# Classification Models in Intensive Care Outcome Prediction-can we improve on current models?

Nicholas A. Barnes,

Intensive Care Unit,
Waikato Hospital, Hamilton, New Zealand.

Lynnette A. Hunt,

Department of Statistics,
University of Waikato, Hamilton, New Zealand.

Michael M. Mayo,

Department of Computer Science,
University of Waikato, Hamilton, New Zealand.

Corresponding Author: Nicholas A. Barnes.

**Abstract**
Classification models ("machine learners" or "learners") were developed using machine learning techniques to predict mortality at discharge from an intensive care unit (ICU) and evaluated based on a large training data set from a single ICU. The best models were tested on data on subsequent patient admissions. Excellent model performance (AUCROC (area under the receiver operating curve) =0.896 on a test set), possibly superior to a widely used existing model based on conventional logistic regression models was obtained, with fewer per-patient data than that model.

## 1 Introduction

Intensive care clinicians use explicit judgement and heuristics to formulate prognoses as soon as reasonable after patient referral and admission to an intensive care unit [1].

Models to predict outcome in such patients have been in use for over 30 years [2] but are considered to have insufficient discriminatory power for individual decision making in a situation where patient variables that are difficult or impossible to measure may be relevant. Indeed even variables that have little or nothing to do with the patient directly (such as bed availability or staffing levels [3]) may be important in determining outcome.

There are further challenges for model development. Any model used should be able to deal with the problem of class imbalance, which refers in this case to the fact

that mortality should be much less common than survival. Many patient data are probably only loosely or indeed not related to outcome and many are highly correlated. For example, elevated measurements of serum urea, creatinine, urine output, diagnosis of renal failure and use of dialysis will all be closely correlated.

Nevertheless, models are used to risk adjust for comparison within an institution over time or between institutions, and model performance is obviously important if this is to be meaningful. It is also likely that a model with excellent performance could augment clinical assessment of prognosis. Furthermore, a model that performs well while requiring fewer data would be helpful as accurate data acquisition is an expensive task.

The APACHE III-J (Acute Physiology and Chronic Health Evaluation revision III-J [4]) model is used extensively within Australasia by the Centre for Outcomes Research of the Australian and New Zealand Intensive Care Society (ANZICS) and a good understanding of its local performance is available in the published literature [4]. It should be noted that death at hospital discharge is the outcome variable usually considered by these models. Unfortunately the coefficients for all variables for this model are no longer in the public domain so direct comparison with new models is difficult. The APACHE (Acute Physiology and Chronic Health Evaluation) models are based largely on baseline demographic and illness data and physiological measurements taken within the first day after ICU admission.

This study aims to explore machine learning methods that may outperform the logistic regression models that have previously been used.

The reader may like to consult a useful introduction to the concepts and practice of machine learning [5] if terms or concepts are unfamiliar.

## 2 Methods

The study is comprised of three parts:

1. An empirical exploration of raw and processed admission data with a variety of attribute selection methods, filters, base classifiers and metalearning techniques (which are overarching models that have other methods nested within them) that were felt to be suitable to develop the best classification models. Metamodels and base classifiers may be nested within other metamodels and learning schemes can be varied in very many ways .These experiments are represented below in Figure 1 where we used up to two metaclassifiers with up to two base classifiers nested within a metaclassifier.

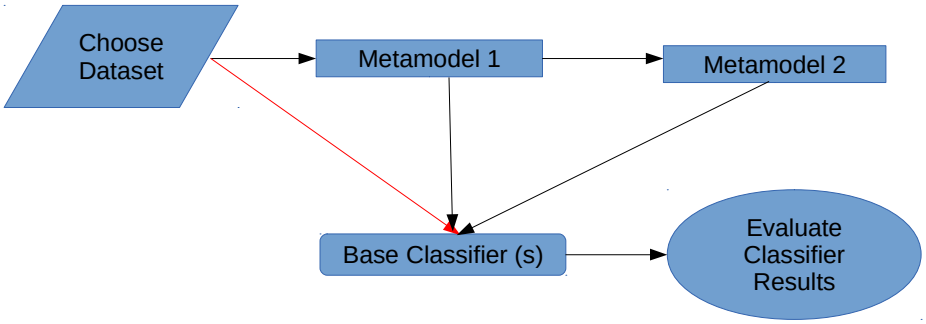


**Fig. 1.** Schematic of phase 1 experiments. Different color arrows indicate that one or more metamodels and base classifiers may optionally be combined in multiple different ways. One or more base classifiers are always required.

