLong and DeLong [127]. No adjustment for multiple hypothesis testing has been applied.*

6.3.2 Severity score discrimination

Performance of the severity scores as measured by the AUROC (which accounts for only model discrimination) is provided in Table 6.3. Pair-wise tests of statistical significance are provided in Table 6.4 using the method of DeLong and DeLong [127]. Note that the tests in Table 6.4 have not been adjusted for multiple hypothesis testing. For hospital mortality, the difference between OASIS and ICNARC was not statistically significant. When evaluating hospital mortality, the APS III had the lowest discrimination (0.751) and was statistically significantly different from all other severity scores. The IPS had the highest discrimination (AUROC = 0.782) though the discrimination of OASIS (AUROC = 0.776) was not statistically significantly different ($p = 0.452$). For ICU mortality, the IPS had the highest discrimination (AUROC = 0.832) and this was statistically significantly different from all other severity scores ($p < 0.001$). The remaining severity scores were not statistically significantly different. A visual comparison of the severity score AUROCs with confidence intervals is provided in Figure 6.2. ROC curves are provided in Appendix D.2.2.

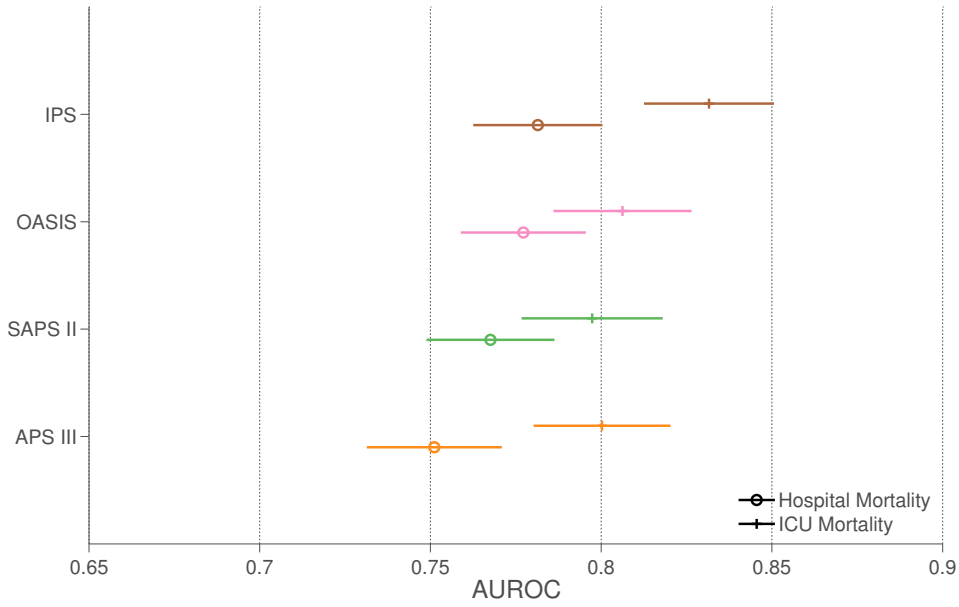

Figure 6.2: *AUROCs for the various severity scores plotted with 95% confidence intervals as calculated using the method of DeLong and DeLong [127]. Circles indicate hospital mortality is the outcome used for the calculation of the AUROC, whereas crosses indicate ICU mortality was used.*

6.3.3 Severity score calibration

The calibration of the SAPS II, APS III and OASIS is reported in Table 6.5. Note that Table 6.5 does not contain calibration information for SAPS II on ICU mortality as these calibration coefficients were not published in the original article [9]. Calibration for SAPS II on hospital mortality was poor, with $HL_{\hat{C}} = 401.4$ and SMR $= 0.704$ indicating overprediction of mortality. Calibration for the APS III and OASIS was good, though both overpredicted mortality with SMR values > 1. $HL_{\hat{C}}$ values were lowest for OASIS and highest for SAPS II. In terms of overall model fit, OASIS had the highest $\mathcal{I}_{\mathcal{L}}$, highest B_{adj} and lowest B for both hospital mortality and ICU mortality, though the APS III had an SMR closer to 1 for hospital mortality.

	Hospital mortality		
Metric	APS III	SAPS II	OASIS
SMR	1.107 [1.042, 1.174]	0.704 [0.663, 0.745]	1.122 [1.057, 1.189]
$HL_{\hat{C}}$	93.8 [63.3, 160.7]	401.4 [316.5, 521.1]	49.8 [30.7, 104.4]
$\mathcal{I}_{\mathcal{L}}$	0.113 [0.081, 0.146]	0.054 [0.010, 0.096]	0.158 [0.128, 0.187]
B	0.149 [0.140, 0.157]	0.163 [0.156, 0.171]	0.141 [0.134, 0.150]
B_{adj}	0.132 [0.098, 0.165]	0.047 [0.004, 0.095]	0.175 [0.145, 0.205]

	ICU mortality		
Metric	APS III	SAPS II	OASIS
SMR	1.073 [0.991, 1.154]	-	1.073 [0.994, 1.155]
$HL_{\hat{C}}$	72.6 [47.9, 128.0]	-	28.7 [17.7, 84.1]
$\mathcal{I}_{\mathcal{L}}$	0.163 [0.126, 0.198]	-	0.188 [0.151, 0.224]
B	0.106 [0.098, 0.114]	-	0.101 [0.094, 0.109]
B_{adj}	0.144 [0.104, 0.183]	-	0.181 [0.145, 0.219]

Table 6.5: *Statistical significance of calibration metrics for the severity scores evaluated on the JR_{DB}. 95% confidence intervals calculated using the bootstrap percentile method [49] and 5,000 bootstrap samples are provided.*

Calibration curves are provided in Figure 6.3 for the severity scores with hospital mortality as the dependent variable. All models exhibited a lack of fit accoding to the $HL_{\hat{C}}$. Both OASIS and APS III have calibration curves which indicate underprediction of hospital mortality in the lower ranges of risk. The APS III overpredicts hospital mortality in the highest decile of patient risk, while OASIS appears to be very well calibrated for this decile. SAPS II appears well calibrated for the first four deciles of risk, and then gradually begins to markedly overpredicting mortality (28.1% lower in the second highest decile and 27.0% lower in the highest decile). Tables containing the

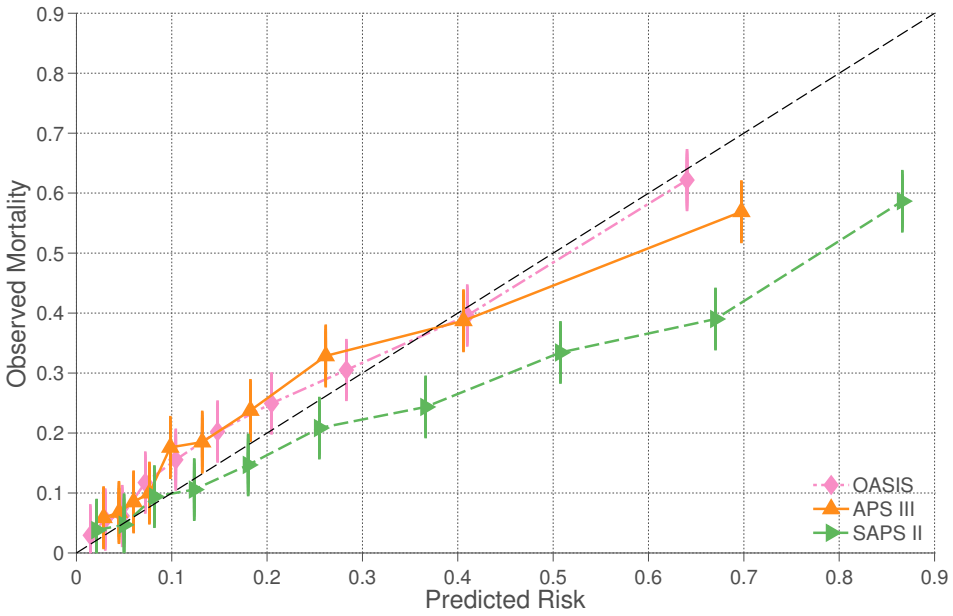

Figure 6.3: *Calibration curves for the three severity scores for which calibration coefficients were available. The dashed line represents the ideal performance of a model where the expected mortality is always equal to the observed mortality. Scores above this line underpredict mortality, while scores below this line overpredict mortality. The calibration curves are calculated across the entire JR_{DB} of 3,366 patients.*

exact values in each decile are provided in Appendix D.1.2.

6.4 Discussion

All severity scores evaluated performed well. In terms of discriminating hospital mortality as measured by the AUROC, the worst performing severity score was the APS III. As APS III was developed on a purely American cohort in 1993 [7], it is not unreasonable to expect lower performance on a UK populaton admitted between 2008-2011. The AUROC = 0.751 still indicates that the APS III discriminates hospital mortality well. In order to contextualise the performance of these models, it it beneficial to examine previous studies evaluating severity scores in UK populations. Livingston *et al.* tested the APACHE III system in a large Scottish database and it achieved an AUROC = 0.845 [65]. As the APACHE III system contains comorbidities, age and diagnosis as covariates it is reasonable to expect the APS III to have lower discrimination on its own. In the large multi-center study in which the ICNARC model was developed [34], the AUROC of the APACHE III system = 0.845. Zimmerman *et al.* [137] assessed the importance of the various parameters in the APACHE IV hospital mortality pre-

diction model and found that physiology accounted for 66% of the model's predictive performance. Other important covariates included diagnosis (16%), age (9%), comorbidities (5%) and other admission variables including length of stay prior to admission and source of admission (3%). While acute physiology clearly accounts for the majority of the predictive power in APACHE IV there is a significant contribution from diagnosis, age and comorbidities, all of which are present in the APACHE III model but not in the APS III. Nevertheless, the performance of severity scores varies from institution to institution, and the lower performance of APS III on the JR_{DB} may simply represent a case-mix not well predicted by the severity score.

OASIS discriminates well on the UK based JR_{DB} with an AUROC = 0.776 for hospital mortality and an AUROC = 0.805 for ICU mortality. There was no statistically significant difference between the discrimination of the IPS and the OASIS (as measured by the AUROC) indicating that both discriminate equally well on the JR_{DB}. The IPS was developed using backward step-wise feature selection and only retained features if they significantly improved model fit. Conversely, OASIS was developed using a hybrid GA and PSO approach which allowed for multivariate feature selection. Given the reduced feature set of OASIS (10 features versus 15 in the IPS) and equivalent performance, it is reasonable to postulate that the hybrid GA and PSO approach has allowed OASIS to capture almost as much information as the IPS while only using two thirds the number of features. Still, the IPS had a statistically significantly higher AUROC for ICU mortality. This could indicate that while OASIS and the IPS capture an equivalent amount of physiologic derangement which is indicative of eventual in hospital mortality, the IPS contains information which is specific to patients dying before ICU discharge.

The calibration of the APS III and OASIS was very good, especially considering the models were originally calibrated on an American dataset. Though neither of the SMRs for the APS III or OASIS contained 1, visual inspection of the calibration curve indicates a reasonable fit (though the highest decile of the APS III appears poorly calibrated). SAPS II greatly overpredicted mortality (SMR = 0.704), especially in the higher ranges of risk. Both the APS III and OASIS benefited from a much more recent sample size (data collected between 2007-2009 [138]) as opposed to SAPS II which calibrated poorly

(data collected in 1993 [9]). Though SAPS II was collected at European and other institutions, case-mix has changed in the subsequent decades and the use of an international population for the development of SAPS II was insufficient to provide good calibration. As the final APS III value was recalibrated to data from 2007-2009, and not the individual score components, the significant changes in practice over two decades will have contributed to the reduction in the calibration of the APS III.

While the statistical significance has been quantified in Table 6.4, the clinical significance of the variation in performance is a matter of debate. As a result practical concerns, such as the burden of data collection, become key factors in the use of a severity score. While the ideal situation can be imagined, with monitors continuously providing the necessary data to computers which automatically calculate the desired severity scores, this vision has yet to materialise. Though the database from which the ICNARC model is impressive in its scale with hundreds of ICUs and over 200,000 admissions [34], many participant ICUs continue to enter all their data manually. Similarly, some hospitals equipped with the APACHE IV system still utilise manual data entry for the various components of the model. Even for large economic entities whose focus is on ICU severity scoring, manual input still plays a critical role. Any reduction in the amount of data entry required would substantially improve the efficiency of the severity score estimation for large cohorts of patients. Data abstraction was studied by Kuzniewicz *et al.*, who found that MPM_0-III required 11.1 minutes, SAPS II required 19.6 minutes and APACHE IV required 37.3 minutes [153]. Though the absolute speed of abstraction is highly dependent on the software utilised, training of the personnel and ease of access to electronic medical records, the relative times still provide some insight into the additional burden of higher dimensional models. The reduction in the number of features between the MPM_0-III and SAPS II (16 in MPM_0-III versus 19 in SAPS II) compared to the reduction in time taken (11.1 minutes for MPM_0-III versus 19.6 minutes for the SAPS II) indicates that a reduction in features may not be linearly related to a reduction in abstraction time. The simplicity of the variables in OASIS, all based on direct measurements of physiology, compared to other severity scores which use potentially ambiguous diagnoses and comorbidities provides a stronger impetus for its

use. However, interrater reliability studies of OASIS are required in order to substantiate this claim.

There are limitations to the presented study. First, the ICNARC Physiology Score (IPS) is not directly used to benchmark UK institutions participating in the CMP. The full ICNARC model, which incorporates the IPS, contains many more covariates including diagnosis and various comorbidies. This model had higher performance than the IPS in the initial publication [34], and would be expected to perform better both in the presented cohort and across the UK as compared to the IPS. However, the coefficients for the ICNARC model are not publicly available. Further study is needed to validate the use of OASIS when compared to the full ICNARC model in addition to the IPS.

A second limitation is the use of a single center for the study. As case-mix can vary widely between institutions depending on the demographics of patients admitted, it is difficult to discuss the generalisability of the results presented. Furthermore, the sample size studied is small relative to the large development populations of the SAPS II (12,997) [9], APS III (17,440) [7] and OASIS (81,007) [138]. A large multi-center study of OASIS in the UK would provide a more robust estimate of its generalisation performance outside of the United States.

Chapter 7

Conclusions and future work

As the births of living creatures at first are ill-shapen, so are all innovations, which are the births of time.

Francis Bacon

This thesis has examined mortality prediction for critically ill patients in detail. A database released for the Physionet/Computing in Cardiology 2012 challenge was described. Various machine learning models were benchmarked against the more common regression models. RLR and RLR^2 were shown to perform competitively with more complex models which have the potential to capture higher level interactions such as SVMs and RFs. Further benchmarking of the same models in a large, multi-center database showed that in some cases the simpler regression models outperformed the SVM and the RF, but that this performance hinged on appropriate regularisation and preprocessing of the data.

The primary conclusion of the model evaluations appears to be that logistic regression is an excellent technique for predicting mortality in ICU patients. However, there are a few key caveats, and the regression models applied in this work have been carefully applied in order to achieve such high performance. There are three key aspects of the regression model which led to the improved performance. First, regularisation played a pivotal role in controlling the complexity of the regression models. The regularisation parameter was learned using cross-validation to ensure that the regularisation parameter

selected would produce a model which generalised well. Without the inclusion of regularisation, the amount of data would have been insufficient to learn stable coefficients for the models in Chapter 2, and the resultant models would have had poor performance. In particular, the use of the L_1 norm provided sparse models which have the added benefit of reduced data burden. Second, the use of a custom data preprocessing algorithm to remove outliers substantially improved model performance. While BCOR provided an elegant and conceptually simple method for removing outliers, the primary concept to take note of is the removal of values at the extrema of the distribution which are unphysiological and highly likely to be erroneous. Due to the chaotic nature of the ICU environment and the extreme severity of illness of the patients, sensors often acquire inaccurate data but do not exclude it for concern of censoring data indicative of a potentially fatal event (i.e. removing a heart rate of zero when the patient is in cardiac arrest). As such the data acquired is noisy and contains large numbers of artefacts. One feasible approach that is often applied is the use of domain knowledge to filter data based upon physiologically reasonable thresholds. However, this is a time consuming task and requires a large amount of domain knowledge. Furthermore, the technical implementation is tedious and sensitive to the unit of measurement which differs between many care providers (including the US and the UK). Nevertheless, implementation of some form of data preprocessing to remove outliers is critical. Finally, the addition of square terms as covariates in the model also improved performance over regular regression models (i.e. RLR^2 was usually equivalent or better than RLR). This is a technically simple step which consistently provided performance gains with very little additional effort over training a regression model without square terms. While it is a logical continuation of this approach to consider interaction terms or higher order interactions, the drawback of these extensions is rapidly increasing model complexity. The addition of square terms is a comfortable balance between increasing model complexity and improving model performance. Overall, the inclusion of these three steps in the regression models substantially improved performance. In the case of the PN_{db} database, the latter two steps improved the AUROC of the model by 0.015, and this required no additional data collection. In actual fact, the gain in performance was much higher as unregularised logistic regression

models failed to accurately estimate the coefficients and had AUROCs < 0.7.

