The School of Mathematics

THE UNIVERSITY
of EDINBURGH

On the use of electronic health records for building clinical prediction models

by

Kevin Maske

Dissertation Presented for the Degree of
MSc in Statistics and Operational Research

August 2018

Supervised by
Dr. Vanda Inacio De Carvalho
Dr. Catalina A. Vallejos

Abstract

Electronic health records routinely collected by health providers have been increasingly used in research, particularly in models that dynamically predict patient outcomes using the information in these records. In this dissertation, data from the Medical Information Mart for Intensive Care (MIMIC-III) dataset was analysed using logistic regression and joint longitudinal and survival modelling techniques to build statistical models that aimed to predict in-ICU mortality within 72 hours of admission.

Logistic regression models built using the first 16 hours of observations and joint models using all available longitudinal information were found to perform poorly in terms of prediction. The multivariate Bayesian joint model that used only the first 16 hours of observation and was fit on balanced subsamples of the cohort was found to also perform poorly. In comparison, the joint model using all observations for a single longitudinal variable fitted on a balanced subsample (where the number of cases and controls were equal) was found to perform well. This suggested that a combination of balanced datasets and longer periods of observation for the longitudinal measurements was necessary to build good predictive models.

Acknowledgements

First and foremost, I would like to thank my dissertation supervisors: Dr. Catalina Vallejos and Dr. Vanda Inacio. Their expertise and passion in the intersecting fields of statistics and healthcare were invaluable to the development of the study. The countless lessons I learned from them throughout this entire journey are ones that I will carry for years to come.

Second, I would like to thank my sister, Pia Maske, for her unwavering support throughout this past year that I've been in Edinburgh. From helping to spot typographical errors in papers, to being willing to lend an ear despite her busy schedule, her presence has been a steady reminder that family was never too far away.

Third, I would like to thank my peers from the Philippines who were willing to help ensure the quality of this dissertation through feedback and consultations, namely: Camille Dee, Micah Alampay, and Ryan Paul Gozum.

Finally, I would like to thank my partner, Dr. Sheldon Wong, MD-MBA. While the insights he provided as a medical practitioner were of great use in the development of this dissertation, more significant was the manner in which he shared of himself throughout this past year that I have been away from home. I would not have accomplished the things I have without him standing beside me, and I will forever be grateful.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text

Kevin Maske

Contents

1 Introduction

Electronic health records (EHRs) are digital editions of patient information, as collected throughout their medical care[1]. Although the information included in these records are collected primarily for hospital bookkeeping purposes, the utilisation of EHRs in research settings has seen growing interest in recent years, despite suggestions in literature that EHR usage suffers problems in accuracy and data quality [19].

One area that makes use of EHRs is dynamic prediction. Dynamic predictions aim to use longitudinal measurements often found in these records to predict patient outcomes, such as death or infection. In this study, data from the Medical Information Mart for Intensive Care (MIMIC-III) dataset was used to build models that aimed to predict the occurrence of death inside the intensive care unit (ICU) within 72 hours of admission. It must be noted, however, that the results shown in this study correspond only to initial exploration, and substantial future work is necessary to deliver a prediction system that is usable in practice.

The dissertation is organised as follows: the remainder of Section 1 will discuss the dataset used and the background of the study. Section 2 will discuss the methodological background applicable to all portions of the study, as well as the procedures done to pre-process the raw data. Section 3 will discuss the application of logistic regression techniques. Section 4 will discuss the process and results of building both frequentist and Bayesian joint longitudinal-survival models on the data. Finally, Section 5 will summarise the findings of the study and discuss future prospects.

1.1 MIMIC-III Database

The MIMIC-III database is a compilation of de-identified health-related information gathered during routine hospital care. As a single-source database, it is populated by data about patients admitted to the ICU of the Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012. The MIMIC-III dataset includes 38,597 unique adult patients (16 years and above), with a total of 49,785 admissions. It includes information ranging from demographic (e.g. age, sex, race), administrative (e.g. admission times, services administered), and longitudinal physiological measurements (e.g. heart rate, respiration rate) [9].

The MIMIC-III database is an update to the 2010 MIMIC-II database, and is the only freely accessible database of its kind and scale. Its wide availability and open nature hopes to allow greater reproducibility in clinical studies, an issue that has been discussed by Johnson et al. in a study where they failed to reproduce the cohort sizes of previous analyses performed on the MIMIC database [8].

The data is freely available online through `mimic.physionet.org`, although access is granted only upon the completion of the CITI "Data or Specimens Only Research" course, an online course on human subject protection training and research ethics, as well as a signed data use agreement. The dataset is provided through comma separated value (CSV) files, accompanied by codes that can be used to build the dataset in systems such as SQL. Pre-processing methods are outlined in the MIMIC-III documentation and the scripts accompanying the database[2].