2. Further testing with the best performing data set (full unimputed training set) and learners with manual hyperparameter setting. A hyperparameter is a particular model configuration that is selected by the user, either manually or following an automatic tuning process. This is represented in a schematic below:

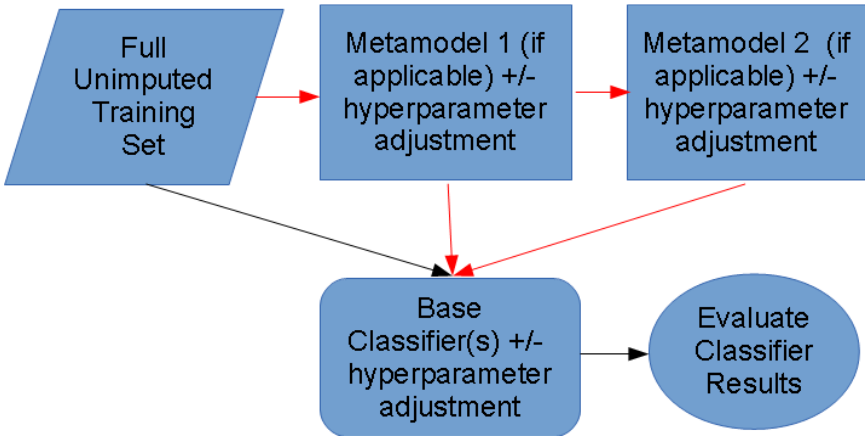


**Fig. 2.** Schematic of phase 2 experiments. As in phase 1, one or more metamodels may be optionally combined with one or more base classifiers.

3. Testing of the best models from phase 2 above on a new set of test data to better understand generalizability of the models. This is depicted in Figure 3 below.

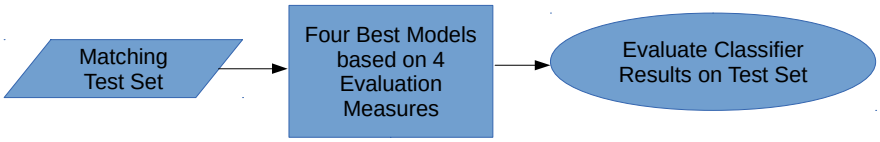


**Fig. 3.** Schematic of phase 3

The training data for adult patients (8122 patients over 16 years of age) were obtained from the database of a multidisciplinary ICU in a tertiary referral centre from a period between July 2004 and July 2012.Data extracted were comprised of a demographic variable (age), diagnostic category (with diagnostic coefficient from the APACHE III-J scoring system, including ANZICS modifications), and an extensive list of numeric variables relating to patient physiology and composite scores based on these, along with the classification variable: either survival, or alternatively, death at ICU discharge (as opposed to death at hospital discharge as in the APACHE models). Much of the data collected is used in APACHE III-J model mentioned above, and represents a subset of the data used in that model. Training data, prior to the imputation process, but following discretization of selected variables are represented in Table 1. Test data for the identical variable set were obtained from the same database for the period July 2012 to March 2013.

Of particular interest is that the data is clearly class imbalanced with mortality during ICU stay of approximately 12%. This has important implications for modelling the data.

There were many strongly correlated attributes within the data sets. Many of the model variables are collected as highest and lowest measures within twenty four hours of admission to the ICU. Correlated variables may bring special problems with conventional modelling including logistic regression. The extent of correlation is demonstrated in Figure 4.

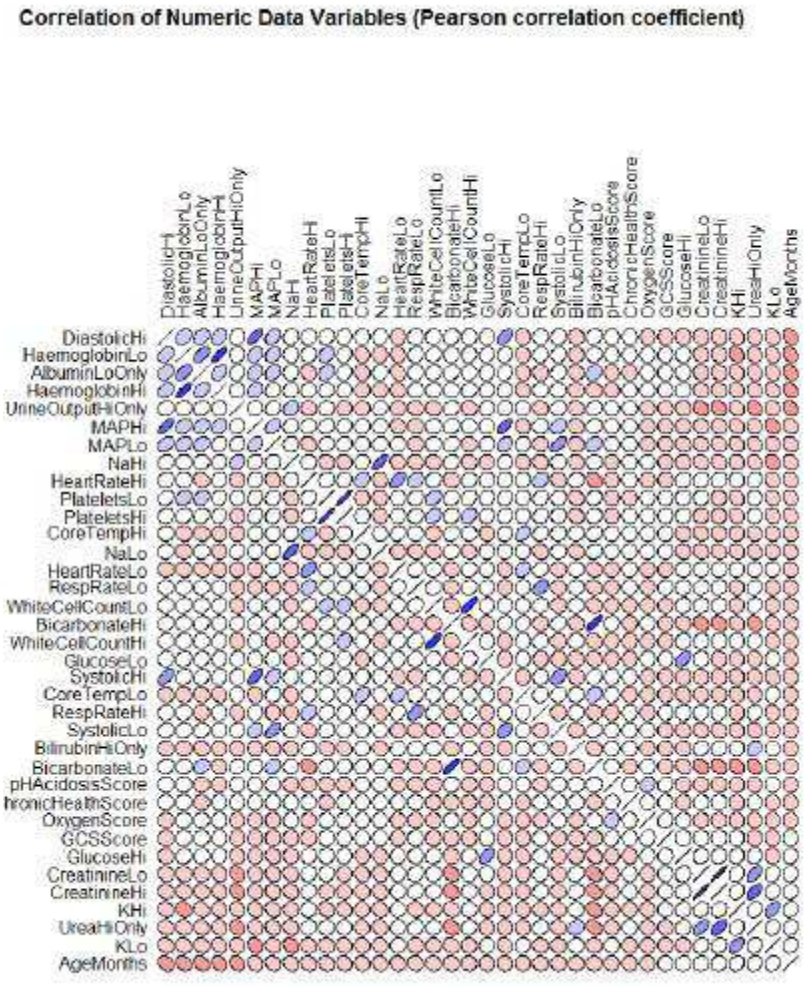


**Fig. 4.** Pearson correlations between variables are shown using colour. Blue colouration indicates positive correlation. Red colouration indicates negative correlation. The flatter the ellipse, the higher the correlation. White circles indicate no significant correlation between variables.