A novel parsimonious severity score, OASIS, was then developed using a hybrid GA and PSO approach which allowed direct optimisation of a severity score in a clinically relevant form with simultaneous multivariate feature selection. The new score discriminated better than the APS III when evaluated in a univariate fashion and equivalently when evaluated as a covariate in a larger risk adjustment model similar to APACHE IV. OASIS also has an extremely low burden for data collection and quality control, requiring only 10 features and does not require laboratory measurements, diagnosis or comorbidity information. These performances were calculated on data from patients admitted to ICUs after the development cohort which provided a temporal validation of the score. In order to further validate the score, data from the MIMIC II database was extracted and five severity scores were compared, including OASIS. OASIS did not have the highest discrimination for hospital mortality (AUROC = 0.790 versus SAPS II AUROC = 0.802), though OASIS had excellent calibration (SAPS II greatly overpredicted mortality). To provide a final robust assessment of OASIS, data from the John Radcliffe hospital in the United Kingdom was acquired and four severity scores were evaluated. The IPS component of the ICNARC model had the highest discrimination for hospital mortality though this was statistically insignificantly different from OASIS (IPS AUROC = 0.782 versus OASIS AUROC = 0.776). OASIS had excellent calibration though it tended to underpredict mortality in lower ranges of risk.

It is interesting to note the comparable performance of the APS III, SAPS II and OASIS despite the variety of features utilised as inputs. The APS III includes arterial pH and serum albumin while SAPS II includes serum bicarbonate and serum potassium. While every version of the APACHE system has included respiration rate, this feature is excluded from the SAPS II model. The fact that these models exclude different variables and yet possess similar levels of discrimination is a strong indicator that there is a large amount of correlation, and redundancy, among these variables.

Overall, OASIS consistently had good discrimination and calibration across three independent evaluations: a large multi center evaluation in the United States, a completely external single center validation in the United States and a final geographically

distinct validation in a single ICU in the United Kingdom. OASIS is a well calibrated severity score with good discrimination which outperformed most severity scores it was compared to while requiring substantially fewer features.

There remains room for improvement in the models developed in this thesis. First, the features extracted from the PN_{db} were simple aggregate features of the time series. Techniques which better summarise the information contained in time series, such as autoregressive models [113], Gaussian processes [154] and in particular Gaussian processes based upon extreme value theory (EVT) [155] may provide the improved performance that more complex classifiers could not attain. Similarly, the BCOR developed utilised a simple critical ratio test with application of the Bonferroni correction. The field of EVT provides a well principled approach for modelling and isolating extreme values from distributions of maximum or minimum values. The BCOR may be improved by instead defining the statistical significance threshold based upon a distribution motivated by EVT, rather then the conservative use of the Bonferroni correction and potentially inappropriate assumption of a normal distribution.

The use of a GA and PSO together to optimise OASIS allowed for a parsimonious severity score which has performed well on a variety of databases. Nevertheless, the GA and PSO suffer from a number of limitations, primarily of which is the many hyperparameters involved in the algorithms. In fact, the optimisation of hyperparameters is itself an ongoing topic of research [156, 157]. One particularly interesting concept is of *random* search which has been shown to have competitive results with more labourious grid and sequential search algorithms while providing advantages in terms of parallelisation and ease of implementation [157]. In this work, the hyperparameters were selected without a rigorous framework, and one avenue of improvement would be the optimisation of the hyperparameters formally. While random search is the simplest approach, other techniques such as gradient descent or Gaussian process regression have been applied to more efficiently search the space of hyperparameters [158].

The goal of models in this thesis was excellent generalisation performance in data collected at an external site. However, variation in patient case-mix and care procedures has been argued to justify institution specific modelling of mortality [147]. Such

approaches become increasingly feasible as the amount of data which is collected and archived increases for individual critical care units [159]. It would be interesting to test this hypothesis by developing a model from the large multi-center AO database and a distinct model using only data from the MIMIC II database. When the guardians of the MIMIC II database release the subsequent cohort of patients who were admitted to the Beth Israel Deaconess Medical Center both models could be evaluated on "future" data (from the model's perspective). The comparison would shed light on whether large multi-center datasets or single center highly customised datasets are preferable for the development and application of mortality prediction models. Interesting analysis which could also be performed concurrently would be the evaluation of the models developed Chapter 2 with the severity of illness scores presented in Chapter 5. Due to missing information in the PN_{db} direct calculation of the severity scores was not possible. However if the same variables present in the PN_{db} were directly extracted from the MIMIC II database (their original source) then such a comparison would be possible and would shed light on the power of generic ICU risk adjustment models as compared to local customised modelling.

The possibility of acuity monitoring using smaller time intervals is raised by OASIS as it does not include any laboratory measurements and does not primarily rely on comorbidity or diagnostic information. Since OASIS was developed using data extracted from the first 24 hours of a patient stay it is currently inappropriate to apply the model to data collected over smaller windows (including real time). Nevertheless, a previous study by Ho *et al.* [64] showed that the APACHE II model calculated using only data captured within an hour of admission performed equivalently to the APACHE II model calculated using data captured over the first 24 hours. This study could feasibly be reproduced using either the MIMIC II database or the JR_{DB} to demonstrate that OASIS could be used in a similar fashion. This would improve the broad applicability of the score as many patients are discharged before their first day in the ICU has completed.

Given the extensive validation of OASIS in multiple institutions in multiple countries, the score shows promise for having captured true physiologic derangement which is independent of the care structure of the given institution. While it is infeasible that

OASIS has adjusted for every possible case-mix, the empirical study in this thesis has shown that OASIS has excellent generalisation performance. The evaluation of regression models against more complex machine learning methods corroborates the notion that with proper feature subsets simpler models can perform competitively with, or better than, more complicated approaches. OASIS provides researchers and clinicians with a simple severity score which provided equivalent or slightly reduced performance as compared to other severity scores. Furthermore, the empirical evaluation of various models has broad applications for ICU prediction as the deluge of data collected and warehoused will provide ample opportunity for future model development.

Bibliography

[1] L. Tarassenko, D. A. Clifton, M. R. Pinsky, M. T. Hravnak, J. R. Woods, and P. J. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, pp. 1013–8, Aug. 2011.

[2] National Health Service, "Guidelines on admission to and discharge from intensive care and high dependency units Department of Health," Jan. 1996.

[3] P. Pronovost, D. Angus, T. R. Dorman, K. A. Dremsizov, and T. T. Young, "Physician Staffing Patterns and Clinical Outcomes in Critically Ill Patient: A Systematic Review," *JAMA*, vol. 288, no. 17, pp. 2151–2162, 2002.

[4] R. Kane, T. Shamliyan, C. Mueller, S. Duval, and T. J. Wilt, "The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis," *Medical Care*, vol. 45, pp. 1195–1204, Dec. 2007.

[5] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, "APACHE-acute physiology and chronic health evaluation: a physiologically based classification system," *Critical Care Medicine*, vol. 9, pp. 591–597, 1981.

[6] W. A. Knaus, "APACHE 1978-2001: the development of a quality assurance system based on prognosis: milestones and personal reflections," *Archives of Surgery*, vol. 137, no. 1, pp. 37–41, 2002.

[7] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, C. A. Bastos, P G snd Sirio, D. J. Murphy, T. Lotring, A. Damiano, and F. E. Harrell Jr., "The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[8] J. R. LeGall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers, "A simplified acute physiology score for ICU patients," *Critical Care Medicine*, vol. 12, pp. 975–977, 1984 1984. PT: J.

[9] J. R. LeGall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS-II) based on a european north-american multicenter study," *JAMA*, vol. 270, pp. 2957–2963, DEC 22 1993 1993. PT: J.

[10] S. Lemeshow, D. Teres, and H. Pastides, "A method for predicting survival and mortality of ICU patients using objectively derived weights," *Critical Care Medicine*, vol. 13, pp. 519–525, 1985.

[11] S. Lemeshow, D. Teres, and J. Klar, "Mortality probability model (MPM II) based on an international cohort of intensive care unit patients," *JAMA*, vol. 270, pp. 2478–2486, 1993.

[12] J.-L. Vincent, F. Ferreira, and R. Moreno, "Scoring systems for assessing organ dysfunction and survival," *Critical Care Clinics*, vol. 16, no. 2, pp. 353–366, 2000.

[13] J.-L. Vincent and R. Moreno, "Clinical review: scoring systems in the critically ill," *Critical care*, vol. 14, p. 207, Jan. 2010.

[14] H. Wunsch, W. T. Linde-Zwirble, and D. C. Angus, "Methods to adjust for bias and confounding in critical care health services research involving observational data," *Journal of Critical Care*, vol. 21, no. 1, pp. 1–7, 2006.

[15] D. Bennett and B. J, "ABC of intensive care: Organisation of intensive care," *British Medical Journal*, vol. 318, no. 7196, pp. 1468–1470, 1999.

[16] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "An Evaluation of Outcome from Intensive Care in Major Medical Centers," *Annals of Internal Medicine*, vol. 104, pp. 410–418, Mar. 1986.

[17] N. A. Christakis, *Death foretold: prophecy and prognosis in medical care*. University of Chicago Press, 2001.

[18] Bristol Royal Infirmary Inquiry, *The report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995*. Stationery Office, 2001.

[19] G. Teasdale, "Learning from bristol: report of the public inquiry into children's heart surgery at bristol royal infirmary 1984-1995," *British Journal of Neurosurgery*, vol. 16, no. 3, pp. 211–216, 2002.

[20] Department of Health, "NHS reference costs 2004-2005," Apr. 2006.

[21] P. S. Halpern NA, "Critical Care Medicine in the United States 2000-2005: An analysis of bed numbers, occupancy rates, payer mix, and costs," *Critical Care Medicine*, vol. 38, no. 1, pp. 65–71, 2010.

[22] P. Griner, "Medical intensive care in the teaching hospital: costs versus benefits. the need for assessment.," *Annals of Internal Medicine*, vol. 78, no. 4, pp. 581–585, 1973.

[23] W. A. Knaus, D. P. Wagner, J. E. Zimmerman, and E. A. Draper, "Variations in mortality and length of stay in intensive care units," *Annals of Internal Medicine*, vol. 118, no. 10, pp. 753–761, 1993.

[24] E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, and M. Tomlanovich, "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *New England Journal of Medicine*, vol. 345, no. 19, pp. 1368–1377, 2001.

[25] G. R. Bernard, J.-L. Vincent, P.-F. Laterre, S. P. LaRosa, J.-F. Dhainaut, A. Lopez-Rodriguez, J. S. Steingrub, G. E. Garber, J. D. Helterbrand, E. W. Ely, *et al.*, "Efficacy and safety of recombinant human activated protein c for severe sepsis," *New England Journal of Medicine*, vol. 344, no. 10, pp. 699–709, 2001.

[26] ARDS Network, "Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome.," *New England Journal of Medicine*, vol. 342, no. 18, pp. 1302–1308, 2000.

[27] A. Anzueto, R. P. Baughman, K. K. Guntupalli, J. G. Weg, H. P. Wiedemann, A. A. Raventós, F. Lemaire, W. Long, D. S. Zaccardelli, and E. N. Pattishall, "Aerosolized surfactant in adults with sepsis-induced acute respiratory distress syndrome," *New England Journal of Medicine*, vol. 334, no. 22, pp. 1417–1422, 1996.

[28] J. E. Zimmerman, D. P. Wagner, W. A. Knaus, J. F. Williams, D. Kolakowski, and E. A. Draper, "The use of risk predictions to identify candidates for intermediate care units implications for intensive care utilization and cost," *Chest*, vol. 108, no. 2, pp. 490–499, 1995.

[29] P. J. Pronovost, D. C. Angus, T. Dorman, K. A. Robinson, T. T. Dremsizov, and T. L. Young, "Physician staffing patterns and clinical outcomes in critically ill patients: a systematic review," *Jama*, vol. 288, no. 17, pp. 2151–2162, 2002.

[30] J. Curtis, R. Engelberg, M. Wenrich, E. Nielsen, S. Shannon, P. Treece, M. Tonelli, D. Patrick, L. Robins, and B. Mcgrath, "Studying communication about end-of-life care during the ICU family conference: Development of a framework," *Journal of Critical Care*, vol. 17, pp. 147–160, Sept. 2002.

[31] J. R. Curtis, "Patients' Perspectives on Physician Skill in End-of-Life Care* : Differences Between Patients With COPD, Cancer, and AIDS," *Chest*, vol. 122, pp. 356–362, July 2002.

[32] P. Singer, D. Martin, and M. Kelner, "Quality End-of-Life Care: Patients' Perspectives," *JAMA*, vol. 281, no. 2, pp. 163–168, 1999.

[33] D. Sackett, "Evidence-based medicine," *Seminars in Perinatology*, vol. 21, pp. 3–5, Feb. 1997.

[34] D. A. Harrison, G. J. Parry, and J. R. Carpenter, "A new risk prediction model for critical care: the intensive care national audit & research centre (ICNARC) model," *Critical Care Medicine*, vol. 35, pp. 1091–1098, 2007.

[35] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients.," *Critical Care Medicine*, vol. 34, pp. 1297–1310, May 2006.

[36] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, Sept. 1995.

[37] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Sept. 2001.

[38] L. I. Iezzoni, *Risk Adjustment for Measuring Health Care Outcomes*. Health Administration Press, 2003.

[39] D. C. Hadorn, E. B. Keeler, W. H. Rogers, and R. H. Brook, "Assessing the performance of mortality prediction models," 1993.

[40] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *Journal of Biomedical Informatics*, vol. 38, pp. 404–15, Oct. 2005.

[41] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons Inc, 2000.

[42] A. A. Kramer and J. E. Zimmerman, "Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited," *Critical Care Medicine*, vol. 35, no. 9, pp. 2052–2056, 2007.

[43] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[44] B. Efron, "Regression and anova with zero-one data: Measures of residual variation," *Journal of the American Statistical Association*, vol. 73, no. 361, pp. 113–121, 1978.

[45] E. W. Steyerberg, *Clinical prediction models*. Springer, 2009.

[46] A. A. Kramer, T. L. Higgins, and J. E. Zimmerman, "Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV Hospital Mortality Models: Implications for National Benchmarking*," *Critical Care Medicine*, vol. 42, no. 3, pp. 544–553, 2014.

[47] M. Mittlböck, M. Schemper, *et al.*, "Explained variation for logistic regression," *Statistics in Medicine*, vol. 15, no. 19, pp. 1987–1997, 1996.

[48] D. McFadden, "Conditional logit analysis of qualitative choice behavior," 1973.

[49] B. Efron and R. Tibshirani, *Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy*. JSTOR, 1986.

[50] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. Springer New York, 2006.

[51] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Clarendon press Oxford, 1995.

[52] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall/CRC, 1994.

[53] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, and E. A. Draper, "APACHE II: a severity of disease classification system," *Critical Care Medicine*, vol. 13, pp. 818–829, 1985.

[54] P. G. H. Metnitz, R. P. Moreno, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, and J.-R. Le Gall, "SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description," *Intensive Care Medicine*, vol. 31, pp. 1336–44, Oct. 2005.

[55] R. P. Moreno, P. G. H. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, and J.-R. Le Gall, "SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission," *Intensive Care Medicine*, vol. 31, pp. 1345–55, Oct. 2005.

[56] T. L. Higgins, D. Teres, W. S. Copes, B. H. Nathanson, M. Stark, and A. A. Kramer, "Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III)," *Critical Care Medicine*, vol. 35, pp. 827–35, Mar. 2007.

[57] J.-L. Vincent, R. Moreno, J. Takala, S. Willats, A. De Mendoca, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs, "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, pp. 707–710, 1996.

[58] K. Strand and H. Flaaten, "Severity scoring in the ICU: a review," *Acta Anaesthesiol Scand*, vol. 52, pp. 467–478, 2008.

[59] A. G. Schneider, M. Lipcsey, M. Bailey, D. V. Pilcher, and R. Bellomo, "Simple translational equations to compare illness severity scores in intensive care trials," *Journal of Critical Care*, vol. 28, pp. 885.e1–8, Oct. 2013.

[60] D. A. Harrison, A. R. Brady, G. J. Parry, J. R. Carpenter, and K. Rowan, "Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom," *Critical Care Medicine*, vol. 34, pp. 1378–1388, May 2006.

[61] S. Brinkman, F. Bakhshi-Raiez, A. Abu-Hanna, E. de Jonge, R. J. Bosman, L. Peelen, and N. F. de Keizer, "External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II," *Journal of Critical Care*, vol. 26, pp. 105.e11–8, Feb. 2011.