1.2 Significance of the Study

Medical professionals make critical care decisions everyday, such as whether or not to proceed with some form of treatment on a certain patient. The effects of such decisions are wide and impactful, ranging from outcomes such as curing a slow-progressing disease, to preventing death from acute phyisiological deterioration. Literature suggests a rise in evidence that medical professionals utilise intuitive strategies, usually informed by years of medical study and experience, to make these critical care decisions [10].

[1] See `www.healthit.gov/faq/what-electronic-health-record-ehr`.

[2] Visit `mimic.physionet.org/tutorials` for details on how this was done.

The employment of intuitive methods is reasonable, given the time-sensitive and high pressure environments that medical professionals operate in. However, evaluating the quality of the decisions made using such methods and whether they are actually beneficial to the patients is an issue of great importance. Intuitive methods often rely on thresholds or trigger points for different physiological markers. However, the values different hospitals assign for these triggers are widely varied, and often fail to account for patient-specific information [1]. Thus, situations arise such as those found in a study by Ospina-Tascn et al., where they found that among 72 ICU admissions that had interventions, only 10 of these interventions showed a positive effect on preventing or delaying death, while 7 actually showed a negative effect [12].

Recent years have seen a rise in the digitalisation of medical healthcare data [9], leading to an increased interest in using data to produce evidence-based insights on the complexities of critical care. While there are currently still very few results from such studies that have found successful implementation in the practical setting, this is expected to change in the coming years due to the growth the field is experiencing [15].

This study aimed to build statistical models to predict in-ICU patient mortality within 72 hours of admission using static information (e.g. sex, race) alongside longitudinal information (e.g. heart rate, respiration rate). The study used the data available in the MIMIC-III database to draw empirical insights into patterns that might help inform the intuitive methods often used in the critical care environment. While the study was by no means exhaustive, using only a subset of the information available in the database and a limited number of methods, it allowed some preliminary conclusions to be drawn about the variables considered, as well as showed the way statistical methods could be employed on the MIMIC-III database.

Excerpts of the scripts used in this dissertation are provided, allowing future studies to make use of and verify the results presented[3].

1.3 Scope and Limitation

This dissertation was a direct follow up to a study performed by Maria Gordillo-Maranon as part of a PhD rotation project supervised by Dr. Catalina Vallejos at the Alan Turing Institute, from February to May 2018[4]. R scripts that were used during data preparation were also provided by Gordillo-Maranon.

The models built in this study made use of the following variables, as extracted from the MIMIC-III database:

- **Static Variables**: Age (upon first admission), Sex, Race, Service Category
- **Longitudinal Variables**: Heart Rate (HR), Respiration Rate (RR), Blood Oxygen Saturation (O2S), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP)

Note that the variables considered in this study were only a subset of those studied by Gordillo-Maranon. In particular, the variables pertaining to pH, white blood cell count, temperature, weight, height, and BMI were removed. The details of these exclusions are discussed at the end of Section 2.1.1.

The major outcome considered was fatality inside the ICU within 72 hours of admission. Fatalities that occurred within 72 hours of admission but after the patient was discharged from the ICU were not considered as cases. The specifics of cohort building (exclusion criteria) are also discussed in Section 2.1.1.

The methods used for analysis were limited to logistic regression and joint modelling of longitudinal and survival outcomes. Treatment data was not used in the study as this, since this study was limited to investigating how mortality predictions changed based solely on physiological measurements rather than interventions.

[3]These scripts can be accessed via `github.com/kevinmaske/Maske_Dissertation`.

[4]The results are unpublished. However, relevant extracts can be accessed via `github.com/kevinmaske/Maske_Dissertation`.

While the availability of MIMIC-III data was what made the study possible, the reliance of its results on the dataset also presented a major limitation. There is discussion in literature about the dangers of over-reliance on a data-driven approach to critical health care, stating that insights drawn purely from big data, rather than in tandem with medical professionals and experts, run the risk of being uninformed or oversimplified [5].