Patterns of missing data are indicated in Table 1 and represented graphically in Figure 5.

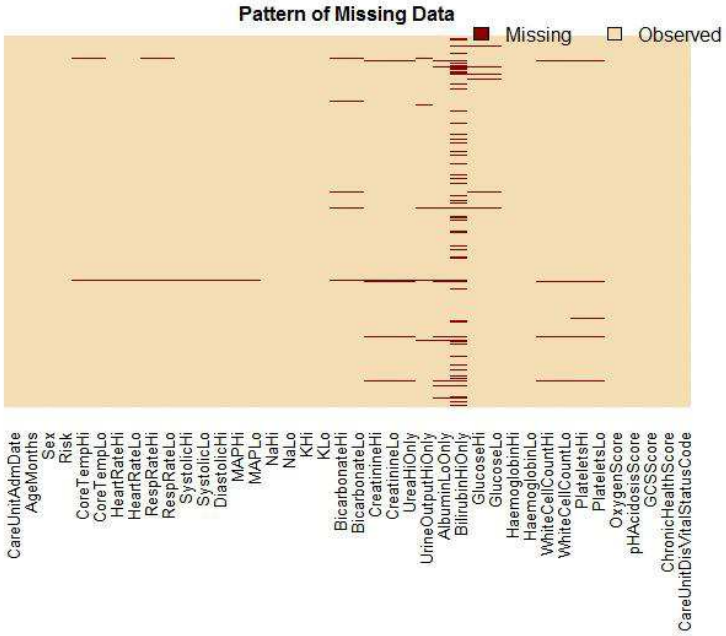


**Fig. 5.** Patterns of missing data in the raw training set. Missing data is represented by red colouration.

Missing numeric data in the training set was imputed using multiple imputation with the R program [6] and the R package Amelia [7], which utilises bootstrapping of non-missing data followed by imputation by expectation maximisation. We initially used the average of five multiple imputation runs.

Using the last imputed set was also trialled, as it may be expected to be the most accurate based on the iterative nature of the Amelia algorithm. No categorical data were missing. Date of admission was discretized to the year of admission, age was converted to months of age, and the diagnostic categories were converted to five to eight (depending on study phase) ordinal risk categories by using coefficients from the existing APACHE III-J risk model.

A summary of data is presented below in Table 1.

**Table 1.** Data Structure

| Variable | Type | Missing | Distinct values | Min. | Max. |
|---|---|---|---|---|---|
| CareUnitAdmDate | numeric | 0 | 9 | 2004 | 2012 |
| AgeMonths | numeric | 0 | 880 | 192 | 1125 |
| Sex | pure factor | 0 | 2 | F | M |
| Risk | pure factor | 0 | 8 | Vlow | High |

| | | | | | |
|---|---|---|---|---|---|
| CoreTempHi | numeric | 50 | 89 | 29 | 42.3 |
| CoreTempLo | numeric | 53 | 102 | 25.2 | 40.7 |
| HeartRateHi | numeric | 25 | 141 | 38.5 | 210 |
| HeartRateLo | numeric | 26 | 121 | 0 | 152 |
| RespRateHi | numeric | 38 | 60 | 8 | 80 |
| RespRateLo | numeric | 40 | 42 | 2 | 37 |
| SystolicHi | numeric | 27 | 161 | 24 | 288 |
| SystolicLo | numeric | 55 | 151 | 11 | 260 |
| DiastolicHi | numeric | 27 | 105 | 19 | 159 |
| MAPHi | numeric | 28 | 124 | 20 | 200 |
| MAPLo | numeric | 43 | 103 | 3 | 176 |
| NaHi | numeric | 46 | 240 | 112 | 193 |
| NaLo | numeric | 51 | 245 | 101 | 162 |
| KHi | numeric | 46 | 348 | 2.7 | 11.7 |
| KLo | numeric | 51 | 275 | 1.4 | 9.9 |
| BicarbonateHi | numeric | 218 | 322 | 3.57 | 48 |
| BicarbonateLo | numeric | 221 | 319 | 2 | 44.2 |
| CreatinineHi | numeric | 130 | 606 | 10.2 | 2025 |
| CreatinineLo | numeric | 134 | 552 | 10 | 2025 |
| UreaHiOnly | numeric | 232 | 433 | 1 | 99 |
| UrineOutputHiOnly | numeric | 184 | 3501 | 0 | 15720 |
| AlbuminLoOnly | numeric | 281 | 66 | 5 | 65 |
| BilirubinHiOnly | numeric | 1579 | 183 | 0.4 | 618 |
| GlucoseHi | numeric | 172 | 255 | 1.95 | 87.7 |
| GlucoseLo | numeric | 177 | 198 | 0.1 | 60 |
| HaemoglobinHi | numeric | 54 | 153 | 1.8 | 25 |
| HaemoglobinLo | numeric | 59 | 151 | 1.1 | 25 |
| WhiteCellCountHi | numeric | 131 | 470 | 0.1 | 293 |
| WhiteCellCountLo | numeric | 135 | 393 | 0.08 | 293 |
| PlateletsHi | numeric | 149 | 653 | 7 | 1448 |
| PlateletsLo | numeric | 153 | 621 | 0.27 | 1405 |
| OxygenScore | numeric | 0 | 8 | 0 | 15 |
| pHAcidosisScore | numeric | 0 | 9 | 0 | 12 |
| GCSScore | numeric | 0 | 11 | 0 | 48 |
| ChronicHealthScore | numeric | 0 | 6 | 0 | 16 |
| Status at ICU Discharge | pure factor | 0 | 2 | A | D |