[62] N. Peek, D. Arts, R. Bosman, P. van der Voort, and N. De Keizer, "External validation of prognostic models for critically ill patients required substantial sample sizes," *Journal of clinical epidemiology*, vol. 60, no. 5, pp. 491–e1, 2007.

[63] D. H. Beck, G. B. Smith, J. V. Pappachan, and B. Millar, "External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study," *Intensive Care Medicine*, vol. 29, no. 2, pp. 249–256, 2003.

[64] K. M. Ho, G. J. Dobb, M. Knuiman, J. Finn, K. Y. Lee, and S. A. R. Webb, "A comparison of admission and worst 24-hour Acute Physiology and Chronic Health Evaluation II scores in predicting hospital mortality: a retrospective cohort study," *Critical Care*, vol. 10, no. 1, p. R4, 2005.

[65] B. M. Livingston, F. N. MacKirdy, J. C. Howie, R. Jones, and J. D. Norrie, "Assessment of the performance of five intensive care scoring models within a large Scottish database," *Critical Care Medicine*, vol. 28, pp. 1820–7, June 2000.

[66] R. Markgraf, G. Deutschinoff, L. Pientka, and T. Scholten, "Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit," *Critical Care Medicine*, vol. 28, pp. 26–33, Jan. 2000.

[67] M. J. Vassar, F. R. Lewis, J. A. Chambers, R. J. Mullins, P. E. O'Brien, J. A. Weigelt, M. T. Hoang, and J. W. Holcroft, "Prediction of outcome in intensive care unit trauma patients: a multicenter study of Acute Physiology and Chronic

Health Evaluation (APACHE), Trauma and Injury Severity Score (TRISS), and a 24-hour intensive care unit (ICU) point system," *The Journal of Trauma*, vol. 47, pp. 324–9, Aug. 1999.

[68] Y. Sakr, C. Krauss, A. C. K. B. Amaral, A. Réa-Neto, M. Specht, K. Reinhart, and G. Marx, "Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit," *British Journal of Anaesthesia*, vol. 101, pp. 798–803, Dec. 2008.

[69] G. Duke, M. Piercy, D. DiGiantomasso, and J. Green, "Comparison of intensive care outcome prediction models based on admission scores with those based on 24-hour data," *Anaesthesia & Intensive Care*, vol. 36, no. 6, pp. 845–850, 2008.

[70] P. Bastos, X. Sun, D. P. Wagner, W. A. Knaus, and J. Zimmerman, "Application of the APACHE III prognostic system in Brazilian intensive care units: A prospective multicenter study," *Intensive Care Medicine*, vol. 22, pp. 564–570, June 1996.

[71] M. Capuzzo, V. Valpondi, A. Sgarbi, S. Bortolazzi, V. Pavoni, G. Gilli, G. Candini, G. Gritti, and R. Alvisi, "Validation of severity scoring systems saps ii and apache ii in a single-center population," *Intensive Care Medicine*, vol. 26, pp. 1779–1785, Dec. 2000.

[72] S. Nouira, M. Belghith, S. Elatrous, M. Jaafoura, M. Ellouzi, R. Boujdaria, M. Gahbiche, S. Bouchoucha, and F. Abroug, "Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units," *Critical Care Medicine*, vol. 26, pp. 852–9, May 1998.

[73] B. Khwannimit and A. Geater, "A comparison of APACHE II and SAPS II scoring systems in predicting hospital mortality in Thai adult intensive care units," *Medical Association of Thailand*, vol. 90, no. 4, p. 643, 2007.

[74] K. M. Ho, K. Y. Lee, T. Williams, J. Finn, M. Knuiman, and S. A. R. Webb, "Comparison of Acute Physiology and Chronic Health Evaluation (APACHE) II score with organ failure scores to predict hospital mortality," *Anaesthesia*, vol. 62, pp. 466–73, May 2007.

[75] D. H. Beck, B. L. Taylor, B. Millar, and G. B. Smith, "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*, vol. 25, no. 1, pp. 9–15, 1997.

[76] R. Moreno and P. Morais, "Outcome prediction in intensive care: results of a prospective, multicentre, portuguese study," *Intensive Care Medicine*, vol. 23, no. 2, pp. 177–186, 1997.

[77] J. Y. Kim, S. Y. Lim, K. Jeon, Y. Koh, C.-M. Lim, S. O. Koh, S. Na, K. M. Lee, B. H. Lee, J.-Y. Kwon, *et al.*, "External Validation of the Acute Physiology and Chronic Health Evaluation II in Korean Intensive Care Units," *Yonsei Medical Journal*, vol. 54, no. 2, pp. 425–431, 2013.

[78] S. Katsaragakis, K. Papadimitropoulos, P. Antonakis, S. Strergiopoulos, M. M. Konstadoulakis, and G. Androulakis, "Comparison of Acute Physiology and Chronic Health Evaluation II (APACHE II) and Simplified Acute Physiology Score II (SAPS II) scoring systems in a single Greek intensive care unit," *Critical Care Medicine*, vol. 28, pp. 426–32, Feb. 2000.

[79] S. Christensen, M. B. Johansen, C. F. Christiansen, R. Jensen, and S. Lemeshow, "Comparison of Charlson comorbidity index with SAPS and APACHE scores for prediction of mortality following intensive care," *Clinical Epidemiology*, vol. 3, no. 1, pp. 203–211, 2011.

[80] P. A. Patel and B. J. B. Grant, "Application of mortality prediction systems to individual intensive care units," *Intensive Care Medicine*, vol. 25, pp. 977–982, Sept. 1999.

[81] E. Paul, M. Bailey, and D. Pilcher, "Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: development and validation of the Australian and New Zealand Risk of Death model," *Journal of Critical Care*, vol. 28, pp. 935–41, Dec. 2013.

[82] J. E. Zimmerman, D. P. Wagner, E. A. Draper, L. Wright, C. Alzola, and W. A. Knaus, "Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database," *Critical Care Medicine*, vol. 26, pp. 1317–26, Aug. 1998.

[83] F. Shann, J. Santamaria, D. Ernest, P. Stow, C. George, G. Duke, and D. Pilcher, "Critical care outcome prediction equation (COPE) for adult intensive care," vol. 10, p. 35, Mar. 2008.

[84] M. T. Keegan, O. Gajic, and B. Afessa, "Comparison of APACHE III, APACHE IV, SAPS 3, and MPM_0III and influence of resuscitation status on model performance," *Chest*, vol. 142, no. 4, pp. 851–858, 2012.

[85] V. Pettilä, M. Pettilä, S. Sarna, P. Voutilainen, and O. Takkunen, "Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill," *Critical Care Medicine*, vol. 30, pp. 1705–11, Aug. 2002.

[86] M. W. Kuzniewicz, E. E. Vasilevskis, R. Lane, M. L. Dean, N. G. Trivedi, D. J. Rennie, T. Clay, P. L. Kotler, and R. A. Dudley, "Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders," *Chest*, vol. 133, pp. 1319–27, June 2008.

[87] J. R. Le Gall, A. Neumann, F. Hemery, J. P. Bleriot, J. P. Fulgencio, B. Garrigues, C. Gouzes, E. Lepage, P. Moine, and D. Villers, "Mortality prediction using SAPS II: an update for French intensive care units," *Critical Care*, vol. 9, pp. R645–52, Jan. 2005.

[88] A. Reiter, W. Mauritz, B. Jordan, T. Lang, A. Pölzl, L. Pelinka, and P. G. H. Metnitz, "Improving risk adjustment in critically ill trauma patients: the TRISS-SAPS Score," *Journal of Trauma*, vol. 57, pp. 375–80, Aug. 2004.

[89] P. Aegerter, A. Boumendil, A. Retbi, E. Minvielle, B. Dervaux, and B. Guidet, "SAPS II revisited," *Intensive Care Medicine*, vol. 31, pp. 416–23, Mar. 2005.

[90] Ø. Haaland, F. Lindemark, H. Flaatten, R. Kvåle, and K. Johansson, "A calibration study of SAPS II with Norwegian intensive care registry data," *Acta Anaesthesiologica Scandinavica*, 2014.

[91] R. Moreno, D. R. Miranda, V. Fidler, and R. Van Schilfgaarde, "Evaluation of two outcome prediction models on an independent database," *Critical Care Medicine*, vol. 26, pp. 50–61, Jan. 1998.

[92] D. Poole, C. Rossi, N. Latronico, G. Rossi, S. Finazzi, and G. Bertolini, "Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better?," *Intensive Care Medicine*, vol. 38, no. 8, pp. 1280–1288, 2012.

[93] P. G. Metnitz, T. Lang, H. Vesely, A. Valentin, and J. Le Gall, "Ratios of observed to expected mortality are affected by differences in case mix and quality of care," *Intensive Care Medicine*, vol. 26, no. 10, pp. 1466–1472, 2000.

[94] K. Strand, E. Søreide, S. Aardal, and H. Flaatten, "A comparison of SAPS II and SAPS 3 in a Norwegian intensive care unit population," *Acta Anaesthesiologica Scandinavica*, vol. 53, no. 5, pp. 595–600, 2009.

[95] P. G. H. Metnitz, A. Valentin, H. Vesely, C. Alberti, T. Lang, K. Lenz, H. Steltzer, and M. Hiesmayr, "Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study," *Intensive Care Medicine*, vol. 25, pp. 192–197, Feb. 1999.

[96] G. Apolone, G. Bertolini, R. D'Amico, G. Iapichino, A. Cattaneo, G. De Salvo, and R. M. Melotti, "The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: Results from GiViTI," *Intensive Care Medicine*, vol. 22, pp. 1368–1378, Dec. 1996.

[97] M. Soares and J. I. Salluh, "Validation of the saps 3 admission prognostic model in patients with cancer in need of intensive care," *Intensive Care Medicine*, vol. 32, no. 11, pp. 1839–1844, 2006.

[98] M. Capuzzo, A. Scaramuzza, B. Vaccarini, G. Gilli, S. Zannoli, L. Farabegoli, G. Felisatti, E. Davanzo, and R. Alvisi, "Validation of SAPS 3 admission score and comparison with SAPS II," *Acta Anaesthesiologica Scandinavica*, vol. 53, no. 5, pp. 589–594, 2009.

[99] S. Y. Lim, C. R. Ham, S. Y. Park, S. Kim, M. R. Park, K. Jeon, S.-W. Um, M. P. Chung, H. Kim, O. J. Kwon, *et al.*, "Validation of the Simplified Acute Physiology Score 3 scoring system in a Korean intensive care unit," *Yonsei Medical Journal*, vol. 52, no. 1, pp. 59–64, 2011.

[100] D. Poole, C. Rossi, A. Anghileri, M. Giardino, N. Latronico, D. Radrizzani, M. Langer, and G. Bertolini, "External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units," *Intensive Care Medicine*, vol. 35, no. 11, pp. 1916–1924, 2009.

[101] B. Metnitz, E. Schaden, R. Moreno, J.-R. Le Gall, P. Bauer, and P. G. H. Metnitz, "Austrian validation and customization of the SAPS 3 Admission Score," *Intensive Care Medicine*, vol. 35, pp. 616–22, Apr. 2009.

[102] B. Khwannimit and R. Bhurayanontachai, "The performance and customization of SAPS 3 admission score in a Thai medical intensive care unit," *Intensive Care Medicine*, vol. 36, pp. 342–6, Feb. 2010.

[103] J. M. Silva Junior, L. M. S. Malbouisson, H. L. Nuevo, L. G. T. Barbosa, L. Y. Marubayashi, I. C. Teixeira, A. P. Nassar Junior, M. J. C. Carmona, I. F. da Silva, A. Júnior, *et al.*, "Applicability of the simplified acute physiology score (SAPS 3) in Brazilian hospitals," *Revista Brasileira de Anestesiologia*, vol. 60, no. 1, pp. 20–31, 2010.

[104] M. Rue, A. Artigas, M. Alvarez, S. Quintana, and C. Valero, "Performance of the mortality probability models in assessing severity of illness during the first week in the intensive care unit," *Critical Care Medicine*, vol. 28, pp. 2819–2824, AUG 2000 2000. PT: J.

[105] P. A. Patel and B. J. B. Grant, "Application of mortality prediction systems to individual intensive care units," *Intensive Care Medicine*, vol. 25, pp. 977–982, SEP 1999 1999. PT: J.

[106] C. Hug and G. D. Clifford, "An analysis of the errors in recorded heart rate and blood pressure in the icu using a complex set of signal quality metrics," in *Computers in Cardiology, 2007*, pp. 641–644, IEEE, 2007.

[107] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[108] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012," *Computing in Cardiology*, vol. 39, pp. 245–248, 2012.

[109] A. E. W. Johnson, A. A. Kramer, and G. D. Clifford, "Pre-processing methods for prognostic models," in *Neural Information Processing Systems: Workshop on Machine Learning for Clinical Data Analysis and Healthcare*, 2012.

[110] A. E. Johnson, A. A. Kramer, and G. D. Clifford, "Data preprocessing and mortality prediction: the physionet/cinc 2012 challenge revisited," in *Computing in Cardiology Conference (CinC)*, 2014.

[111] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211–252, 1964.

[112] J. P. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, no. 1, pp. 561–584, 1995.

[113] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009.

[114] R. P. Dellinger, M. M. Levy, J. M. Carlet, J. Bion, M. M. Parker, R. Jaeschke, K. Reinhart, D. C. Angus, C. Brun-Buisson, R. Beale, *et al.*, "Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008," *Intensive Care Medicine*, vol. 34, no. 1, pp. 17–60, 2008.

[115] A. W. Haider, M. G. Larson, S. S. Franklin, and D. Levy, "Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the framingham heart study," *Annals of Internal Medicine*, vol. 138, no. 1, pp. 10–16, 2003.

[116] V. N. Vapnik and S. Kotz, *Estimation of dependences based on empirical data*, vol. 41. Springer-Verlag New York, 1982.

[117] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998.

[118] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[119] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[120] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, vol. 338, 2009.

[121] P. G. Bastos, X. Sun, D. P. Wagner, A. W. Wu, and W. A. Knaus, "Glasgow Coma Scale score in the evaluation of outcome in the intensive care unit: findings from the Acute Physiology and Chronic Health Evaluation III study," *Critical Care Medicine*, vol. 21, no. 10, pp. 1459–1465, 1993.

[122] T. G. Buchman, "Nonlinear dynamics, complex systems, and the pathobiology of critical illness," *Current Opinion in Critical Care*, vol. 10, no. 5, pp. 378–382, 2004.

[123] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," *Advances in neural information processing systems*, vol. 16, no. 1, pp. 49–56, 2004.

[124] J. R. Le Gall, S. Lemeshow, and F. Saulnier, "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study," *JAMA*, vol. 270, no. 24, pp. 2957–63, 1993.

[125] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. H. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database," *Critical Care Medicine*, vol. 39, no. 5, p. 952, 2011.

[126] A. A. Kramer, D. Shumate, and M. Stark, "White Paper Report," tech. rep., Cerner, 2005.

[127] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[128] J.-L. Vincent, S. M. Opal, J. C. Marshall, and K. J. Tracey, "Sepsis definitions: time for change," *Lancet*, vol. 381, no. 9868, pp. 774–775, 2013.

[129] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[130] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *The Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.

[131] N. Barricelli, "Symbiogenetic evolution processes realized by artificial methods," *Methodos*, pp. 143–182, 1957.

[132] H. JH, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, 1975.

[133] M. M. Davis LD, *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, 1991.

[134] J. Kennedy and R. Everhart, "Particle Swarm Optimization," *IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948, 1995.

[135] I. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Information Processing Letters*, vol. 85, pp. 317–325, Mar. 2003.

[136] K. G. Moons, D. G. Altman, Y. Vergouwe, and P. Royston, "Prognosis and prognostic research: application and impact of prognostic models in clinical practice," *BMJ*, vol. 338, 2009.

[137] J. E. Zimmerman and A. A. Kramer, "Outcome prediction in critical care: the Acute Physiology and Chronic Health Evaluation models," *Current Opinion in Critical Care*, vol. 14, pp. 491–497, 2008.

[138] A. E. W. Johnson, A. A. Kramer, and G. D. Clifford, "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy," *Critical Care Medicine*, vol. 41, no. 7, pp. 1711–1718, 2013.

[139] L. Mayaud, P. S. Lai, G. D. Clifford, L. Tarassenko, L. A. G. Celi, and D. Annane, "Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension," *Critical Care Medicine*, vol. 41, no. 4, p. 954, 2013.

[140] L. Mayaud, *Prediction of mortality in septic patients with hypotension*. PhD thesis, University of Oxford, Oxford, United Kingdom, 2014.

[141] I. Neamatullah, M. M. Douglass, L. wei H Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford, "Automated de-identification of free-text medical records," *BMC medical informatics and decision making*, vol. 8, no. 1, p. 32, 2008.