2 Preliminaries

2.1 Data Preparation

2.1.1 Cohort Building

Results from Gordillo-Maranon

The full MIMIC-III dataset contained a total of 61,532 admissions. The first level of exclusion performed was based on what Johnson et al. deemed as "fundamental for all studies." Subjects that satisfied the following criteria were excluded [8]:

1. Patients below 15 years of age at the time of ICU admission
2. Patients with "no charted observations, no measurements of heart rate, or an incomplete administrative recording of ICU admission and discharge"
3. Patients with ICU records due to organ donation
4. Patients that stayed less than 4 hours in the ICU

After the first level of exclusion, what can be called the "Johnson Cohort" was obtained, containing a total of 52,085 admissions. Further exclusion criteria were applied by Gordillo-Maranon, wherein subjects that satisfied the following conditions were also removed from the study:

1. Patients that had been to the ICU more than once
2. Patients with more than one hospital admission record
3. Patients above 89 years of age

(1) and (2) were to ensure that each subject in the study corresponded to a unique admission, so that a single person might not have more influence over the results than others. (3) was due to issues arising from how patients older than 89 years of age were being de-identified. The MIMIC-III dataset de-identified these individuals by shifting their birth dates to exactly 300 years before their first admission. Such a shifting would cause issues in modelling when the age variable is included.

After the second level of exclusion, there were 36,227 remaining samples. These remaining samples were merged with the table that contained static data records (that is, a further filtering was performed based on availability of static data), and Gordillo-Maranon ended with a final cohort size of 36,198.

The next steps of data preparation involved the longitudinal data. Each of the individual longitudinal variables was contained in a separate data table that was in long format. Each dataset was filtered to remove entries that were outside the range of values deemed to be "phyisiologically valid". The construction of these valid ranges was performed following the method outlined by Johnson et al. [7]. For each longitudinal variable, each observation for each patient at all times available were collected into a single set of data. The Box-Cox transformation was then applied to these unified datasets, making their distributions approximately normal. Observations that had values below and above certain extreme tail thresholds were then removed; the values of these tail thresholds were set as the limits for valid observations.

Table 1 summarises the valid ranges obtained in Gordillo-Maranon's analyses, along with a column comparing her results with figures obtained in consultations with clinicians.

Variable	Domain Knowledge	Gordillo-Maranon
Heart Rate	[40, 200]	[22, 257]
Respiration Rate	[10, 40]	[2, 91]
Blood Oxygen Saturation	[40, 100]	[70, 100]
Systolic Blood Pressure	[50, 250]	[33, 341]
Diastolic Blood Pressure	[0, 160]	[5, 308]

Table 1: Valid ranges for longitudinal variables.

Preparation specific to this study

First, longitudinal measurements were removed based on observation time. Specifically, measurements that had chart times less than 0 (before ICU admission) and greater than the discharge time recorded for each patient (after ICU admission) were removed. The study is focused solely on information obtained while patients were inside the ICU; the removal made no assumptions on the irregularity of the recorded times.

Second, longitudinal measurements were aggregated into time bins. One reason for aggregating longitudinal data is that both joint modelling and logistic regression required some level of uniformity on the observation times of the longitudinal variables considered. Unlike in studies designed with specific follow-up times for which the issue of uniformity is trivial, the MIMIC-III dataset has observation times which were not fixed. As a solution, data was aggregated using the mean of the measurements for a given subject over some interval of time.

To choose an appropriate length for the time bins over which information would be aggregated, the trade-off between sample size and precision was considered. Choosing larger granularities would result in more individuals having records for each time interval at the cost of losing data variability; the reverse is also true of choosing finer smaller intervals. This property of the data is illustrated in Figure 1.

Ultimately, the data was binned into 4-hour intervals. 4-hour bins were chosen as a compromise between sample size and precision. Smaller bins were not chosen since the predictions of relevance to the study were for longer time spans. Bins larger than 4 hours were not chosen as too much of the information would have been be lost, and the imputation strategy used in this study would have been difficult to justify over long time spans.

Of the two main modelling methods employed, logistic regression had the more stringent requirements for its data. The data had to be complete: all subjects that were to be entered into the cohort needed to have data for all the bins considered. This meant that for some specified data cut-off bin T, the cohort would only be composed of subjects who had measurements for all the longitudinal variables for all of bins 1 to T. Note that as T increases, the number of subjects eligible to be part of the cohort decreases, since as time passes, more and more patients die or get discharged, leaving them with no more data for later time bins.

It was deemed that the restriction of having data for all of bins 1 to T was too stringent. Thus, a scheme was developed that would allow the cohort to include subjects that required only minimal imputation. In particular, the following criterion was used: for a subject to be included in the dynamic cohort, they must not have had missing data for two or more consecutive bins. This allowed the values of missing bins to be imputed using the measurements from the bins before and after it. The selected method of imputation was linear interpolation, appealing to the intuition that longitudinal variables would tend to behave in a continuous manner. The criterion is illustrated in Table 2.