Phase 1 consisted of an exploration of machine learning techniques thought suitable to this classification problem, and in particular those thought to be appropriate to a class imbalanced data set. Attribute selection, examining the effect of using imputed and unimputed data sets and application of a variety of base learners and metaclassifiers without major hyperparameter variation occurred in this phase. The importance of

attributes was examined in multiple ways including using random forest methodology for variable selection, using improvement in Gini index using particular attributes. This information is displayed in figure 6.

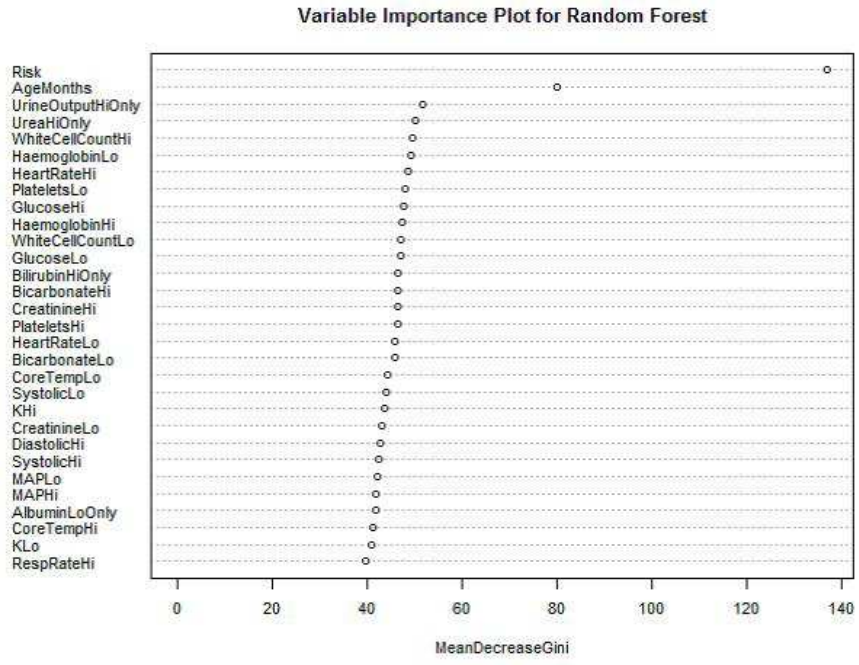


**Fig. 6.** Variable importance as measured by Gini index using random forest methodology. A substantial decrease in Gini index indicates better classification with variable inclusion. Variables used in the study are ranked by their contribution to Gini index.

A comprehensive evaluation of all techniques is nearly impossible given the enormous variety of techniques and the ability to combine up to several of these at a time in any particular model. Techniques were chosen based on the likely success of their application. WEKA [8] was used to apply learners and all models were evaluated with tenfold cross validation. WEKA default settings were commonly used in phase 1 and the details of these defaults are widely available [9]. Unless otherwise stated all settings in all study phases were the default settings of WEKA for each classifier or filter. Two results were used to judge overall model performance during phase 1. These were:

1. Area under the receiver operating curve (AUC ROC)
2. Area under the precision recall curve (AUC PRC)

The results are presented in Table 3 in the results section.

Phase 2 of our study involved training and evaluation on the same data sets with learners that had performed well in phase 1. Hyperparameters were mostly selected manually, as automatic hyperparameter selection in any software is limited and hampered by a lack of explicitness. Class imbalance issues were addressed with appropriate WEKA filters (spread subsample and SMOTE, a filter which generates a synthetic data set to balance the classes [10]), or the use of cost sensitive learners [11]. Unless otherwise stated in Table 3, WEKA default settings were used for each filter or classifier. Evaluation of these models proceeded with tenfold cross-validation and the results were examined in light of four measures:

1. Area under the receiver operating curve with 95% confidence intervals by the method of Hanley and McNeill [12]
2. Area under the precision recall curve
3. Matthews correlation coefficient and,
4. F-measure

Additionally, scaling the quantitative variables by standardizing or normalizing the data was explored as this is known to sometimes improve model performance [13].
The results of phase 2 are presented in Table 2 in the results section.
Phase 3 involved evaluating the accuracy of the best classification models from phase 2 on a new test set of 813 patient admissions. Missing data in the test set were not imputed. Results are shown in Table 3.