[142] C. N. Sessler, M. S. Gosnell, M. J. Grap, G. M. Brophy, P. V. O'Neal, K. A. Keane, E. P. Tesoro, and R. Elswick, "The richmond agitation–sedation scale: validity and reliability in adult intensive care unit patients," *American journal of respiratory and critical care medicine*, vol. 166, no. 10, pp. 1338–1344, 2002.

[143] W. A. Knaus, "APACHE 1978-2001: The Development of a Quality Assurance System Based on Prognosis: Milestones and Personal Reflections," *Arch Surg*, vol. 137, no. 1, pp. 37–41, 2002.

[144] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical care*, vol. 36, no. 1, pp. 8–27, 1998.

[145] A. A. Kramer, "Predictive mortality models are not like fine wine," *Critical Care*, vol. 9, pp. 636–7, Jan. 2005.

[146] B. W. Silverman, *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[147] L. A. G. Celi, R. J. Tang, M. C. Villarroel, G. A. Davidzon, W. T. Lester, and H. C. Chueh, "A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury," *Journal of healthcare engineering*, vol. 2, no. 1, pp. 97–110, 2011.

[148] L. wei H Lehman, R. Adams, L. Mayaud, G. Moody, A. Malhotra, R. Mark, and S. Nemati, "A physiological time series dynamics-based approach to patient monitoring and outcome prediction," *Computing in Cardiology*, 2014.

[149] D. Teres and S. Lemeshow, "As American as apple pie and APACHE," *Critical Care Medicine*, vol. 26, no. 8, pp. 1297–1298, 1998.

[150] K. G. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman, "Prognosis and prognostic research: what, why, and how?," *BMJ*, vol. 338, 2009.

[151] D. A. Harrison, A. R. Brady, and K. Rowan, "Case mix, outcome and length of stay for admissions to adult, general critical care units in england, wales and northern ireland: the intensive care national audit & research centre case mix programme database," *Critical Care*, vol. 9, no. Suppl 3, p. S1, 2004.

[152] J. D. Young, C. Goldfrad, and K. Rowan, "Development and testing of a hierarchical method to code the reason for admission to intensive care units: the ICNARC Coding Method," *British Journal of Anaesthesia*, vol. 87, no. 4, pp. 543–548, 2001.

[153] M. W. Kuzniewicz, E. E. Vasilevskis, R. Lane, M. L. Dean, N. G. Trivedi, D. J. Rennie, T. Clay, P. L. Kotler, and R. A. Dudley, "Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders," *Chest*, vol. 133, pp. 1319–27, June 2008.

[154] C. E. Rasmussen, *Gaussian processes for machine learning*. Citeseer, 2006.

[155] D. A. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko, "An extreme function theory for novelty detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 1, pp. 28–37, 2013.

[156] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, pp. 2546–2554, 2011.

[157] J. S. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.

[158] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, pp. 2951–2959, 2012.

[159] L. A. Celi, E. Hassan, C. Marquardt, M. Breslow, and B. Rosenfeld, "The eICU: it's not just telemedicine," *Critical Care Medicine*, vol. 29, no. 8, pp. N183–N189, 2001.

[160] A. E. W. Johnson, N. Dunkley, L. Mayaud, A. Tsanas, A. A. Kramer, and G. D. Clifford, "Patient Specific Predictions in the Intensive Care Unit Using a Bayesian Ensemble," *Computing in Cardiology*, vol. 39, pp. 249–252, 2012.

[161] L. Citi and R. Barbieri, "Physionet 2012 Challenge: Predicting mortality of ICu patients using a cascaded SVM-GLM paradigm," *Computing in Cardiology*, vol. 39, pp. 257–260, 2012.

[162] S. Vairavan, L. Eshelman, S. Haider, A. Flower, and A. Seiver, "Prediction of Mortality in an Intensive Care Unit using Logistic Regression and a Hidden Markov Model," *Computing in Cardiology*, vol. 39, pp. 393–396, 2012.

[163] M. Macas, J. Kuzilek, T. Odstrcilik, and M. Huptych, "Linear Bayes Classification for Mortality Prediction," *Computing in Cardiology*, vol. 39, pp. 473–476, 2012.

[164] H. Xia, B. J. Daley, A. Petrie, and X. Zhao, "A Neural Network Model for Mortality Prediction in ICU," *Computing in Cardiology*, vol. 39, pp. 261–264, 2012.

[165] S. L. Hamilton and J. R. Hamilton, "Predicting in-hospital death and mortality percentage using logistic regression," *Computing in Cardiology*, vol. 39, pp. 489–492, 2012.

[166] C. H. Lee, N. M. Arzeno, J. C. Ho, H. Vikalo, and J. Ghosh, "An Imputation-Enhanced Algorithm for ICU Mortality Prediction," *Computing in Cardiology*, vol. 39, pp. 253–256, 2012.

[167] S. McMillan, C.-C. Chia, A. V. Esbroeck, I. Rubinfeld, and Z. Syed, "ICU Mortality Prediction using Time Series Motifs," *Computing in Cardiology*, vol. 39, pp. 265–268, 2012.

[168] D. Bera and M. M. Nayak, "Mortality risk assessment for ICU patients using logistic regression," *Computing in Cardiology*, vol. 39, pp. 493–496, 2012.

[169] M. Kayaalp, "ICU outcome predictions using physiologic trends in the first two days," *Computing in Cardiology*, vol. 39, p. 977, 2012.

[170] A. Bosnjak and G. Montilla, "Predicting mortality of ICU patients using statistics of physiological variables and support vector machines," *Computing in Cardiology*, vol. 39, pp. 481–484, 2012.

[171] T. J. Pollard, L. Harra, D. Williams, S. Harris, D. Martinez, and K. Fong, "2012 PhysioNet Challenge: An Artificial Neural Network to Predict Mortality in ICU Patients and Application of Solar Physics Analysis Methods," *Computing in Cardiology*, vol. 39, pp. 485–488, 2012.

[172] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference Committee.," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[173] L.-w. H. Lehman, S. Nemati, R. P. Adams, and R. G. Mark, "Discovering shared dynamics in physiological signals: Application to patient monitoring in icu," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 5939–5942, IEEE, 2012.

[174] R. Dybowski, P. Weller, R. Chang, and V. Gant, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *Lancet*, vol. 347, pp. 1146–50, Apr. 1996.

[175] G. Clermont, D. Angus, S. DiRusso, M. Griffin, and W. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models," *Critical Care Medicine*, vol. 29, no. 2, pp. 291–296, 2001.

[176] G. Doig, K. Inman, W. Sibbald, C. Martin, and J. Robertson, "Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression.," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, p. 361, American Medical Informatics Association, 1993.

[177] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. Moons, "Prognosis and prognostic research: validating a prognostic model," *BMJ*, vol. 338, 2009.

[178] S. Finazzi, D. Poole, D. Luciani, P. E. Cogo, and G. Bertolini, "Calibration belt for quality-of-care assessment based on dichotomous outcomes," *PloS one*, vol. 6, no. 2, p. e16110, 2011.

[179] D. K. McClish, "Analyzing a portion of the ROC curve," *Medical Decision Making*, vol. 9, no. 3, pp. 190–195, 1989.

[180] N. A. Halpern, S. M. Pastores, and R. J. Greenstein, "Critical Care Medicine in the United States 1985–2000: An analysis of bed numbers, use, and costs," *Critical Care Medicine*, vol. 32, no. 6, pp. 1254–1259, 2004.

[181] S. Ridley and S. Morris, "Cost effectiveness of adult intensive care in the UK," *Anaesthesia*, vol. 62, no. 6, pp. 547–554, 2007.

[182] N. Hawkes, "Royal college recommends national system to recognise deteriorating patients," *BMJ*, vol. 345, p. e5041, 2012.

Appendix A

The PhysioNet/Computing in Cardiology 2012 Challenge for predicting mortality

And now for something completely different.

Monty Python

The severity scores described in Chapter 1 are all primarily used for adjustment of patient severity across cohorts. In cooperation with the annual Computing in Cardiology conference, PhysioNet hosts an annual challenge which aims to spur interest in difficult clinically relevant problems. The Physionet/Computing in Cardiology 2012 Challenge (Challenge) aimed to improve patient-specific predictions of in-hospital mortality [108].

The hypothesis tested was that the increased granularity of the information available in the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II database[1] would allow for more accurate predictions applicable to a single patient at the bedside. For example, while severity scores such as the APS rely on a single extreme value to represent a patient's heart rate, the Challenge database contains all heart rates measured for the first 48 hours.

[1]See Chapter 5 for a description of MIMIC II, the database which was the source of the Challenge data.

The Challenge database is split into three sets: set a (PN_a), set b (PN_b) and set c (PN_c). Each set contains 4,000 records, and each record describes a single patient's first 48 hours in the ICU. All patients were admitted to an ICU at the Beth Israel Deaconess Medical Center in Boston, USA between 2001 and 2006. These records were released as anonymised text files. Each row of the text file contained a measurement description (e.g. heart rate), the time of the measurement in minutes since admission, and the value of the measurement. All records strictly ended at 48 hours post ICU admission, and consequently all patients survived to 48 hours.

After data preparation, the organisers released PN_a and PN_b to the public. A single additional text file which contained detailed outcomes of each patient record was released only for PN_a. Outcomes included a binary indicator of hospital mortality, length of stay, date of death (if applicable), and two severity scores. Outcomes were not made available for PN_b. The lack of outcome information for PN_b prohibited its use as a part of the development dataset. However, it still allowed for validating that the code for a model was working correctly on test data. Finally, neither data nor outcomes were made available for PN_c, and this set was used for the final evaluation in the competition. The competition lasted for a finite period of time (January 2012 to August 2012) and consequently results on PN_c were only calculated for certain models developed during this time period. Furthermore, as the challenge has ended, it was not possible to correct all errors detected at a later point in time.

A.1 Scoring

The two official scoring metrics for the Challenge are referred to as Score 1 (s_1) and Score 2 (s_2). The first, s_1, aimed to quantify the discrimination and utility of the prediction model simultaneously. The s_1 is calculated as the minimum of the Sensitivity (Se) and Positive Predictive Value (PPV) and is defined as follows:

$$s_1 = min(Se, PPV) = min(\frac{TP}{TP + FN}, \frac{TP}{TP + FP}), \tag{A.1}$$

where TP indicates a true positive, FP indicates a false positive and FN indicates a

false negative.

s_1 quantifies the reliability of a classifier for positive outcomes (which indicate patient death). For example, if an algorithm obtained $s_1 = 0.5$, this would indicate that the algorithm correctly classifies non-surviving patients as moribund with probability 0.5 and incorrectly classifies survivors as moribund with probability 0.5.

s_2, in contrast, aimed to quantify the calibration of the model and reflect how accurately the predictions matched the overall prevalence of patient mortality. s_2 is a range-scaled version of the $HL_{\hat{C}}$ from Equation 1.1 and is calculated as follows:

$$s_2 = \frac{1}{p_D - p_1} \sum_{j=1}^{D} \frac{(O_j - E_j)^2}{n_j p_j (1 - p_j) + 0.001} = \frac{1}{p_D - p_1} HL_{\hat{C}}. \tag{A.2}$$

Equation A.2 shows the calculation of the statistic, where p_j is the estimated probability of mortality in the jth decile (equivalent to $\frac{E_j}{n_j}$, where n_D is the number of observations in decile j), p_D is the probability of mortality in the 10th (Dth) decile, and p_1 is the probability of mortality in the 1st decile.

Scores (both s_1 and s_2) on $\mathrm{PN_b}$ were provided at two stages during the Challenge. The final ranking was determined by s_1 and s_2 by the organisers using predictions on $\mathrm{PN_c}$. In order to compete, a user had to submit code which the Challenge organizers could evaluate on their own. While the competition has ended, the Challenge organizers continue to allow submissions and provide performance metrics on $\mathrm{PN_b}$. Thus, while the official Challenge performance metrics are available on $\mathrm{PN_c}$ (as these models were submitted during the Challenge), most performance metrics are only available on $\mathrm{PN_a}$ or $\mathrm{PN_b}$.

A.2 Bayesian ensemble of additive sigmoidal trees

A novel tree based classifier referred to as a Bayesian Ensemble of Additive Sigmoidal Trees (BEAST) was utilised for predicting mortality on the Challenge database. This model was originally proposed and developed by Nic Dunkley Hutchinson and has been previously described [140,160]. During the empirical study presented here, the author of this thesis contributed to the i) optimisation of the algorithm in terms of speed, ii) dis-

cussions regarding the level of complexity of the trees and iii) empirical development and evaluation of the algorithm. The BEAST has many advantages, including high overall performance, invariance to data scale and automatic handling of missing data. Each tree selects a subset of observations using a depth two decision tree. All observations corresponding to a single node are assigned a contribution proportional to the observation's value plus an intercept. Furthermore, the tree also assigns a separate contribution to missing values for these observations. Many trees with the same structure are created, and the contributions across all trees are summed to provide the contribution for a single "forest" (plus an intercept term based on the prevalence of the training data outcome). During model development, multiple forests are generated, and the final output predictions involve summing each forests' contribution and transforming this value using the inverse logit, shown the following equation:

$$p = \frac{1}{1 + e^{-\sum c + \beta}} \tag{A.3}$$

where p is the final probability, c is the contribution for each forest and *beta* is an intercept term. The weights and features selected are determined using a custom Markov chain Monte Carlo (MCMC) sampler.

A.2.1 Data normalisation

Before learning the model, each feature in the data is converted into a percentile representing its relative order in the data, as follows:

$$Z_{k,j} = P(\mathbf{X}_j < X_{i,j}) \leq \frac{k}{N}, \tag{A.4}$$

where N is the number of rows in the design matrix $\mathbf{X}$. Given the above definition, it follows that $\{\mathbf{Z} : \mathbf{Z} \in \mathbb{R} \wedge \mathbf{Z} \geq 0 \wedge \mathbf{Z} \leq 1\}$. These percentiles are then treated as the probabilities of the cumulative density function of a standard normal distribution; that is we update $\mathbf{Z}$ as follows:

$$Z_{i,j} = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{Z_{i,j}}{\sqrt{2}}\right)\right), \tag{A.5}$$

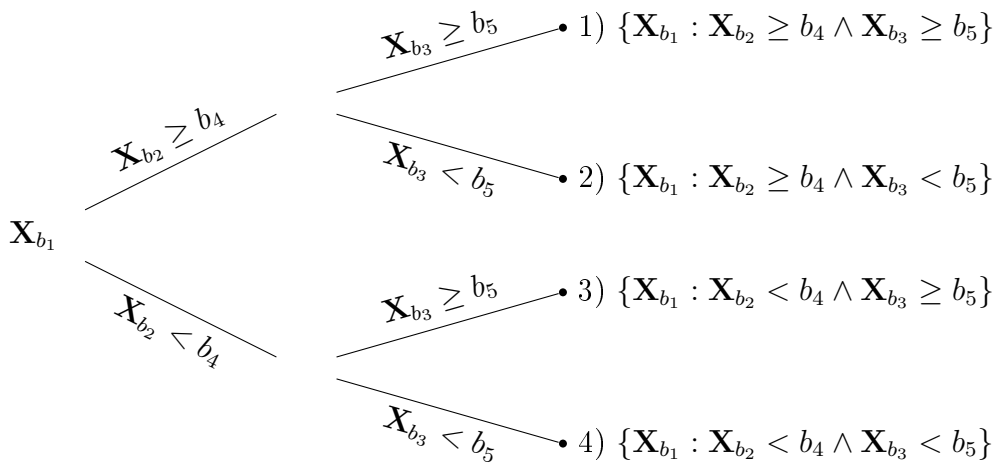

Figure A.1: *Diagram of a single tree in the BEAST. The result is four leaf nodes, each of which contains a subset of the design matrix* $\mathbf{X}$ *and each subset is disjoint. For example, node 1 contains the values of feature* b_1 *where the values of* $\mathbf{X}_{b_2}$ *are greater than or equal to some threshold* b_4 *and the values of* $\mathbf{X}_{b_3}$ *are greater than or equal to some threshold* b_5*.*

where we adopt the standard formula for the error function; $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \text{d}t$. This formula is equivalent to the inverse cumulative density of the value $Z_{i,j}$ given a normal distribution $\mathcal{N} \sim (0, 1)$.

This approach distorts the scale of the input feature (e.g. logarithmically distributed values become normally distributed), but provides robustness to outliers as extreme values are shifted closer to the mean of the distribution.

A.2.2 Tree structure

Each tree performs two tasks: it randomly selects a subset of observations (patients), generates a feature dependent contribution for that subset and generates a unique contribution for missing values in that subset. These two tasks will be addressed in turn.