For the purposes of logistic regression, a data cut-off time of 16 hours, or 4 bins, was specified. The combination of the data cut-off time and the leniency criteria resulted in a dynamic cohort of 14,113. Of these subjects, 169 were cases, making up 1.20% of the cohort. As a consequence of having a data cut-off time of 16 hours, the results presented in this project were not generalisable to to the entire population of the MIMIC-III dataset. Rather, along with the exclusion criteria described in this section, the criteria imposed an implicit condition that patients must have had

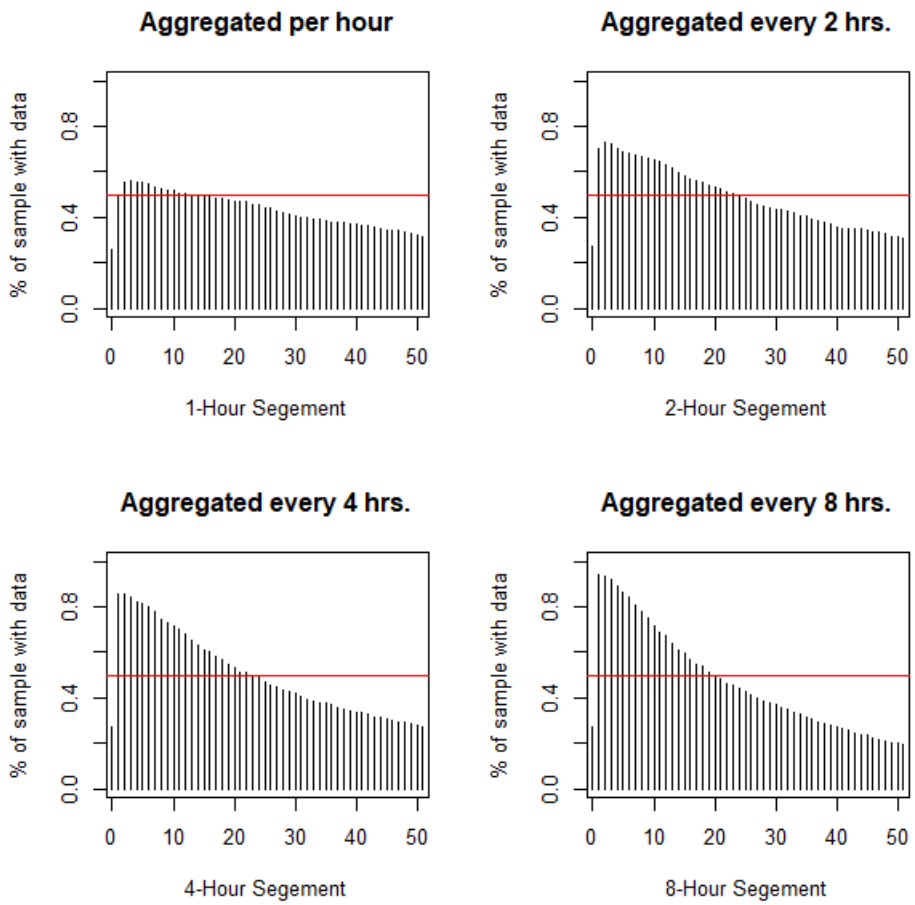

Figure 1: Data availability for heart rate by data granulation. Each plot shows the proportion of the cohort built by Gordillo-Maranon (N = 36,198) that had any observations across different time bins. A horizontal line is superimposed on each graph to indicate 50%.

survived at least 16 hours inside the ICU.

In the interest of being able to compare the different techniques, the same cohort built for the logistic models was used in the joint longitudinal survival models. However, as the joint modelling framework is not as rigid as the logistic framework, more longitudinal data could be included in the joint models. In particular, since the R package used (JM) allowed for patients with missing observations, all measurements available in the 72-hour period considered were used for the frequentist joint models, which focused on single longitudinal outcome models only.

Exclusion of Variables

The variables that were removed from consideration were pH, white blood cell count (WBC), temperature, weight, BMI, and height. For all these variables, the main reason for their exclusion was the effect their inclusion would have in sample size.

Note that since the MIMIC-III dataset exhibited heavy class imbalance, it was preferable to have as many samples as possible to give the models more power. Due to the stringent requirements of logistic regression, the inclusion of variables with low data availability would dramatically decrease the eligible cohort for the study. Further, too much missing data prevents meaningful imputation without over-modifying the data.