## 3 Results

Table 2 presents the results following tenfold cross validation on a variety of techniques thought suitable for trial in the modelling problem. These are listed in order of descending area under the curve of the receiver operating curve and the area under the precision recall curve is also presented.

**Table 2.** Phase 2 of study.

| Data | Preprocess | Meta Model 1 | Meta model 2 | Meta model 3 | Base classifier 1 | Base classifier 2 | ROC | PRC |
|---|---|---|---|---|---|---|---|---|
| Unimputed all variables | NA | Cost Sensitive Classifier matrix 0,5;1,0 | NA | NA | Random Forest 500 trees | NA | 0.895 | 0.629 |
| Unimputed all variables | NA | Cost Sensitive Classifier matrix 0,5;1,0 | NA | NA | Random Forest 200 trees | NA | 0.894 | 0.416 |
| Unimputed all variables | NA | Cost Sensitive Classifier matrix 0,5;1,0 | NA | NA | Naïve Bayes | NA | 0.864 | 0.418 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Unimputed all variables | Spread-subsample uniform | Filtered Classifier | Attribute selected classifier 20 variables selected on info. Gain and ranked | Vote | J4.8 tree | Naïve Bayes | 0.854 | 0.439 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | Logistic regression | NA | Logistic Regression | NA | 0.766 | 0.283 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | NA | NA | SimpleLogis-tic | NA | 0.766 | 0.28 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | Random Comm | NA | REP tree | NA | 0.753 | 0.259 |
| Imputed ten variables | NA | Filtered Classifier | NA | NA | Naïve Bayes | NA | 0.742 | 0.248 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | Adaboost M1 | NA | J48 | NA | 0.741 | 0.254 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | Vote | NA | Random Forest 10 trees | Naïve Bayes | 0.741 | 0.252 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | Bagging | NA | J48 | NA | 0.736 | 0.258 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | Decorate | NA | Naïve Bayes | NA | 0.735 | 0.238 |
| Imputed all variables | Spread-subsample uniform | Filtered Classifier | Attribute selected classifier 20 variables selected on info. Gain and ranked | Vote | J4.8 tree | Naïve Bayes | 0.735 | 0.238 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | NA | NA | J4.8 tree | NA | 0.734 | 0.234 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | NA | NA | Random Forest 10 trees | NA | 0.713 | 0.221 |
| Imputed ten variables | Spread-subsample uniform | Filtered Classifier | SMO | NA | SMO | NA | 0.5 | 0.117 |

ROC-area under receiver operating characteristic curve
CI-confidence interval
PRC-area under precision-recall curve
NA-not applicable

Table 3 presents the results of tenfold cross validation on the best models from phase 1 trained on the training set in phase 2 of our study. Models are listed in descending order of AUC ROC. The data set used in the modelling is indicated, along with any pre-processing of data, base learners, metalearners if applicable, and other evaluation tools as listed in the methods section above. The model which performs

best of all models on any of the four classification methods is shaded in red to emphasise that no one performance measure dominates a classifier's overall utility.