First, let us define three parameters of the tree: b_1, b_2 and b_3. These three parameters index a specific feature in the design matrix $\mathbf{X}$, that is $b_1 \in 1, 2, 3, ..., D$, $b_2 \in 1, 2, 3, ..., D$ and $b_3 \in 1, 2, 3, ..., D$ where D is the number of features. Second, let us define two more parameters of the tree: b_4 and b_5. These parameters define thresholds applied to the feature values to select subsets of the design matrix. Figure A.1 provides a schematic of the tree used to create four leaves which each contain disjoint subsets of the design matrix $\mathbf{X}$.

The depth two tree results in four subsets of the design matrix $\mathbf{X}$. The tree then

disregards three of the four subsets, that is only node $b_{10} \in 1, 2, 3, 4$ is retained, the remaining data is disregarded. Consequently, our tree defines a function $t(\cdot)$ as follows:

$$t(\mathbf{X}_{b_1}) = \begin{cases} 1 \\ b_{10} = 1 \wedge \mathbf{X}_{b_2} \geq b_4 \wedge \mathbf{X}_{b_3} \geq b_5 1 \\ b_{10} = 2 \wedge \mathbf{X}_{b_2} \geq b_4 \wedge \mathbf{X}_{b_3} < b_5 1 \\ b_{10} = 3 \wedge \mathbf{X}_{b_2} < b_4 \wedge \mathbf{X}_{b_3} \geq b_5 1 \\ b_{10} = 4 \wedge \mathbf{X}_{b_2} < b_4 \wedge \mathbf{X}_{b_3} < b_5 0 \quad \text{otherwise,} \end{cases} \tag{A.6}$$

and consequently $t(\mathbf{X}_{b_1})$ returns one for rows to be assigned a contribution and zero otherwise.

Given the splitting process, the tree now considers data for a single feature $\mathbf{X}_{b_1}$ and a subset of rows. The contribution for these rows is then calculated as $b_6 x + b_{11}$. It is worth noting that b_6 and b_{11} are *not* directly calculated using any supervised method, they are randomly generated and the entire tree is later accepted or rejected (further detailed later).

Each observation for node b_{10} has now been assigned a "contribution", i.e. the tree has made a prediction for these observations. If there are no missing values in the feature $\mathbf{X}_{b_1}$, this completes the calculation of predictions for a single tree in the BEAST algorithm. However, if there are missing values in the data $\mathbf{X}$, two additional steps of the algorithm ensue. First, a single value b_7 is sampled from the uniform distribution; $b_7 \, \mathcal{U}(0, 1)$. If the proportion of missing values in feature b_1 is less than b_7, no missing value imputation occurs. If the proportion is larger, then the $b_8{}^{\text{th}}$ value of the transformed data $\mathbf{Z}$ is extracted and the contribution is equal to $b_7 Z_{b_8, b_1} + b_{11}$. The goal of this method is to allow the algorithm to learn impact of a missing value for each feature on the final prediction, rather than always defaulting missing values to "normal" or some other fixed value.

To summarise, the contribution for each row of design matrix $\mathbf{X}$ by a single tree is

given as follows:

$$c(X_{i,b_1}) = \begin{cases} b_6 Z_{i,b_1} + b_{11} & t(\mathbf{X}_{b_1}) = 1 \wedge X_{i,b_1} \in \mathbb{R} \\ b_7 Z_{b_8,b_1} + b_{11} & t(\mathbf{X}_{b_1}) = 1 \wedge X_{i,b_1} \notin \mathbb{R} \wedge b_7 > \frac{||\mathbf{X}_{b_1}||^0}{N} \\ 0 & t(\mathbf{X}_{b_1}) = 1 \wedge X_{i,b_1} \notin \mathbb{R} \wedge b_7 \leq \frac{||\mathbf{X}_{b_1}||^0}{N} \\ 0 & \text{otherwise.} \end{cases} \tag{A.7}$$

where we define $||\cdot||^0$ as the L^0 norm (the number of elements in a vector) and only rows i which have been selected by the tree splitting function $t(\mathbf{X}_{b_1})$ are given a contribution. Each tree is represented by a total of eleven parameters: $b_1, b_2, ..., b_{11}$.

A.2.3 Forest structure

A forest in the context of the BEAST consists of N_t trees which assign contributions to design matrix $\mathbf{X}$ plus a single parameter set at model initialisation to centre the contributions about the mean outcome occurrence. This allows the individual trees to generate contributions without being improperly penalized by a lack of calibration. This offset, β_0, is calculated as follows:

$$\beta_0 = \log\left(\frac{1}{-1 + \left(\frac{\sum_{i=1}^{N} y_i}{N}\right)^{-1}}\right), \tag{A.8}$$

where $\mathbf{y}$ is the vector of binary outcomes. An additional parameter, β_1, controls the width of the forest and is used to scale the summed contributions from all trees in the forest. β_1 is defined as:

$$\beta_1 = \frac{2\sqrt{\text{var}(\frac{1}{1+e^{\hat{\mathbf{p}}}})}}{N_t} \tag{A.9}$$

where $\hat{\mathbf{p}}$ are predictions on the same dataset using a regularised logistic regression (see Section 2.3.1). The output of a forest $\hat{\mathbf{y}}$ given the design matrix $\mathbf{X}$ is calculated as:

$$\mathcal{F}(\mathbf{X}) = \hat{\mathbf{y}} = \frac{1}{1 + e^{-\beta_0 + \beta_1 \sum_{k=1}^{K} c_k}} \tag{A.10}$$

where $\mathcal{F}$ represents the forest and $\mathbf{c}_k$ is the contribution for the k^{th} tree as calculated in Equation $A.7$.

A.2.4 Initialisation

The forest is initialised using three parameters: N_t, N_r, N_s and N_i. N_t is the number of trees, N_r is the number of Markov Chain Monte Carlo (MCMC) resets (more detail later) and N_i is the number of iterations in the MCMC. N_s defines how the interval, in iterations, at which the forest is saved (the final algorithm applies all saved forests to the design matrix $\mathbf{X}$ to attain the final prediction). The parameters of the N_t trees are randomly generated using distributions defined in Table A.1. Note that the features are selected using a uniform distribution inversely weighted by the frequency of each feature currently contained in the forest, preventing the forest from converging on a single predictive feature and encouraging searching of the feature space.

Parameter	**Purpose**	**Update Scheme**
b_1	Feature used for 1^{st} split	Weighted Uniform distribution, [1,D]
b_2	Feature used for 2^{nd} split	Weighted Uniform distribution, [1,D]
b_3	Feature used to calculate contribution	Weighted Uniform distribution, [1,D]
b_4	Percentile used to segment data, (1^{st} split)	Uniform distribution, [0,1]
b_5	Percentile used to segment data, (2^{nd} split)	Uniform distribution, [0,1]
b_6	Contribution slope	$\mathcal{N}(0, W)$
b_7	Fraction of MVs required before surrogate values are used (1^{st} split)	Uniform distribution, [0,1]
b_8	Fraction of MVs required before surrogate values are used (2^{nd} split)	Uniform distribution, [0,1]
b_9	Feature value used as surrogate for MV	Uniform distribution [1-4]
b_{10}	Which node in the tree is used	Uniform distribution [1-4]
b_{11}	Contribution intercept	$\mathcal{N}(0, W)$

Table A.1: *The parameters used in each tree and the method in which they are updated during optimization.* MV = *missing value.*

A.2.5 Updating

The optimisation of the forest proceeds by updating exactly two trees per iteration. The goal of this process is to generate a completely new tree which results in the overall forest having a higher likelihood of generating the outcomes $\mathbf{y}$. This update is performed in the same manner for every iteration, though the trees affected vary. The update process regenerated all the parameters of the two trees affected and evaluates the likelihood of the forest's predictions using the newly generated trees versus the likelihood of the forest's predictions using the unmodified trees. That is, we aim to maximise the likelihood of the outcomes given our forest $\mathcal{F}$:

$$\underset{\mathcal{F}}{\text{argmax}}\ \mathcal{L}(y|\mathcal{F}(\mathbf{X})). \tag{A.11}$$

As we cannot estimate the distribution $P(\mathbf{y};\mathcal{F},\mathbf{X})$ directly we use sampling techniques, more specifically MCMC with a Metropolis-Hastings (MH) acceptance step. Each iteration of the MCMC updates the forest as described earlier and calculates the likelihood of the outcomes given the data, that is:

$$\mathcal{L}(\mathbf{y}|\mathcal{F},\mathbf{X}) = \mathcal{L}(\mathbf{y}|\hat{\mathbf{y}}) \tag{A.12}$$

where $\hat{\mathbf{y}} = \mathcal{F}(\mathbf{X})$. The MH step then involves calculating how much more likely the current forest is compared to the previous forest. If we define $\mathcal{F}^{(k)}$ as the current forest and $\mathcal{F}^{(k-1)}$ as the previous forest, then we first calculate:

$$\frac{\mathcal{L}(\mathbf{y}|\mathcal{F}^{(k)})}{\mathcal{L}(\mathbf{y}|\mathcal{F}^{(k-1)})} = \frac{\log\ \mathcal{L}(\mathbf{y}|\mathcal{F}^{(k)})}{\log\ \mathcal{L}(\mathbf{y}|\mathcal{F}^{(k-1)})} = \frac{\sum_{i=1}^{N}\Big(y_i\log(\hat{y}_i^{(k)}) + (1-y_i)\log(1-\hat{y}_i^{(k)})\Big)}{\sum_{i=1}^{N}\Big(y_i\log(\hat{y}_i^{(k-1)}) + (1-y_i)\log(1-\hat{y}_i^{(k-1)})\Big)}, \tag{A.13}$$

and the MH step involves accepting the new forest with probability defined by Equation A.13. Here "acceptance" refers to retaining the modifications done to the forest at iteration k. Explicitly we can define the probability of acceptance p as:

$$p = e^{-\left(\log\ \mathcal{L}(\mathbf{y}|\mathcal{F}^{(k)}) - \log\ \mathcal{L}(\mathbf{y}|\mathcal{F}^{(k-1)})\right)} \tag{A.14}$$

If the likelihood of the new forest is much higher than the old forest, it is very likely to be accepted and conversely if the likelihood of the new forest is much lower it is unlikely to be accepted.

If the update is accepted, the two trees are kept in the forest, otherwise they are discarded and the forest remains unchanged. After completion of 20% of the iterations, a set burn-in period which allows stabilisation to a forest of reasonable likelihood, the algorithm begins storing forests at a fixed interval (every N_s iterations). Once the number of user-defined iterations are reached (N_i), the forest is re-initialized as before (up to N_r times), and the iterative process restarts. Again after the set burn-in period, the forests begin to be saved again at a fixed interval. The final result of this algorithm is a set of forests, each of which will contribute to the final model prediction.

Note that N_t, N_r, N_s and N_i are hyperparameters which must be set prior to model development. In this work $N_t = 500$, $N_r = 4$, $N_i = 100000$ and $N_s = 2000$.

A.3 Overview of model development

Data from PN_a concerning 4,000 patients was first preprocessed using Domain Knowledge (DK) preprocessing. This preprocessing method is described in Section A.3.1. After data preprocessing a BEAST was used to predict whether a patient died within their hospital stay. Preliminary assessment of model performance involved randomly splitting PN_a into training sets of size 3,000 and test sets of size 1,000. The performance across 32 model developments is reported, allowing for an estimate of the generalization performance. For the Challenge submission, the model was redeveloped using all 4,000 observations in PN_a. All evaluation metrics are available for PN_b. As this model was submitted during the competition, s_1 and s_2 are available for PN_c.

A.3.1 Preprocessing

The data used for the Challenge was preprocessed primarily using arbitrary rules which removed unphysiological values (such as heart rates above 300) or artefactual values (many measurements are recorded as 0 instead of missing). This preprocessing method is referred to as DK preprocessing. DK consisted of two parts: transformations accounting

for known transcription errors and physiologic constraints which could be reasonably applied to the data. The transformations, if known, were applied to the data to shift values from obviously impossible values back to the presumed correct value. For example, temperatures which were in the range of 90-110 were assumed to be erroneously recorded in degrees Farenheit, and the appropriate conversion was applied to return them to degrees Centigrade. Another example is order of magnitude errors. The physiologic range of pH is, conservatively, between 6.5-8. Values which occur between 65-80 are likely to be transcription errors which have shifted the values by an order of magnitude. Finally, the transformation stage also involved the replacement of numeric values which are used to indicate missing values (such as -1 for height and weight). A list of the transformations for each variable is shown in Table A.2.

Physiologic thresholds were then applied to the data. Any value outside the threshold range was set to missing. A list of these thresholds are shown in Table A.3.

The transformation steps detailed in Table A.2 and the thresholding steps described in Table A.3 fully comprise the DK preprocessing stage.

A.4 Threshold calculation

As s_1 required binary classifications it was necessary to select a threshold which would be used to round the probabilistic predictions into binary values. Optimally this threshold would maximize $s_1 = min(Se, PPV)$ on the unseen test set, though this optimal cut off is not trivially calculated when only the training set is available. The threshold was determined by first calculating the prevalence of mortality in $\mathrm{PN_a}$, $\frac{1}{N}\sum_{i=1}^{N}\hat{y}_i = 0.1385 = 13.85\%$. The model was then applied to $\mathrm{PN_b}$, and the threshold was selected as $100\% - 13.85\% = 86.15\%$ percentile of these probabilities.

A.5 Challenge benchmark

The benchmark for the Challenge was based on the SAPS described in Section 1.3.2. As SAPS is an integer severity score it is inappropriate to directly apply as s_2 requires probabilistic predictions. To provide the benchmark, SAPS was recalibrated by the

Variable	Inclusion criteria	Transformation	Reasoning
Age	≥ 100	$x \leftarrow 105$	Replace anonymised ages with 105
Gender	$= -1$	$x \leftarrow NaN$	Replace unphysiological values with NaN
Height	$= -1$	$x \leftarrow NaN$	Replace unphysiological values with NaN
	$\geq 0, < 10$	$x \leftarrow 100x$	Correcting transcription error
	$\geq 10, < 25$	$x \leftarrow 10x$	Correcting transcription error
	$\geq 25, < 100$	$x \leftarrow 2.20x$	Correcting conversion error
	$> 250, \leq 1000$	$x \leftarrow 0.4536x$	Correcting conversion error
	> 1000	$x \leftarrow 10x$	Correcting transcription error
Weight	$= -1$	$x \leftarrow NaN$	Replace unphysiological values with NaN
BUN	$= 0$	$x \leftarrow NaN$	Replace unphysiological values with NaN
RespRate	$= 0$	$x \leftarrow NaN$	Replace unphysiological values with NaN
HR	$= 0$	$x \leftarrow NaN$	Replace unphysiological values with NaN
	$= 300$	$x \leftarrow NaN$	Replace unphysiological values with NaN
Temp	$\geq 10, < 20$	$x \leftarrow \frac{9x}{5} + 32$	Reverse incorrect conversion
	$> 0, < 10$	$x \leftarrow \frac{9(\frac{9x}{5}+32)}{5} + 32$	Reverse incorrect double conversion
PaO_2	$= 0$	$x \leftarrow NaN$	Replace unphysiological values with NaN
	$> 0, < 10$	$x \leftarrow 7.5006x$	Correct conversion error
pH	$\geq 0.65, \leq 0.8$	$x \leftarrow 10x$	Correcting transcription error
	$\geq 65, \leq 80$	$x \leftarrow \frac{x}{10}$	Correcting transcription error
	$\geq 650, \leq 800$	$x \leftarrow \frac{x}{100}$	Correcting transcription error
	$> 8.0, < 65$	$x \leftarrow NaN$	Replace unphysiological values with NaN
	$> 80, < 650$	$x \leftarrow NaN$	Replace unphysiological values with NaN

Table A.2: *List of the transformations applied to the data. The inclusion criteria indicate the range of values which were used in the transformation. A value of NaN stands for not a number, and this value is used to represent missing data.*

Variable	Rejection rules
DiasABP	≤ 0 or > 200
MAP	< 1
NIDiasABP	< 1
NIMAP	< 1
NISysABP	< 1
PaCO2	< 1
SysABP	< 1
Temp	≤ 0
WBC	≤ 1
Weight	< 35

Table A.3: *List of the thresholds applied to the data. Values outside these limits were deemed unphysiological and removed.*

authors using a univariate logistic regression across all of PN_a and evaluated on PN_b and PN_c [108]. This involved two steps. First, SAPS was calculated on the design matrix from PN_a: $\mathbf{X}$. This vector of scores is denoted by $\mathbf{a}$. Second, a single independent variable (SAPS, $\mathbf{a}$) was used as the input to a Logistic Regression (LR) to predict in-hospital mortality (the dependent variable). That is, we define the following function:

$$g(a) = \frac{1}{1 - e^{-\beta_0 - \beta_1 s}}, \tag{A.15}$$

where β_0 and β_1 are two coefficients to be maximised and a is the SAPS for a single record (i.e. single row vector in $\mathbf{X}$). Coefficients β_0 and β_1 are selected to maximise the following:

$$\log(\mathcal{L}(g(\mathbf{a}); \mathbf{y})) = \sum_{i=1}^{N} \left(y_i \times \log(g(\mathbf{a})) + (1 - y_i) \times \log(1 - g(\mathbf{a}))\right), \tag{A.16}$$

where $\mathbf{y}$ is in-hospital mortality. The threshold which maximises the s_1 on the training set is selected to convert the probabilistic predictions to binary predictions. Note that as this model development is performed once on PN_a, the value of s_1 and s_2 will be optimistic when evaluated on PN_a (in contrast to BEAST which is repeatedly developed using training subsets and evaluated on validation subsets).