Figure 2 shows plots that summarise the data availability per hour for pH, WBC, and temperature. Note that for pH and WBC, the data availability was very low, where even the use of 12 hour bins was unable to produce cohorts of larger than approximately 7,000. Although

	Mean Heart Rate				
	0 – 4 hrs.	4 – 8 hrs.	8 – 12 hrs.	12 – 16 hrs.	Included?
Patient A	100	98	101	95	**Yes**; data is complete.
Patient B	NA	90	NA	105	**Yes**; data can be interpolated (in the case of 0 – 4 hours, extrapolated).
Patient C	80	**NA**	**NA**	85	**No**; consecutive bins with missing data (in red) prevent interpolation.
Patient D	NA	76	100	**NA**	**Depends**; if the patient has data for 16 – 20 hours, the missing data from 12 – 16 hours can be interpolated, and the patient is included. Otherwise, the patient is excluded.

Table 2: An illustration of the criterion used to determine if a subject had a sufficient amount of observations. The example shows an invented set of patients and measurements, and assumes a T = 16-hour cut-off time.

not as extreme, a similar situation was also true of the temperature observations, the inclusion of which would have cut the cohort down by around 25%.

Along the same vein, of the 14,113 subjects, 1,697 (12%) had missing weights, 7702 (54%) had missing heights, and 7728 (55%) had missing BMIs. Due to the large effect their inclusion would have in the cohort size, they were also removed from consideration all together. Note that the removal was done only due to the fact that the aims of the study were academic and exploratory by nature; metrics such as weight and BMI, when used in a formal clinical settings, would no doubt be invaluable to building risk models for patients.

2.1.2 Descriptive Analysis

Table 3 shows summary statistics of the static information pertaining to the cohort built in Section 2.1.1[5].

Initial exploration of the longitudinal variables was performed using the cohort built in Section 2.1.1. Figure 3 shows some characteristics of the heart rate measurements of the entire cohort over the first 72 hours of ICU admission. Figure 4 compares the behaviour of the heart rates of the patients depending on whether they were cases (in-ICU death within 72 hours) or controls.

From the plots shown in Figure 3, it can be seen that the distribution of heart rate remained relatively flat over time. On the other hand, Figure 4 suggests that while the average heart rates seemed to differ between cases and controls, the dataset still exhibited high enough variance that any differences would not be statistically discernible. In particular, the extreme heart rate plots shown in the same figure suggest that the ranges of values taken by heart rate tended to be the same regardless of a patient's case status.

While the plots shown in figures 3 and 4 pertained only to heart rate measurements, the same observations were also found to be true of the other variables - flat, and non-discriminatory of case status. Given this, a reasonable expectation was that models built using these variables would exhibit low predictive power on the cohort.

[5]For all modelling techniques, age was centred around the cohort mean to help control the magnitude of the intercepts of the fitted models.