**Table 3.** Phase 2 results

| Preprocess | Metamodel1 | Metamodel2 | Base Model 1 | Base model 2 | ROC | ROC 95% CI's | PRC | MCC | F-measure |
|---|---|---|---|---|---|---|---|---|---|
| Spread subsample uniform | Filtered classifier | Rotation forest 100 iterations | Alternating decision tree 100 iterations | NA | 0.903 | (0.892,0.912) | 0.622 | 0.47 | 0.51 |
| NA | Cost sensitive classifier 0,5;1,0 | NA | Rotationforest 500 iterations | J 48 | 0.901 | (0.881,0.921) | 0.625 | 0.482 | 0.481 |
| Spread subsample uniform | Filtered classifier | Rotationforest 200 iterations | NA | J 48 | 0.897 | (0.888,0.906) | 0.606 | 0.452 | 0.494 |
| Spread subsample uniform | Filtered classifier | NA | Rotationforest 500 iterations | J 48 | 0.897 | (0.888,0.906) | 0.608 | 0.45 | 0.493 |
| Spread subsample uniform | Filtered classifier | NA | Rotation forest 500 iterations | J48 graft | 0.897 | (0.888,0.906) | 0.611 | 0.456 | 0.5 |
| Spread subsample uniform | Filtered classifier | Rotation forest 50 iterations | Alternating decision tree 50 iterations | NA | 0.896 | (0.887,0.905) | 0.608 | 0.452 | 0.495 |
| Spread subsample uniform | Filtered classifier | NA | Rotation forest 100 iterations | J 48 | 0.895 | (0.886,0.904) | 0.602 | 0.443 | 0.488 |
| NA | Cost sensitive classifier 0,5;1,0 | NA | Random forests (RF) 1000 trees 2 features each tree | NA | 0.893 | (0.879,0.907) | 0.599 | 0.506 | 0.561 |
| NA | Cost sensitive classifier 0,5;1,0 | NA | RF 500 trees 2 features each tree | NA | 0.892 | (0.878.0.906) | 0.598 | 0.511 | 0.567 |
| NA | Cost sensitive classifier 0,1;1,0 | NA | RF 500 trees 2 features each tree | NA | 0.891 | (0.867,0.915) | 0.602 | 0.416 | 0.398 |
| NA | Cost sensitive classifier 0,1;1,0 | NA | RF 1000 trees 2 features each tree | NA | 0.891 | (0.867,0.915) | 0.603 | 0.422 | 0.391 |
| NA | Cost sensitive classifier 0,10;1,0 | NA | RF 500 trees 2 features each tree | NA | 0.891 | (0.878,0.904) | 0.594 | 0.497 | 0.558 |
| NA | Cost sensitive classifier 0,5;1,0 | NA | Rotation Forest 50 iterations | J48 | 0.891 | (0.871,0.911) | 0.606 | 0.479 | 0.485 |
| Spread subsample | Filtered classifier | Bagging 150 iterations | J 48 C 0.25 M 2 | NA | 0.89 | (0.869,0.911) | 0.609 | 0.474 | 0.471 |
| Spread subsample | Filtered classifier | Bagging 200 iterations | J 48 C 0.25 M 3 | NA | 0.889 | (0.868,0.910) | 0.61 | 0.474 | 0.473 |
| NA | Cost sensitive classifier 0,1;1,1 | NA | RF 200 trees 2 features each tree | NA | 0.889 | (0.865,0.913) | 0.598 | 0.425 | 0.395 |
| Spread subsample | Filtered classifier | Bagging 100 iterations | J 48 C 0.25 M 2 | NA | 0.888 | (0.867,0.909) | 0.605 | 0.47 | 0.467 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| NA | Cost sensitive classifier 0,5;1,0 | NA | RF 100 trees 2 features each tree | NA | 0.888 | (0.864,0.912) | 0.594 | 0.42 | 0.396 |
| Spread subsample uniform | Filtered classifier | NA | Random committee 500 iterations | Random tree | 0.887 | (0.879,0.895) | 0.578 | 0.373 | 0.409 |
| Spread subsample | Filtered classifier | Adaboost M1 150 iterations | J 48 C 0.25 M 2 | NA | 0.886 | (0.865,0.907) | 0.584 | 0.48 | 0.476 |
| Spread subsample | Filtered classifier | Adaboost M1 100 iterations | J 48 C 0.25 M 2 | NA | 0.884 | (0.863,0.905) | 0.577 | 0.469 | 0.467 |
| Spread subsample | Filtered classifier | Bagging 50 iterations | J 48 C 0.25 M 2 | NA | 0.883 | (0.862,0.904) | 0.597 | 0.465 | 0.465 |
| Spread subsample uniform | Filtered classifier | NA | Random subspace 100 iterations | REP tree | 0.877 | (0.868,0.886) | 0.563 | 0.423 | 0.473 |
| Spread subsample uniform | Filtered classifier | NA | Multiboost AB 50 iterations | J 48 | 0.874 | (0.864,0.884) | 0.428 | 0.435 | 0.482 |

RF-random forest
REP-representative
NA-not applicable
MCC-Matthews correlation coefficient

Normalizing or standardizing the data did not improve model performance and indeed tended to moderately worsen it.

Table 4 presents the results of applying four of the best models from phase 2 on a test data set of 813 patient admissions which should be from the same population distribution (if date of admission is not a relevant attribute). Evaluation is based on AUC ROC, AUC PRC, Matthews's correlation coefficient and F-measure. These evaluations were obtained by WEKA's knowledge flow interface.

**Table 4.** Model results with new test set in Phase 3

| Data preprocessing | Metamodel 1 | Metamodel 2 | Base Classifer 1 | Base Classifier 2 | ROC | 95% CI ROC | PRC | MCC | F-meas |
|---|---|---|---|---|---|---|---|---|---|
| Spread subsample uniform | Filtered classifier | Rotation forest 100 iterations | Alternating decision tree 100 iterations | NA | 0.896 | (0.854,0.938) | 0.592 | 0.401 | 0.426 |
| Spread subsample uniform | Filtered classifier | Rotation forest 200 iterations | NA | J 48 | 0.893 | (0.863,0.923) | 0.571 | 0.525 | 0.534 |
| NA | Cost sensitive classifier 0,5;1,0 | NA | Rotation forest 500 iterations | J 48 | 0.887 | (0.821,0.953) | 0.561 | 0.386 | 0.411 |
| NA | Cost sensitive classifier 0,5;1,0 | NA | Random forest 500 trees, 2 features each tree | NA | 0.885 | (0.855,0.915) | 0.551 | 0.51 | 0.555 |

ROC-area under receiver operating characteristic curve
CI-confidence interval
PRC-area under precision-recall curve
MCC-Matthews correlation coefficient
F-meas-F-measure

## 4 Discussion

It is unrealistic to expect models to perfectly represent such a complex reality as that of survival from critical illness. Perfect classification is impossible because of the limitations of any combination of currently available measurements made on such patients to accurately reflect survival potential. Patient factors such as attitudes towards artificial support and presumably health practitioner and institution related factors are important. Additionally non-patient related factors which may be purely logistical will continue to thwart perfect prediction by any future model. For instance, a patient may die soon after discharge from the ICU if a ward bed is available and conversely will die within the ICU if a ward bed is not available and transfer cannot proceed. Models currently employed generally consider death at hospital discharge, but new factors that increase randomness can enter in the hospital stay following ICU discharge, so problems are not necessarily decreased with this approach.