Variable	Candidate values	Observations affected Set a	Set b
Age	≥ 100	0	0
Gender	$= -1$	3	5
Height	$= -1$	2106	2070
	$\geq 0, < 10$	1	1
	$\geq 10, < 25$	8	2
	$\geq 25, < 100$	0	0
	$> 250, \leq 1000$	7	2
	> 1000	0	0
Weight	$= -1$	326	347
BUN	$= 0$	1	1
RespRate	$= 0$	59	40
Heart Rate	$= 0$	11	21
Heart Rate	$= 300$	1	0
Temperature	$> 0, < 10$	24	17
	$\geq 10, < 20$	105	60
PaO_2	$= 0$	1	
	$> 0, < 10$	2	2
pH	$\geq 0.65, \leq 0.8$	0	0
	$\geq 65, \leq 80$	0	0
	$\geq 650, \leq 800$	3	0
	$> 8.0, < 65$	2	1
	$> 80, < 650$	6	1

Table A.4: *List of the number of observations affected by the transcription process for set a (development data) and set b (evaluation data) for the Physionet/CinC 2012 challenge.*

A.6 Results

A.6.1 Domain knowledge preprocessing

Table A.4 shows the number of observations in the training and test set which were modified by the DK preprocessing. A large number of negative values, which are used to represent missing data, were deleted (e.g. 2,106 for height in PN_a). There were a large number of erroneous temperature values in PN_a (105) which were the likely result of an inappropriate conversion from °F to °C applied to data which was already recorded in °C. There were a smaller number of temperature values which underwent this inappropriate conversion twice. There were only 3 order of magnitude errors for pH in PN_a, but none in PN_b. There were 8 pH values in unphysiologic ranges for PN_a and 2 in similar unphysiologic ranges for PN_b. For respiratory rate, heart rate and BUN there were a small number of observations which had been recorded as zero.

Variable	Candidate values	Observations affected Set a	Set b
DiasABP	≤ 0 or > 200	688	550
MAP	< 1	31	34
NIDiasABP	< 1	103	92
NIMAP	< 1	45	30
NISysABP	< 1	1	4
PaCO2	< 1	1	2
SysABP	< 1	685	547
Temp	≤ 0	86	68
WBC	≤ 1	89	90
Weight	< 35	648	503

Table A.5: *List of the number of observations affected by the thresholding process for PN_a (development data) and PN_b (evaluation data) for the Physionet/CinC 2012 challenge.*

Table A.5 shows the number of observations which were affected by the thresholding based preprocessing. The primary erroneous data corrected were zeros.

A.6.2 Model performance

The results on PN_a reported across 32 hold out repetitions are shown in Table A.6. The BEAST had a much higher average AUROC across the repetitions (0.8602) as compared to the benchmark LR which uses SAPS I as the only input (0.6668)[2].

Metric	SAPS I Set A	Model Set A*	95% CI
AUROC	0.6668	0.8602 ($\pm$ 0.0142)	[0.8551, 0.8653]
$\mathcal{I}_{\mathcal{L}}$	0.4023	0.2891 ($\pm$ 0.0165)	[0.2832, 0.2951]
s_1	0.2957	0.4846 ($\pm$ 0.0316)	[0.4732, 0.4960]
s_2	69.001	16.825 ($\pm$ 9.7226)	[13.3196, 20.3304]

Table A.6: *Evaluation metrics of SAPS and the developed model on PN_a (1000 out of sample observations). None of the observations evaluated by the metrics were used in the model development.*
**Statistics on PN_a are presented as the mean and standard deviation from 32 jackknife repetitions.*

Table A.7 shows scores obtained from the final Challenge submission using a BEAST developed with all data available in PN_a. The performance of the model on PN_b was higher than on the PN_a jackknife validation sets in terms of AUROC (0.868 vs. 0.860), $\mathcal{I}_{\mathcal{L}}$ (0.299 vs. 0.402) and s_1 (0.531 vs. 0.484). The s_2 values are not comparable as the

[2]Note that SAPS I only uses the first 24 hours of data.

statistic is sensitive to sample size (See Equation A.2 and Equation 1.1).

Metric	PN_b	PN_c
s_1	0.531002	0.535304
s_2	26.4442	29.8559
AUROC	0.868	-
$\mathcal{I}_{\mathcal{L}}$	0.299	-
SMR	0.912	-
$HL_{\hat{C}}$	20.4	-
B	0.087	-

Table A.7: *Evaluation metrics of the model on PN_b (4000 observations) and PN_c (4000 observations) of the competition. None of the data used in this evaluation were used for model development.*

A.6.3 Comparison to other entries

For reference, Table A.8 displays results of other competitors in the Challenge as measured by s_1 and s_2. In terms of s_1, the BEAST described here outperformed all other methods with a score of 0.5353 [160]. The BEAST also had the fifth best s_2 (29.86). The model by Citi et al. [161] had the highest s_2 (17.88).

A.7 Discussion

Mortality prediction for patients admitted to the ICU has traditionally been performed by logistic regression models, either applied directly or after converting coefficients into an integer score [9, 10, 53]. More recent redevelopments of such models continued the use of logistic regression as the primary prediction model [35, 54–56]. These models also exclusively use data from the first hour of the patient's stay (MPM_0-III [56], SAPS III [54, 55]) or the first 24 hours of the patient's stay (APACHE II-IV [7, 35, 53], SAPS I [8]). Here we have shown that a more complicated model using data from the first 48 hours of a patient's stay strongly outperforms the severity score SAPS I [8]. There are three aspects of the Challenge entry described here which may have led to its good performance as compared to SAPS I and severity scores in general, and these will be discussed in turn.

	PN_b results		PN_c results	
First author and reference	s_1	s_2	s_1	s_2
†Johnson *et al.* [160]	<u>0.5310</u>	26.44	<u>0.5353</u>	29.86
*Citi *et al.* [161]	0.5270	<u>13.24</u>	0.5345	<u>17.88</u>
‡Kang *et al.*				20.58
Vairavan *et al.* [162]	0.50	15.2	0.5009	78.9
Macas *et al.* [163]	0.475	12.820	0.4928	24.70
‡Chidube Ezeozue				24.93
Xia *et al.* [164]	0.5088	82.211	0.4923	
Hamilton *et al.* [165]			0.4872	
Lee *et al.* [166]	0.516	14.4	0.4821	51.69
McMillan *et al.* [167]	0.50	36.63	0.4564	56.45
‡Pantelopoulos *et al.*			0.4544	
Bera *et al.* [168]	0.4436	45.4347	0.4513	45.01
Mehmet Kayaalp [169]			0.39	36.38
Bosnjak *et al.* [170]	0.3504	35.147	0.3333	48.61
Pollard *et al.* [171]	0.24	22.83	0.23	38.23
Sample and random predictors				
SAPS-I (in m-code) [108]			0.3125	68.58
SAPS-I (in C code) [108]			0.3097	35.21
86% randomly predicted to survive			0.1386	10137.7

Table A.8: *The top 10 performances as measured by s_1 reported by the Challenge authors and sorted by s_1 on set c. Corresponding s_2 shown, if available. As scores are only made available if the entry is among the top 10, some are unavailable.*
†This is the BEAST entry which used domain knowledge preprocessing.
** Distinct entry used for set b s_1 and s_2.*
‡No reference available.

A.7.1 Improvement due to model

The first source of improved performance could be attributed to the model itself. In clinical practice, a very low temperature is clinically significant as hypothermia is an indicator of organ failure, while a very high temperature is clinically significant as it is a sign of pyrexia. The ideal model would have the capability of learning this non-monotonic and potentially non-linear mapping of temperature to risk. As the BEAST is a tree based classifier, it has the flexibility for features to have non-monotonic contributions to the final estimate of mortality.

Extending this concept to the multivariate case is exemplified by the systemic inflammation response syndrome (SIRS) [172]. SIRS occurs when the immune system dysfunctions and inflammation, a process that is locally desirable to combat external elements, becomes systemic and threatens the body's homoeostasis. In the worst case, SIRS leads to septicaemia, septic shock and death. SIRS is defined as two or more of the following four criteria: temperature < 36°C or > 38°C, heart rate > 90 bpm, respiratory rate > 20 bpm or $PaCO_2 < 32$ mmHg and the white blood cell count $> 12.10\times10^9$/L or $< 4.10\times10^9$/L or 10% immature band forms [172]. While the definitions of SIRS and sepsis have been since updated [114] and continue to be debated [128], it is clear that illness often presents as a derangement of multiple physiologic parameters in tandem. The ideal model would have the capability of capturing the Gestalt nature of illness: while a temperature < 36°C, a heart rate > 90 bpm, and a respiratory rate > 20 bpm should individually increase the risk of mortality in the model the existence of all three should, using clinical intuition, even further increase the risk. The structure of each tree in the BEAST, which increases or decreases the risk of subsets of patients using similar bivariate thresholds, allows for capturing interactions of this sort. These interactions are not intrinsically captured by the simpler regression models. Nevertheless, attribution of the high performance of the BEAST to the non-linear interactions it may capture is purely speculative.

A final benefit of the model is the learning of a contribution for missing values from the data itself. This is opposed to the standard mean value imputation, which when used in combination with logistic regression, results in missing values being assigned the

average contribution of the feature. As data in the ICU is not missing at random (i.e. the probability of the value being missing *is* dependent on the value) [120] the existence of a missing value may provide additional predictive information and the use of the mean value may distort or destroy this information.

A.7.2 Improvement due to extended time interval

Most severity scores only use data from the first 24 hours and aim to capture the patient physiology on admission. Conversely, the Challenge data spans the first 48 hours of a patient's stay in the ICU. As a result of the longer period of monitoring a more accurate assessment of the patient's physiological state may be attainable. Mayaud *et al.* explored this phenomenon in detail using the MIMIC II database, the source of the Challenge data (Chapter 7, [139]). After excluding all patients who died within the first 72 hours, the authors developed predictive models using 24 hour windows across the first 48 hours. The first window replicated the standard time interval used for severity scores, while the last window consisted of the same features but measured between the 48^{th} and 72^{nd} hour of the patients' ICU stay. The ability to identify non-survivors (specificity) was statistically significantly related to prediction time, but only in patients who died within 17.2 days of ICU admission. They further found a statistically significant increase in AUROC as the window moved from the time of admission to the 48^{th} hour across all patients. As the Challenge data includes the first 48 hours of patient data, whereas SAPS I utilises data from the first 24 hours, it is likely that some performance improvement was due to the extended time interval of monitoring.

A.7.3 Improvement due to additional features

The final potential source of improvement in the model derives from the features extracted. While most severity of illness scores use the worst value across the first ICU day, the entry in this work utilised the first, last, highest, lowest and median value from the data. An example record is shown in Figure A.2 with indication of the features extracted for the Challenge entry and those more commonly extracted by severity of illness scores.

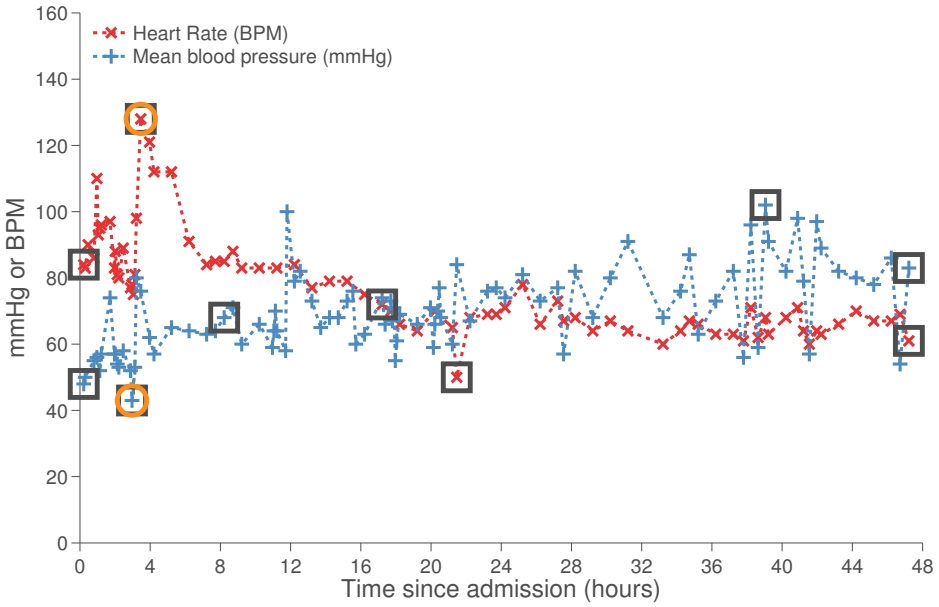

Figure A.2: *Example of the features derived from all measurements for two variables using patient record 132765. Most severity scores use the worst value over the first 24 hours (marked by orange circles), whereas the entry here used the first, last, median, maximum and minimum measurement (marked by gray squares).*

When using only the information available in the worst values, it appears as though patient 132765's prognosis is poor. The combination of a low blood pressure and high heart rate usually indicate the body struggling to adequately perfuse vital organs. However, by the end of the ICU stay, the mean arterial pressure and heart rate have returned to more reasonable values, a likely indicator that the intensive therapy provided has been effective (the volatility in the blood pressure towards the end of the stay is likely due to titration of vasopressors). The inclusion of blood pressure dynamics have been shown to improve the performance of SAPS I on the MIMIC II database [148, 173].

A.8 Use of SAPS as a benchmark

A logistic regression based on SAPS I was utilised as a benchmark for comparison. The benchmark is equivalent to transforming the integer severity scores of SAPS I monotonically into a probability of in-hospital mortality (i.e. there is no difference in discrimination between SAPS I and the benchmark presented, only a difference in calibration). While the model drastically outperforms SAPS I in terms of AUROC, SAPS I was developed in 1984; 17 years prior to the start of data collection in the database [8]. Due

to significant shift in patient demographics and clinical care, it is expected that SAPS I would deteriorate in performance. Furthermore, SAPS I was developed using solely clinical judgement when databases of patient ICU stays were not available. For these reasons, it is perhaps not unexpected to have higher discrimination than SAPS I, and a more relevant benchmark may be the entries of other competitors.

A.9 Complex models in the ICU

One downside of complex machine learning models is the lack of a clinical interpretation. Indeed, this may be a barrier to their acceptance in the clinical community [143]. Logistic regression is very well accepted as it is a simple linear combination of features with an associated ranking of feature importance. While similar feature importance measures can be calculated for most models, their architecture is inevitably less clear and clinicians may distrust a tool they do not comprehend.

Nevertheless, there have been attempts to utilise more sophisticated modelling techniques. In particular Artificial Neural Networks (ANNs), sometimes referred to as connectionist models, have been applied in the past for mortality prediction [174–176]. Both Clermont *et al.* and Doig *et al.* found that LR models performed competitively, or better, than the more complicated ANN models. Dybowski *et al.* used a GA to optimise the architecture of their ANN and achieved an AUROC of 0.863 which was superior to LR (AUROC 0.753) [174]. However, given the number of free parameters in their ANN (122) was high compared to the number of observations (168), and given the use of the test set to optimise the architecture it is likely the ANN would not generalise as well as its reported performance. In the Challenge, a linear model (Naive Bayes) [163] similarly outperformed a ANN model [164], though this comparison is not as appropriate due to different feature extraction methods. The gain in performance of more complex machine learning techniques versus linear models is empirically quantified on the Challenge database in Chapter 2.

A.9.1 Evaluation metrics

s_2, which consists of a modified $HL_{\hat{C}}$, is aimed at measuring the calibration of the model. However, it suffers from a number of limitations. Initially s_2 was directly equivalent to the $HL_{\hat{C}}$. Several entrants contacted the Challenge authors and notified them that the score was susceptible to a perverse entry which involved generating a set of predictions as uniform noise distributed around the prevalence of mortality. An entry such as this one would produce equally spaced deciles all with expected and observed mortality approximately equal to the prevalence, resulting in an essentially perfect s_2 nearing 0. This is simple to demonstrate: if the predictions are generated as $\hat{\mathbf{y}} \sim \mathcal{U}[\bar{\mathbf{y}} - \epsilon, \bar{\mathbf{y}} + \epsilon]$, then $E[E_j] = n_j\bar{\mathbf{y}} = O_j$, $E_j - O_j = 0$ and consequently $HL_{\hat{C}} \approxeq 0$. This led to the later modification of s_2 through addition of the range of the predictions to the denominator as seen in Equation A.2. Note that this modification removes the statistical properties of the s_2 which was approximately χ^2 distributed, and thus prohibits statistical hypothesis testing.