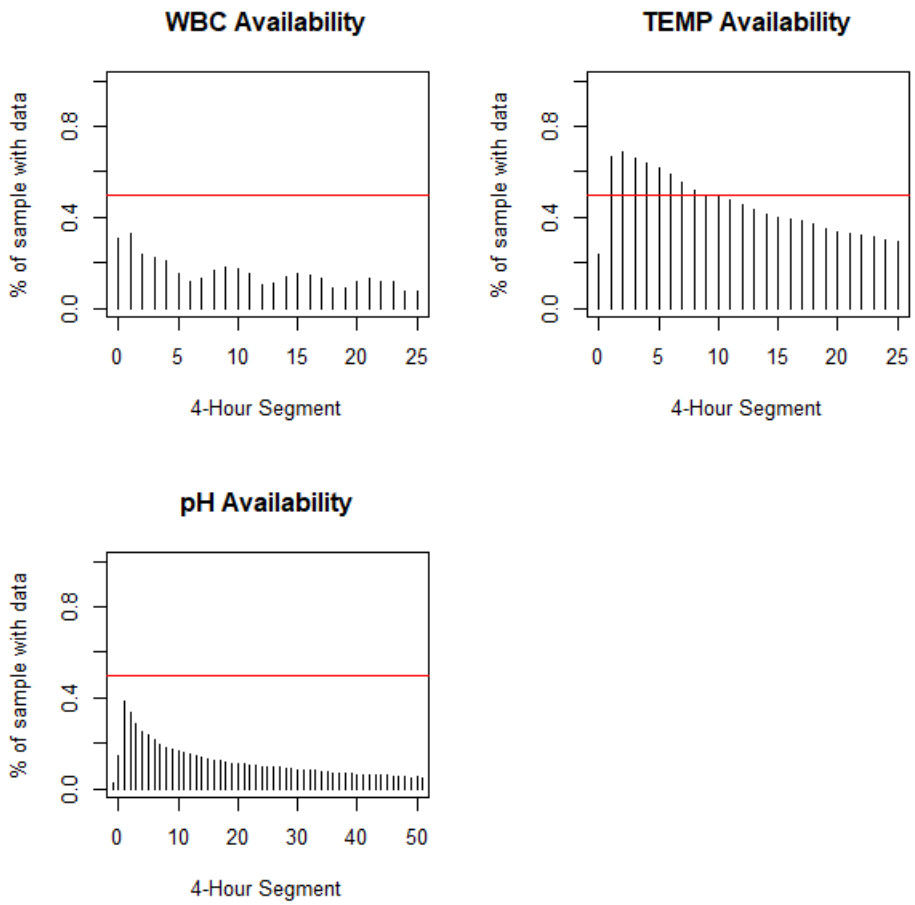

Figure 2: Plots showing the proportion of the cohort built by Gordillo-Maranon (N = 36,198) that had measurements for WBC, TEMP, and pH over each 4-hour interval. A horizontal line indicating 50% is superimposed on each plot.

2.2 Assessing Prediction Performance

2.2.1 Data Partitioning and Classifications

Generalisability is a good quality to have in models, and a way to quantify generalisability is by assessing how well models predict data points that were not used in their development. In this study, these assessments were done using k-fold cross-validation. Due to the low number of cases in the cohort, stratified sampling was used in all cases, such that each data partition had approximately the same proportion of cases and controls.

Further, the same folds were used among different model specifications within the same method. That is, so that model comparison would be possible, all logistic models made use of the same ten folds for cross validation, while all joint models made use of the same five folds for cross validation (less folds were used since each run of the model was computationally intensive). Comparisons are performed using the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, which are built by categorising the predicted probabilities based on whether they are smaller or larger than various thresholds.

In particular, let $p \in [0, 1]$ be a fixed probability, and p_i the predicted probability that subject i is a case. Subject i is classified as a case if $p_i >= p$, and as a control otherwise. This gives rise to the following classification concepts (for some specified threshold p):

- True Positive (TP): Number of cases classified as cases
- False Negative (FN): Number of cases classified as controls

Variable	Controls (n = 13,944)	Cases (n = 169)	Entire Cohort (n = 14,113)
Age (mean ± sd)	63.98 ± 16.09	65.57 ± 17	64 ± 16.1
Male (% admission)	58%	56%	58%
Female (% admission)	42%	44%	42%
Ethnicity (% admission)			
White	70%	68%	70%
Asian	3%	3%	2%
Black	7%	7&	7%
Hispanic	3%	3%	3%
Other	17%	20%	17%
Service Category (% admission)			
Medical - general service for internal medicine	32%	37%	32%
Cardiac Medical - non-surgical, cardiac related	16%	12%	16%
Neurologic Medical - non-surgical, neuro. related	5%	8%	5%
Neurologic Surgical	7%	9%	7%
Cardiac Surgical	13%	5%	13%
Surgery - general unclassified surgery	9%	8%	9%
Trauma - injury caused by external physical harm	7%	13%	7%
Others - all other service categories	10%	8%	10%

Table 3: Summary of cohort demographic.

- True Negative (TN): Number of controls classified as controls
- False Positive (FP): Number of controls classified as cases

2.2.2 Receiver Operating Characteristic Curve

One of the ways by which predictive performance is evaluated is through the Receiver Operating Characteristic (ROC) curve. The ROC curve of a model plots 1-specificity against sensitivity across all possible classifying thresholds. Sensitivity is defined as the ability of a model to correctly identify cases, whereas specificity is its ability to correctly identify controls.

$$sensitivity = \frac{TP}{TP + FN} \tag{2.1}$$

$$specificity = \frac{TN}{TN + FP} \tag{2.2}$$

The most common way of comparing classifier models is by comparing the area under their respective ROC curves, often referred to as their ROC-AUC. It is trivial to show that a perfect classifier (one able to discriminate perfectly between cases and controls) would have a sensitivity and specificity of 1, resulting in a constant line for an ROC with an ROC-AUC of 1.

Consider a completely random classifier classifies a subject as a case with probability p. Suppose that there are N samples, of which c are cases. A random classifier that will produce an expected result of pc true positives, $(1-p)(N-c)$ true negatives, $p(N-c)$ false positives, and $(1-p)c$ false negatives. Substituting these values into equations 2.1 and 2.2 yields

$$sensitivity = \frac{pc}{pc + (1-p)c} = p$$

$$specificity = \frac{(1-p)(N-c)}{(1-p)(N-c) + (N-c)p} = 1 - p$$

This suggests that the random classifier ROC would simply be the line with $sensitivity =$