The best models we have studied have excellent performance when evaluated following tenfold cross validation in the single ICU setting with use of fewer data points than the current gold standard model. Machine learning techniques usually make few distributional assumptions about the data when compared with the traditional logistic regression model. Missing data are often dealt with effectively with machine learning techniques while complete cases are generally used in traditional general linear modelling such as logistic regression. Clinical data will never be complete, as some data will not be required for a given patient, while some patients may die prior to collection of data which cannot subsequently be obtained. Imputation may be performed on data prior to modelling but has limitations. It is interesting that models trained on unimputed data tend to perform better than imputed data, both in phase 2 and with the test set in phase 3.

The best comparison we can make in the published literature is the work of Paul et al [4] which demonstrates that the AUC ROC of the APACHE-III-J model has varied between 0.879 and 0.890 when applied to over half a million adult admissions to Australasian ICUs between 2000 and 2009. Routine exclusions in this study included readmissions, transfers to other ICUs, and missing outcome and other data, and admission post coronary artery bypass grafting prior to introduction of the ANZICS modification to APACHE-III-J for this category. None of these were exclusions in our study. The Paul et al paper looks at outcome at hospital discharge, while ours examines outcome at ICU discharge. For these reasons the results are not directly com-

parable but our results for AUC ROC of up to 0.896 on a separate validation set clearly demonstrate excellent model performance.

The techniques associated with the best performance involve addressing class imbalance (i.e. pre-processing data to create a dataset with similar numbers of those who survive and those that die). This class imbalance is a well-known problem in classification. Mortality data from any healthcare setting tend to be class imbalanced. Our study shows that any approach to class imbalance in the data greatly enhance model performance. Cost sensitive metalearners [11], synthetic minority generation techniques (SMOTE [10]) and creating a uniform class distribution by subsampling across the data all improve model performance.

A cost sensitive learner indicates a technique that reweights cases according to a cost matrix that the user sets to reflect differing "cost" of misclassification of positive and negative cases. This intuitively lends itself to the intensive care treatment process where such a framework is likely implemented at least subconsciously by the intensive care clinician. For instance the cost of clinically "misclassifying" a patient may be substantial and clinicians would likely try hard to avoid this situation.

In our study, the ensemble learner random forests [14] with or without a technique to address class imbalance tends to outperform many more complex metalearners, or enhancements of single base classifiers such as bagging [15] and boosting [16]. Random forests involve generation of many different tree models, each of which splits the cases based on different variables and a criterion to increase information gain. Voting then occurs across the "forest" to decide on the best way to split the cases and this produces the model. The term ensemble simply represents the fact that multiple learners are involved, rather than a single tree. As many as 500 or 1000 trees are commonly required before the error of the forest is at a minimum. The number of variables to be considered by each tree may also be set to try and improve performance. The other techniques that produced excellent results were rotation forests either alone, with a cost sensitive classifier, or in combination with a technique known as alternating decision tree. Alternating decision tree takes a "weak" classifier (such as a tree classifier) and uses a technique similar to boosting to improve performance.

The reason extensive experimentation may be required to produce the best model is attributed to Wolpert [17] and described as the "no free lunch theorem", meaning that there is no one single technique that will model the best in every given scenario. Of course the same is true of any conventional statistical technique applied to multidimensional problems. Data processing and model selection are crucial to performance although if prediction alone is important, a pragmatic approach can be taken to the usual statistical assumptions. Machine learning techniques are generally not a "black box" approach however and deserve the same credibility as any older method, if application is appropriate.

Similarly, no single evaluation measure can summarize a classifier's performance and different model strengths and weaknesses may be more or less tolerable depending on the circumstances of model use and hence a range of measures are usually presented as we have done.

There are several weaknesses to our study. It is clearly from a single centre and may not generalize to other ICUs in other healthcare systems. Mortality remains a

crude measure of ICU performance but remains simple to measure and of great relevance nevertheless. The existing gold standard models usually measure classification of survival or death at hospital discharge, so are not necessarily directly comparable to our models which measures survival or death at ICU discharge.

We are unable to directly compare our models with what may be considered gold standards as some of these (e.g. APACHE IV) are only commercially available, and as mentioned before, even the details of APACHE-III-J are not in the public domain. The best comparison involving Australasian data using APACHE-III-J comes from the paper of Paul et al. [4] but as with all APACHE models, this predicts death at hospital discharge. Additionally, re-admissions were excluded which may be a significant factor beyond what are often relatively small numbers of re-admissions in any given ICU, as re-admissions suffer a disproportionately high mortality.