While a perfect s_2 could no longer be obtained by random predictions, the existence of the problem highlights the difficulty of s_2, and indeed the $HL_{\hat{C}}$ statistic in general, to properly assess the calibration of a model. The $HL_{\hat{C}}$ suffers from sensitivity to sample size [42], making it incomparable across datasets and studies. As incremental improvement of models is a hallmark of scientific progress, this makes it extremely difficult for independent groups to compare with and improve upon past results. Furthermore, one cannot compare a model's calibration across datasets using the $HL_{\hat{C}}$ or s_2, making it impossible to assess the loss of calibration when applying a model on an external dataset. As the most robust assessment of a model's predictive performance is evaluating it on a future dataset collected at a distinct institution from the one which collected the development data [177], this makes the $HL_{\hat{C}}$ and s_2 statistics useless for proper validation of a model. Indeed, the primary use of the $HL_{\hat{C}}$ is a frequentist based p-value which indicates fit or lack of fit. This p-value can be extremely misleading however: Kramer *et al.* [42] showed with simulations that when a perfect model's predictions are slightly modified to produce a 0.4% deviation from the true risk, the $HL_{\hat{C}}$ always indicates a significant deviation with a sample size of 50,000. The authors also showed that with 5,000

patients, comparable to the dataset size for the challenge datasets, a 0.4% deviation from perfect fit caused 34% of the $HL_{\hat{C}}$'s values to be significant (where significance implies poor model calibration). This casts strong doubt on the use of the $HL_{\hat{C}}$ as a measure of model performance, especially as datasets continue to grow in size. There have been attempts to create calibration metrics other than the $HL_{\hat{C}}$ for mortality prediction, such as the GiViTI belt [92,178].

The $\mathcal{I}_{\mathcal{L}}$ used in this work appears to be a useful alternative measure of model fit. The likelihood of a model is a standard metric in the machine learning community and very commonly used for the optimisation of predictive models, including logistic regression, though is less frequently used for ICU risk model assessment. Chapter 2 further discusses the use of this metric in the context of comparing models.

The s_1 of 53.53% on $\mathrm{PN_c}$ indicates that the model had at least a PPV of 53.53% and a sensitivity of 53.53%. This implies for patients with a predicted positive outcome, almost half of these positives are false. Furthermore, for every one of these positive outcomes that is correctly predicted by the model, one is incorrectly predicted as a negative outcome. Even though the model had objectively good performance, the accuracy of its actionable predictions is equivalent to a coin flip. While s_1 has effectively summarised the capability of models to predict hospital mortality, the utility of the binary predictions is not sufficiently high enough for this ranking to be a primary factor in clinical decision making.

Many models, such as the BEAST and logistic regression, output risks of mortality ranging from zero to one rather than binary outcomes. The use of s_1 in the Challenge required an additional step where the risks were converted into dichotomous outcomes. This is a very appropriate step; the definition of a threshold is a key component for creating a predictive model which has actionable decisions. While a higher AUROC is always desirable, it does not necessary guarantee a higher sensitivity and specificity at the desirable operating range. A heavily explored rectification of this flaw in the AUROC is the use of partial ROC curves [179], and the area under pROC curves may be an alternative to the s_1 for evaluating clinically relevant model discrimination.

It is worth noting that the estimate of risk can also be clinically useful. First,

an unexpectedly high risk of mortality for a patient would capture the attention of care providers and encourage a secondary review of the patient, reducing the chance of missed patient deterioration. Secondly, the risk could be used to optimise the efficiency of patient care when few ICU beds are available by providing an objective assessment of relative patient severity. As the ICU is consistently the most expensive area of the hospital, costing almost 0.56% of the GDP in the United States [180] and £1,328 per patient day in the United Kingdom [181], this could lead to substantial savings.

It is worth noting that the $\mathcal{I}_{\mathcal{L}}$ also evaluates the discrimination of a set of predictions, and as such provides a reasonable alternative of model performance in the mathematical sense and model efficacy as it pertains to clinical care.

A.9.2 Domain knowledge preprocessing

The domain knowledge preprocessing presented here was a key component to the overall Challenge entry. Only one other entry [161] preprocessed the data prior to using it in classification, and this preprocessing method was primarily intended to transform feature distributions to better resemble a normal distribution. Clinical contextualisation of the data used for modelling allows for exclusions of erroneous values which would otherwise reduce algorithm performance. In particular blood pressure and heart rate measurements are prone to containing erroneous zeros due to the automated nature of the data capture for these variables.

Critically, the preprocessing was applied to the Challenge data *before* the feature extraction. When an erroneous value (such as a zero for heart rate) was removed, the features extracted consequently contained relevant information regarding patient health (such as the actual observed lowest heart rate). This key distinction is revisited in Chapter 2, and is the reason why preprocessing remained important to the BEAST even though the model automatically handles missing data.

While the preprocessing increased the likelihood that erroneous information would not be used to learn a model of mortality, there exist a number of limitations with the described implementation. Due to the time sensitive nature of the Challenge, a few oversights in the DK preprocessing exist. For example, since a limit of 200 was applied

to diastolic blood pressure, it would have been reasonable to also apply similar (though likely larger) limits to systolic and mean blood pressure. Other Challenge entrants faced similar issues due to time constraints (such as Citi *et al.* [161] omitting mechanical ventilation). Furthermore, all serum measurements must be positive definite, but only a few (such as BUN) were processed in this way.

One additional source of prior knowledge was in the common types of errors made by humans. Though some of the data in the Challenge database was automatically collected, a large portion was at some stage input or over-read by a human operator. Although one might expect such data to contain fewer errors (such as inappropriate zeros), this is not always the case since lapses in vigilance can lead to errors in both transcription and judgement [106]. Knowledge of common errors thus allowed correction, as opposed to deletion, of erroneous values and this process likely increased the information content of the data. In particular, temperature had many erroneous values corrected due to unit of measurement issues (°C versus °F). A total of 148 observations were modified in this way for $\mathrm{PN_a}$, and a total of 81 observations were modified for $\mathrm{PN_b}$. As this is a very small proportion of the total observations (148 out of 86405 for $\mathrm{PN_a}$), it is possible that this was not an integral step to the preprocessing.

Disregarding implementation issues, there remain unresolved issues with the principle of preprocessing medical data using domain knowledge. Primary of these issues is the lack of standardised thresholds for any variables. The choice of an upper threshold of 200 for diastolic blood pressure was arbitrary, based off a conservative estimate of the most extreme diastolic blood pressure that could plausibly be recorded in a critically ill patient. There is no reason why a diastolic blood pressure of 199 would not be equally erroneous. Another example of a variable to which a threshold is commonly applied (though not herein) is $\mathrm{SpO_2}$, the oxygen saturation of blood calculated by a sensor monitoring the periphery of a patient. This variable represents the fraction of oxygenated haemoglobin to total haemoglobin, and thus values over 100% are impossible. This is an example of a straightforward threshold which can be applied. However, the choice of a lower threshold for $\mathrm{SpO_2}$ is less clear. The Royal College of Physicians [182] published an early warning score which treats values of oxygen saturation below 91% as

extremely abnormal and warranting attention. A common issue with the sensors used to measure the SpO_2 is accidental removal from the patient, which can result in erroneously low SpO_2 values. Thus it would be desirable to remove these artefacts with a low value threshold. However certain conditions, such as Raynaud's phenomenon, can cause low oxygen saturation measurements in the periphery which appear artefactual. This is an unfortunately common scenario in medical data: no threshold which only removes artefacts can be determined. This problem is exacerbated in variables with log-normal distributions such as urine output.

Another issue for domain knowledge preprocessing resides in the technical implementation. Since measurement units are not standardised across institutions, or even within institutions, care must be taken that the thresholds applied are appropriate for the distribution in question. The thresholds applied here for creatinine measurements, made in mg/dL, would be inappropriate for data collected in the United Kingdom where creatinine is commonly measured in μmol/L.

Appendix B

Mathematical exposition and pseudocode

B.1 Kernel density estimation

Kernel density estimation [146] involves smoothing an empirical distribution function with a kernel to provide a smooth estimate of the probability density function. If we represent the kernel function by K, then the density estimate can be calculated as:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{B.1}$$

Here x is an observation, N is the number of observations h represents the *bandwidth*. The bandwidth smooths the density estimate. As h increases, the estimated density function becomes smoother, whereas as h approaches zero the estimated density function approaches a histogram of the data. In this work h was calculated as:

$$h = \left(\frac{4}{3N}\right)^{-5}\hat{\sigma} \tag{B.2}$$

where $\hat{\sigma}$ is calculated as follows:

$$\hat{\sigma} == \frac{||\mathbf{x} - \tilde{\mathbf{x}}||}{0.6745} \tag{B.3}$$

Note that here $\tilde{\mathbf{x}}$ indicates the median of the vector $\mathbf{x}$ and $\hat{\sigma}$. The numerator above

is sometimes referred to as the median absolute deviation and the factor of 0.6745 is used to transform the median absolute deviation into an approximation of the standard deviation while maintaining the robust properties of the median. The calculation of the bandwidth h in this fashion has been shown to be an optimal choice if the underlying density is Gaussian [146].

B.2 Genetic algorithm pseudocode

for $n = 1 \rightarrow N$ **do**

  for $d = 1 \rightarrow D$ **do**

    $g^{(n,d)} \leftarrow \text{round}(\mathcal{U} \sim (0, 1))$

  end for

  $c^{(n,1)} \leftarrow 0$

end for

$n_{c-o} = \lfloor 0.45 \times N \rfloor$

$\boldsymbol{trainingData} \leftarrow$ loadTrainingData

$\boldsymbol{testData} \leftarrow$ loadTestData

repeat

  for $n = 1 \rightarrow N$ **do**

    $w \leftarrow \text{trainWrapper}(\boldsymbol{trainingData}(\boldsymbol{g}^{(n)} == 1))$

    $c \leftarrow \text{evaluateWrapper}(w, \boldsymbol{testData}(\boldsymbol{g}^{(n)} == 1))$

  end for

  $\boldsymbol{g} \leftarrow \text{sortByCost}(\boldsymbol{c}, \boldsymbol{g})$

  for $n = 1 \rightarrow \lfloor 0.1 \times N \rfloor$ **do**

    $\boldsymbol{g}^{(N-n+1)} \leftarrow \boldsymbol{g}^{(n)}$

  end for

  for $n = 1 \rightarrow n_{c-o}$ **do**

    $\boldsymbol{g_{parent_1}} \leftarrow \boldsymbol{g}^{(n)}$

    $n_{child_2} \leftarrow \underset{n}{\text{argmax}} \left[\sum_{m=1}^{n_{c-o}} (\boldsymbol{g} - \boldsymbol{g_{parent_1}})^2\right]$

    $\boldsymbol{g_{parent_2}} \leftarrow \boldsymbol{g}^{(n_{child_2})}$

    $CP \leftarrow \lceil \mathcal{U} \sim (0, D) \rceil$

    for $d = 1 \rightarrow D$ **do**

if $d < CP$ **then**

$\boldsymbol{g}^{(n,d)} \leftarrow \boldsymbol{g_{parent_1}}$

$\boldsymbol{g}^{(n+n_{c-o},d)} \leftarrow \boldsymbol{g_{parent_2}}$

else

$\boldsymbol{g}^{(n,d)} \leftarrow \boldsymbol{g_{parent_2}}$

$\boldsymbol{g}^{(n+n_{c-o},d)} \leftarrow \boldsymbol{g_{parent_1}}$

end if

end for

end for

until error gradient $< E$ {Note: E is user defined}

B.3 Particle swarm optimisation pseudocode

Note that $\otimes$ indicates element-wise multiplication.

for $d = 1 \rightarrow D$ **do**

for $n = 1 \rightarrow N$ **do**

$x^{(n,d)} \leftarrow \mathcal{U} \sim (0, 100)$

$p^{(n,d)} \leftarrow x^{(n,d)}$

$v^{(n,d)} \leftarrow 0$

end for

$g^{(1,d)} \leftarrow 0$

end for

repeat

for $n = 1 \rightarrow N$ **do**

if $f(\boldsymbol{x}^{(n)}) > f(p^{(n)})$ **then**

$\boldsymbol{p}^{(n)} \leftarrow \boldsymbol{x}^{(n)}$

end if

if $f(\boldsymbol{p}^{(n)}) > f(\boldsymbol{g})$ **then**

$\boldsymbol{g} \leftarrow \boldsymbol{p}^{(n)}$

end if

end for

$\boldsymbol{x} \leftarrow \boldsymbol{x} + \boldsymbol{v}$

for $n = 1 \rightarrow N$ **do**

for $d = 1 \rightarrow D$ **do**

$c_1^{(1,d)} \leftarrow \mathcal{U} \sim (0, 2.05)$

$c_2^{(1,d)} \leftarrow \mathcal{U} \sim (0, 2.05)$

end for

$\boldsymbol{v}^{(n)} \leftarrow i^{(k)}\boldsymbol{v}^{(n)} + \boldsymbol{c}_1 \otimes (\boldsymbol{p}^{(n)} - \boldsymbol{x}^{(n)}) + \boldsymbol{c}_2 \otimes (\boldsymbol{g} - \boldsymbol{x}^{(n)})$

$\boldsymbol{v}^{(k)} \leftarrow w^{(k)} \cdot \boldsymbol{v}^{(k)}$ {Velocity weight}

if $\boldsymbol{v}^{(k)} > v_{max}$ **then**

$\boldsymbol{v}^{(k)} \leftarrow -\boldsymbol{v}^{(k)}$ {Particle bounce}

end if

end for

until error gradient $< E$ {Note: E is user defined}

Appendix C

MIMIC II *itemids*

C.1 APACHE Outcomes diagnostic categories

Table C.1 provides the definitions associated with the diagnostic binary indicator variables used in the AO dataset. These categories are used to represent primary admission diagnosis.

Acronym	Description	Post-operative?
ACUTMI	Acute myocardial infarction	
AIROBS	Mechanical airway obstruction	
ALLERGY	Asthma	
ASPPNEU	Aspiration pneumonia	
BACVPNEU	Bacterial/viral pneumonia	
CABG	Peripheral artery bypass graft	Post-operative
CARDARR	Cardiac arrest	
CARDIOG	Cardiogenic shock	
CHFD	Congestive heart failure	
COAGTHRO	Coagulopathy/neutropenia/thrombocytopenia	
COMAMETU	Metabolic coma	
COPDD	Chronic obstructive pulmonary disease	
CVOTH	Other cardiovascular diseases	
DIABETIC	Diabetic ketoacidosis	
GENOTH	Other medical diseases	

GIBLEED	GI bleeding due to diverticulosis	
GIBLEUL	GI bleeding due to ulcer/laceration	
GIBLVAR	GI bleeding due to varices	
GIINFLA	GI inflammatory disease (ulcerative colitis/crohn's/pancreatitis)	
GIOTHER	Other GI diseases	
GIPERF	GI perforation/obstruction	
HEADTR	Head trauma (with/without multiple trauma)	
HEMAMISC	Other hematologic diseases	
HEPATF	Hepatic failure	
HYPERT	Hypertension	
ICHMED	Intracerebral hemmorhage	
LUNGSTRAN	Lung transplant	Post-operative
LUNGTRAN	Lung transplant	
MEDAORT	Aortic aneurysm	
METAMISC	Other metabolic diseases	
MULTRAUM	Multiple trauma (excluding head trauma)	
NEONEUR	Neurologic neoplasm	
NEURINF	Neurologic infection	
NEURMUS	Neuromuscular disease	
NEUROTH	Other neurologic diseases	
OD	Drug overdose	
OTHTRANS	Other transplant	
PARAPNEU	Parasitic pneumonia	
PERIART	Peripheral vascular disease	
PULEDEM	Pulmonary edema (non-cardiogenic)	
PULEMB	Pulmonary embolism	
RENOTH	Renal diseases	
RESPARR	Respiratory arrest	

RESPCA	Respiratory neoplasm (including larynx, trachea)	
RESPOTH	Other respiratory diseases	
RHYTHM	Rhythm disturbance	
S-KPTRANSP	Kidney transplant	Post-operative
S-LUNGSTR	Lungs transplant	Post-operative
S-OTHTRANS	Transplant (other)	Post-operative
S-PANTRAN	Pancreatic transplant	Post-operative
SAHMED	Subarachnoid Hemmorhage	
SAORTDIS	Aortic dissection	Post-operative
SCARDOTH	Other cardiovascular diseases	Post-operative
SCAROTID	Carotid endarterectomy	Post-operative
SCRANNEO	Craniotomy for neoplasm	Post-operative
SEIZ	Seizures	
SELAORT	Elective abdominal aneurysm repair	Post-operative
SEPSIS	Sepsis (non-urinary)	
SEPTICUT	Sepsis (urinary tract origin)	
SFEMAORT	Hip or extremity fracture	Post-operative
SGIBLEE	GI bleeding	Post-operative
SGICA	GI neoplasm (cancer)	Post-operative
SGICHOL	GI cholecystitis/cholangitis	Post-operative
SGIINFL	GI inflammatory disease	Post-operative
SGIOBS	GI obstruction	Post-operative
SGIOTH	Other GI diseases	Post-operative
SGIPERF	GI perforation/rupture	Post-operative
SHEADTR	Head trauma (with/without multiple trauma)	Post-operative
SICH	Intracerebral hemmorhage	Post-operative
SLAMINE	Laminectomy/other spinal cord surgery	Post-operative
SLIVERTR	Liver transplant	Post-operative

SMULTR	Multiple trauma (excluding head trauma)	Post-operative
SNEUROTH	Other neurologic diseases	Post-operative
SOBHYST	Hysterectomy	Post-operative
SPERISC	Peripheral vascular disease (no bypass graft)	Post-operative
SRENCA	Renal neoplasm	Post-operative
SRENOTH	Other renal diseases	Post-operative
SRENTRAN	Renal transplant	Post-operative
SRESOTH	Other respiratory diseases	Post-operative
SRESPCA	Lung neoplasm	Post-operative
SRESPINF	Respiratory infection	Post-operative
SRESPLAR	Respiratory/neoplasm (mouth, sinus, larynx, trachea)	Post-operative
SSAH	Subarachnoid hemmorhage	Post-operative
SSDH	Subdural/epidural hematoma	Post-operative
STROKE	Stroke	
SVALVE	Valvular heart surgery	Post-operative

Table C.1: *Definition of the diagnostic categories for the covariates available in the AO dataset.*

C.2 Severity Score Variable *itemids* in MIMIC II

A table of the *itemid*s and tables associated with each severity score variable is shown in Table C.2.