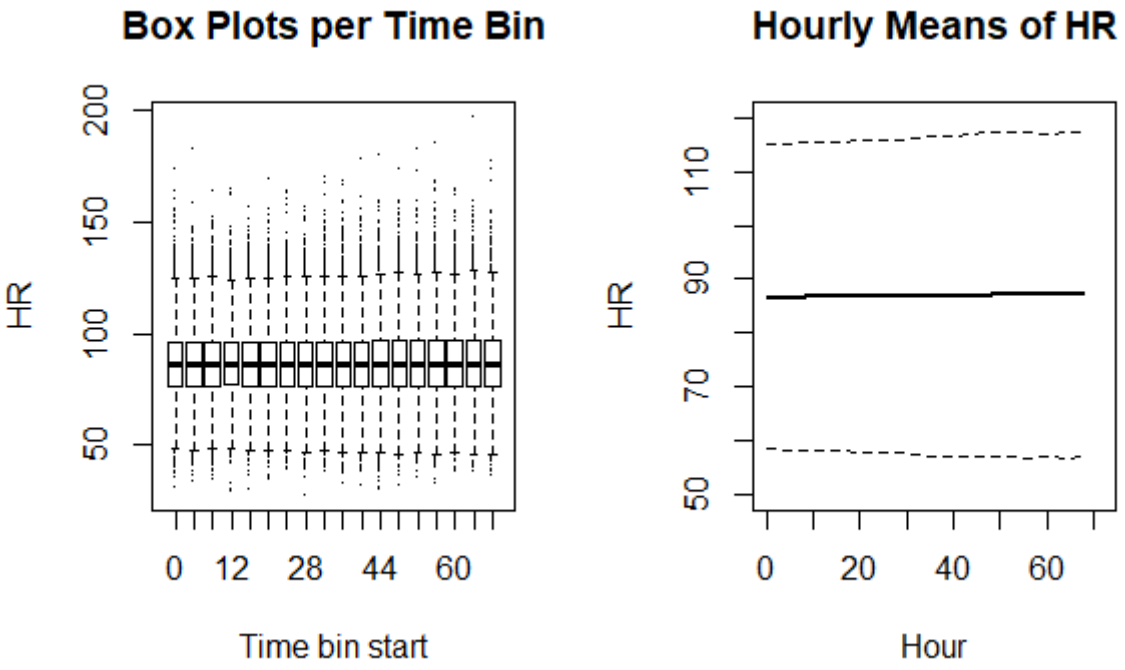

Figure 3: Descriptive plots of the heart rate of the entire cohort over time. Left: Box plots per time-bin showing spread of the observations. Right: Mean heart rate with confidence interval built using the normal distribution.

$1 - specificity$. This is the diagonal line, with an AUC of 0.5.

Thus, based on ROC-AUC alone, a good classifier model would have an AUC as close to 1 as possible, and higher than 0.5 at the very least.

2.2.3 Precision-Recall Curve

An issue with the ROC-AUC analysis is that it does not perform well when the data used suffers from class imbalance, which occurs when the number of either the cases or controls is much larger than that of the other. This is of particular interest in the study since the MIMIC-III database has been shown to have a very low proportion of cases. An ROC-AUC analysis for MIMIC-III runs the risk of being overly optimistic due to the way specificity is defined and the large amounts of true negatives the dataset will tend to produce [3].

An alternative validation metric that is more appropriate for unbalanced datasets is the Precision-Recall curve, which plots a model's recall against its precision across all classifying thresholds. Note that recall is the same as sensitivity, and is thus computed using 2.1. Precision, on the other hand, is the proportion of true cases among those identified as cases. It is sometimes also referred to as positive predictive value (PPV). Mathematically,

$$precision = \frac{TP}{TP + FP} \tag{2.3}$$

It is trivial to show that a perfect classifier would have a precision and recall of 1, and a PR-AUC of 1. For a random classifier, using the same notation as in 2.2.2,

$$recall = sensitivity = p$$
$$precision = \frac{pc}{pc + p(N - c)} = \frac{c}{N}$$

This indicates that a random classifier would have a precision equal to the prevalence of the