Exploration of the available hyperparameters of the many models examined has been relatively limited. The ability to do this automatically, and explicitly or in a reproducible way in WEKA and indeed any available software is limited although this may be changing [18]. Yet minor changes to these hyperparameters may produce meaningful enhancements in model performance. Tuning hyperparameters runs the risk of overfitting a model, but we have tried to guard against this by testing the data on a separate validation set.

Likewise, the ability to combine models with the best characteristics [19], which is becoming more common in prediction of continuous variables [20] is not yet easily performed with the available software.

We have not examined the calibration of our models. Good calibration is not required for accurate classification. Accurate performance across all risk categories is highly desirable in a model. Similarly, performance including calibration for different diagnostic categories that may become more significant in an ICU's case mix is not accounted for.

Modelling using imputed data in every phase of our study tends to show inconsistent or suboptimal performance. It may be that imputation could be applied more accurately by another approach that would improve model performance.

The major current use of these scores is in quality improvement activities. Once a score is developed which accurately quantitates risk, the expected number of deaths may be compared to those observed [21]. The exact risk for a given integer valued number of deaths may be derived from the Poisson binomial distribution and compared to the number observed [22]. A variety of risk adjusted control charts can be constructed with confidence intervals [23].

## 5 Conclusions

We have presented alternative approaches to the classification problem involving prediction of mortality at ICU discharge using machine learning techniques. Such techniques may hold substantial advantage over traditional logistic regression approaches and should be considered to replace these. Complete clinical data may be unnecessary when using machine learning techniques, and in any case are frequently

not available. Out of the techniques studied, random forests seems to be the modelling approach with the best performance and has an advantage that it is relatively easy to conceptualise and implement with open source software. During model training a method to address class imbalance should be used.

## 6 Bibliography

[1]. Downar, J. (2013, April 18). Even without our biases, the outlook for prognostication is grim. Available from ccforum: http://ccforum.com/content/13/4/168

[2]. Knaus WA, W. D. (1981). APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med, 591-597.

[3]. Tucker, J. (2002). Patient volume, staffing, and workload in relation to risk-adjusted outcomes in a random stratified sample of UK neonatal intensive care units: a prospective evaluation. Lancet, 99-107.

[4]. Paul, E., Bailey, M., Van Lint, A., & Pilcher, D. (2012). Performance of APACHE III over time in Australia and New Zealand: a retrospective cohort study. Anaesthesia and Intensive Care, 980-994.

[5]. Domingos, P. (2013, May 6). A few useful things to know about machine learning. Available from Washington University: http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

[6]. R Core Team. (2013, April 25). Available from CRAN: http://www.R-project.org/.

[7]. Honaker, J., King, G., & Blackwell, M. (2013, April 25). Amelia II: a program for missing data. Available from Journal of Statistical Software: http://www.jstatsoft.org/v45/i07/.

[8]. Hall, M., Eibe, F., Holmes, G., Pfahringer, B., & Reutemann, P. (2009, 1). The WEKA Data Mining Software: An Update. SIGKDD Explorations.

[9]. Weka overview. (2013, April 25). Available from Sourceforge: http://weka.sourceforge.net/doc/

[10]. Chawla, N. O., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Ariticial Intelligence Research, 321-357.

[11]. Ling, C. X., & Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem. In C. Sammat; G. Webb, editors. Encyclopaedia of Machine Learning. Springer.p.231-235.

[12]. Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 29-36.

[13]. Aksoy, S., & Haralick, R. M. (2013, May 20). Feature Normalization and Likelihood-based Similarity Measures for Image Retrieval. Available from cs.bilkent.edu: http://www.cs.bilkent.edu.tr/~saksoy/papers/prletters01_likelihood.pdf

[14]. Breiman, L. (2001). Random Forests. Machine Learning, 5-32.

[15]. Breiman, L. (1996). Bagging predictors. Machine Learning, 123-140.

[16]. Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm.In Machine Learning:Proceedings of the Thirteenth International Conference on Machine Learning, (pp. 148-156). San Francisco.

[17]. Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. Neural computation, 1341-1390.

[18]. Thornton, C., Hutter, F., Hoos, H., & Leyton-Brown, K. (2013, April 21). Auto-WEKA: Combined selection and hyperparameter optimisation of classification algorithms. Available from arxiv.org: http://arxiv.org/pdf/1208.3719.pdf

[19]. Carauna, R., Nikilescu-Mizil, A., Crew, G., & Ksikes, A. (2013, May 20).[Internet]Ensemble selection from libraries of models. Available from cs.cornell.edu: http://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml04.icdm06long.pdf

[20]. Meyer, Z. (2013, April 21). New package for ensembling R models [Internet]. Available from Modern Toolmaking: http://moderntoolmaking.blogspot.co.nz/2013/03/new-package-for-ensembling-r-models.html

[21]. Gallivan, S; (2003) How likely is it that a run of poor outcomes is unlikely? European Journal of Operational Research , 150 46 - 52.

[22]. Hong, Y. (2013) On computing the distribution function for the Poisson binomial distribution. Computational Statistics and Data Analysis 59 41–51

[23]. Sherlaw-Johnson C. 2005 A method for detecting runs of good and bad clinical outcomes on Variable Life-Adjusted Display (VLAD) charts. Health Care Manag Sci. Feb;8(1):61-5.