*itemid*s with corresponding descriptions which were used to calculate the urine output over the first 24 hours are shown in Table C.3.

Table Name	itemids	Variable	Notes
Albumin	772	Albumin (>3.2)	Chemistry
Albumin	1521	Albumin	Chemistry
Albumin	3727	Albumin (3.9-4.8)	Chemistry
Arterial PaCO2	778	Arterial PaCO2	ABG
Arterial PaO2	779	Arterial PaO2	ABG
Arterial pH	780	Arterial pH	ABG
Blood Pressure - arterial line	51	Arterial BP	
Blood pressure - non-invasive	455	NBP	
Blood Urea Nitrogen	781	BUN (6-20)	Chemistry
Blood Urea Nitrogen	1162	BUN	
Blood Urea Nitrogen	3737	BUN (6-20)	Chemistry
Blood Urea Nitrogen	5876	bun	
Creatinine	791	Creatinine (0-1.3)	Chemistry
Creatinine	1525	Creatinine	Chemistry
Creatinine	3750	Creatinine (0-0.7)	Chemistry
Direct Bilirubin	1527	Direct Bili	Chemistry
GCS - Eye component	184	Eye Opening	
GCS - Motor Component	454	Motor Response	
GCS - Total	198	GCS Total	
GCS - Verbal component	723	Verbal Response	
Glucose	811	Glucose (70-105)	Chemistry
Glucose	3744	Blood Glucose	Chemistry
Glucose	3745	BloodGlucose	Quick Admit
Haematocrit	813	Hematocrit	Hematology
Haematocrit	3761	Hematocrit (35-51)	ABG'S
Haematocrit - ABG	986	ABG Hct	
Heart Rate	211	Heart Rate	
pH - Arterial line	1126	Art.pH	ABG
pH - Arterial line	4753	pH (Art)	ABG
Potassium	829	Potassium (3.5-5.3)	Chemistry
Potassium	1535	Potassium	Chemistry
Potassium	3792	Potassium (3.5-5.3)	Chemistry
Respiratory Rate	618	Respiratory Rate	
Sodium	837	Sodium (135-148)	Chemistry
Sodium	1536	Sodium	Chemistry
Sodium	3803	Sodium (135-148)	Chemistry
Temperature °C	676	Temperature C	
Temperature °C	677	Temperature C (calc)	Derived from itemid 678
Temperature °F	678	Temperature F	
Temperature °F	679	Temperature F (calc)	Derived from itemid 676
Temperature °F	3654	Temp Rectal [F]	
Total Bilirubin	848	Total Bili (0-1.5)	Chemistry
Ventilator setting	190	FiO2 Set	
Ventilator setting	720	Ventilator Mode	
Ventilator setting	721	Ventilator No.	
Ventilator setting	722	Ventilator Type	
White Blood Cell Count	861	WBC (4-11,000)	Hematology
White Blood Cell Count	1127	WBC (4-11,000)	Hematology
White Blood Cell Count	1542	WBC	Hematology
White Blood Cell Count	3834	WhiteBloodC 4.0-11.0	Heme/Coag
White Blood Cell Count	4200	WBC 4.0-11.0	Heme/Coag

Table C.2: *List of the itemids used to extract variables from the MIMIC II database. The descriptions provided match those which appear in the database for these features.*

Item ID	Description	Item ID	Description
55	Urine Out Foley		
2119	urine output-angio		
56	Urine Out Lt Nephrostomy	2130	ANGIO URINE O/P
57	Urine Out Rt Nephrostomy	2366	angio urine output
61	OR Out OR Urine	2463	cath lab urine
65	OR Out PACU Urine	2507	TRUE URINE
69	Urine Out Void	2510	Urine in cath lab
85	Urine Out Incontinent	2592	VICU URINE OUT
94	Urine Out Condom Cath	2676	ANGIO URINE OUTPUT
96	Urine Out Ureteral Stent # 1	2810	angio urine out
288	PACU Out PACU Urine	2859	ANGIO URINE
405	Urine Out Other	3053	URINE OUT
428	Urine Out Straight Cath	3175	Urine .
473	Urine Out IleoConduit	3462	urine
651	Urine Out Ureteral Stent # 2	3519	urine amnt
715	Urine Out Suprapubic	3966	real urine output
1922	angio urine	3987	urine out or
2042	ANGIO URINE OUT	4132	Procedure urine out
2068	angio Urine output	4253	Urine out angio
2111	ANGIO FOLEY URINE	5927	True Urine

Table C.3: *List of the itemids in the MIMIC II database which were cumulatively summed to create the total urine output over 24 hours feature.*

Appendix D

Detailed performance comparisons

D.1 Hosmer-Lemeshow tables

D.1.1 MIMIC II database

.

Three tables are provided which provide the expected and observed number of deaths in equally spaced deciles for each severity score. Table D.1 shows the values for the OASIS, Table D.2 shows the values for the APS III and Table D.3 shows the values for the SAPS II.

	OASIS			
	Observed		Expected	
0-10%	19	(0.89)	35.4	(1.65)
10-20%	45	(2.10)	60.4	(2.82)
20-30%	63	(2.94)	85.6	(4.00)
30-40%	132	(6.17)	116.1	(5.42)
40-50%	129	(6.03)	154.5	(7.22)
50-60%	187	(8.73)	205.1	(9.58)
60-70%	266	(12.42)	271.5	(12.68)
70-80%	351	(16.39)	371.6	(17.36)
80-90%	506	(23.63)	538.5	(25.15)
90-100%	888	(41.48)	968.1	(45.22)

Table D.1: *List of the observed and expected deaths in equally sized deciles across OASIS with calibration coefficients for in hospital mortality. The severity scores were evaluated across 21,416 admissions to the MIMIC II database.*

	APS III			
	Observed		Expected	
0-10%	55	(2.57)	62.4	(2.91)
10-20%	47	(2.20)	85.3	(3.98)
20-30%	73	(3.41)	105.0	(4.91)
30-40%	103	(4.81)	126.5	(5.91)
40-50%	150	(7.01)	152.1	(7.10)
50-60%	177	(8.27)	184.8	(8.63)
60-70%	239	(11.16)	227.8	(10.64)
70-80%	320	(14.95)	293.8	(13.72)
80-90%	483	(22.56)	424.5	(19.83)
90-100%	938	(43.81)	879.0	(41.06)

Table D.2: *List of the observed and expected deaths in equally sized deciles across APS III with calibration coefficients for in hospital mortality. The severity scores were evaluated across 21,416 admissions to the MIMIC II database.*

	SAPS II			
	Observed		Expected	
0-10%	19	(0.89)	35.4	(1.65)
10-20%	45	(2.10)	60.4	(2.82)
20-30%	63	(2.94)	85.6	(4.00)
30-40%	132	(6.17)	116.1	(5.42)
40-50%	129	(6.03)	154.5	(7.22)
50-60%	187	(8.73)	205.1	(9.58)
60-70%	266	(12.42)	271.5	(12.68)
70-80%	351	(16.39)	371.6	(17.36)
80-90%	506	(23.63)	538.5	(25.15)
90-100%	956	(44.65)	1477.7	(69.02)

Table D.3: *List of the observed and expected deaths in equally sized deciles across SAPS II with calibration coefficients for in hospital mortality. The severity scores were evaluated across 21,416 admissions to the MIMIC II database.*

D.1.2 JR_{DB}

Three tables are provided which provide the expected and observed number of deaths in equally spaced deciles for each severity score. Table D.4 shows the values for the OASIS, Table D.5 shows the values for the APS III and Table D.6 shows the values for the SAPS II.

	OASIS			
Decile	Observed		Expected	
0-10%	11	(3.27)	4.9	(1.46)
10-20%	18	(5.36)	10.2	(3.03)
20-30%	21	(6.25)	16.2	(4.82)
30-40%	40	(11.90)	24.3	(7.22)
40-50%	53	(15.77)	35.0	(10.42)
50-60%	67	(19.94)	49.7	(14.78)
60-70%	84	(25.00)	68.7	(20.45)
70-80%	102	(30.36)	94.8	(28.22)
80-90%	129	(38.39)	137.0	(40.78)
90-100%	209	(62.20)	213.2	(63.45)

Table D.4: *List of the observed and expected deaths in equally sized deciles across the OASIS using calibration coefficients for in hospital mortality. The severity scores were evaluated across 3,336 admissions to the JR_{DB}.*

	APS III			
Decile	Observed		Expected	
0-10%	20	(5.95)	9.5	(2.84)
10-20%	23	(6.85)	15.0	(4.46)
20-30%	29	(8.63)	20.1	(5.98)
30-40%	34	(10.12)	25.7	(7.64)
40-50%	60	(17.86)	32.9	(9.81)
50-60%	60	(17.86)	44.1	(13.12)
60-70%	81	(24.11)	61.0	(18.15)
70-80%	108	(32.14)	87.3	(25.99)
80-90%	130	(38.69)	135.3	(40.27)
90-100%	190	(56.55)	231.7	(68.95)

Table D.5: *List of the observed and expected deaths in equally sized deciles across the APS III using calibration coefficients for in hospital mortality. The severity scores were evaluated across 3,336 admissions to the JR_{DB}.*

	SAPS II			
	Observed		Expected	
0-10%	13	(3.87)	7.2	(2.13)
10-20%	16	(4.76)	16.8	(5.01)
20-30%	32	(9.52)	27.5	(8.19)
30-40%	36	(10.71)	41.5	(12.36)
40-50%	50	(14.88)	60.3	(17.95)
50-60%	68	(20.24)	85.4	(25.42)
60-70%	81	(24.11)	122.7	(36.52)
70-80%	112	(33.33)	170.3	(50.68)
80-90%	130	(38.69)	224.4	(66.79)
90-100%	199	(59.23)	289.6	(86.20)

Table D.6: *List of the observed and expected deaths in equally sized deciles across the SAPS II using calibration coefficients for in hospital mortality. The severity scores were evaluated across 3,336 admissions to the* JR_{DB}.

D.2 Receiver Operator Characteristic (ROC) curves

D.2.1 MIMIC database

The ROC curves for the SOFA, SAPS, SAPS II, APS III and OASIS are provided in Figure D.1.

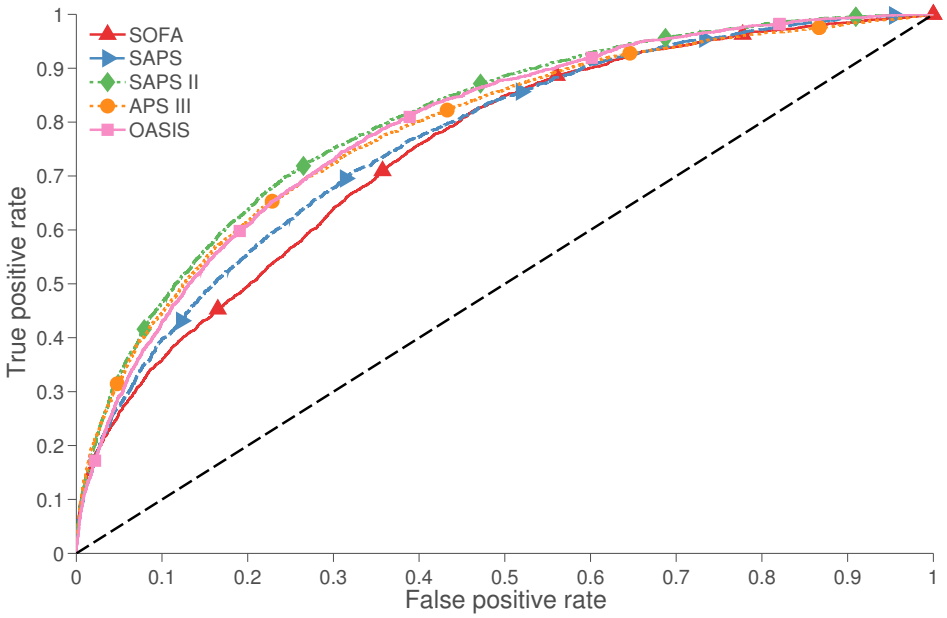

Figure D.1: *ROC curves for the severity scores showing the true positive rate (sensitivity) across varying levels of the false positive rate (1 - specificity). The ideal ROC curve would be two straight lines: one along $x = 0$ and one along $y = 1$. The dashed line indicates chance performance. The ROC curves are calculated across a dataset of 21,416 patients.*

D.2.2 JR_{DB}

The ROC curves for the SAPS II, APS III, IPS and OASIS are provided in Figure D.2.

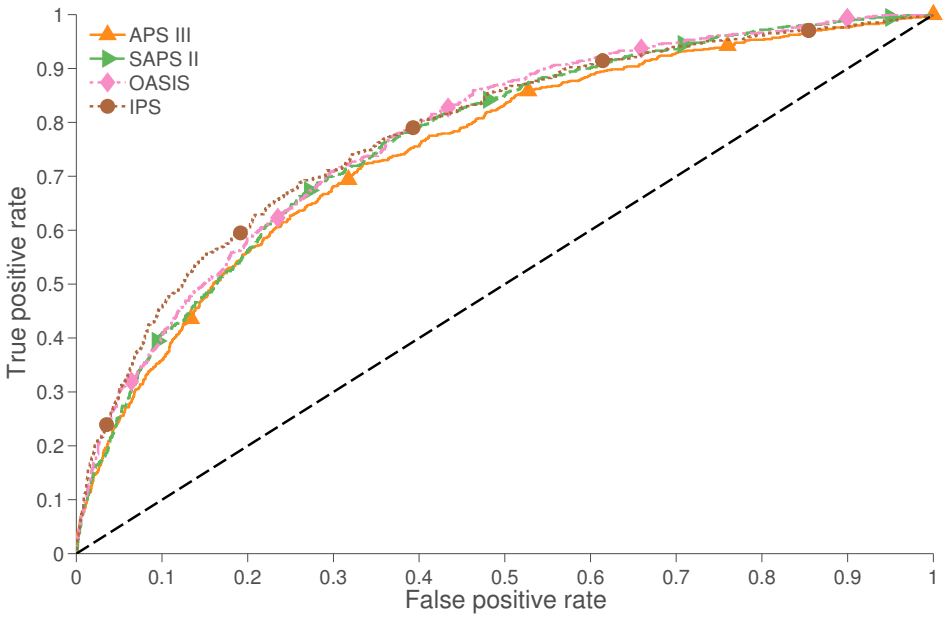

Figure D.2: *ROC curves for the severity scores showing the true positive rate (sensitivity) and the false positive rate (1 - specificity) across varying decision thresholds. The ideal ROC curve would be two straight lines: one along $x = 0$ and one along $y = 1$. The dashed line indicates chance performance. The ROC curves are calculated across the full JR_{DB} of 3,366 patients.*