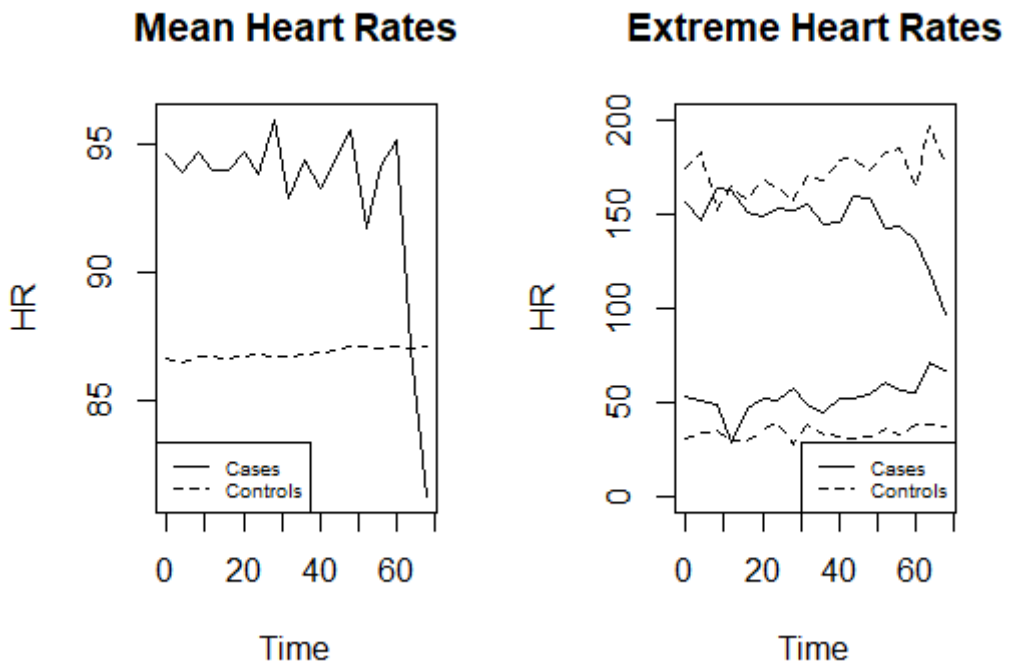

Figure 4: Descriptive plots of the heart rate of the cohort, stratified based on case-control status. Left: Average heart rates. Right: Minimum and maximum heart rates for each strata over time.

dataset, that is, the proportion of the entire cohort that are cases. The PR curve would then be a horizontal line with height equal to the prevalence, resulting in a random classifier having PR-AUC equal to its prevalence.

Thus, a good classifier would have a PR-AUC as close to 1 as possible and larger than the prevalence of the dataset at the very least.

3 Logistic Modelling

3.1 Formulation

Multiple logistic regression is a procedure that aims to describe some outcome as a linear combination of predictors. It is most often used when the outcome of interest is binary, such as whether a patient died within 72 hours of ICU admission. Let the vector y contain the binary outcome that takes value 1 on cases and 0 on controls. Let X be the design matrix containing the measurements for the predictors. Let p denote the conditional probability of being a case given some collection of predictor measurements, and β be the collection of regression parameters to be estimated.

The linear relationship defined by logistic regression is of the form

$$\log \frac{p}{1-p} = \beta X \tag{3.1}$$

and β is estimated most commonly using maximum likelihood.

The left hand side of equation 3.1 is the logistic transformation of the conditional probability p. The term $\frac{p}{1-p}$ is called the odds ratio, since it gives insight into how much more (or less) likely being a case is compared to being a control. The natural logarithm of the odds ratio is called the logit or log odds ratio.

Note that when a subject is equally likely to be a case as they are to be a control, the odds ratio is 1, and the log odds ratio is 0. The sign of the log odds ratio shows whether the subject is more likely to be a case (positive log odds) or a control (negative log odds), while the magnitude indicates how much more (or less) likely the subject is to be a case (dependent on sign).

Solving for p allows the computation of the probability of some subject being a case given their predictor measurements. Specifically:

$$p = \frac{e^{\beta X}}{1+e^{\beta X}} = \frac{1}{1+e^{-\beta X}} \tag{3.2}$$

Computing p for a subject whose predictors are all 0 (whether the predictors are categorical or continuous) gives the probability of being a case for a baseline individual. For some predictor q with estimated parameter β_q, the interpretation is that a unit increase in predictor q results in a β_q increase in the log odds of being a case, with respect to the baseline. This is equivalent to an increase in the odds ratio of e^{β_q}.

Consider the simple logistic model using some continuous predictor x.

$$\log \frac{p}{1-p} = 0 + 2x$$

A basis individual would be someone for whom $x = 0$. Such an individual would have a log odds ratio of 0, indicating that they are as likely to be a case as they are to be a control. Suppose another subject has $x = 1$. The log odds ratio for this subject would be 2, corresponding to an odds ratio of $e^2 \approx 7.39$, suggesting that this new subject is 7.39 times more likely to be a case than a control.

Upon estimating the β, predictions are made by plugging in new X values and, for the purposes of constructing ROC and PR curves, computing p for each subject (using equation 3.2).

3.2 Results

The questions that the analyses in this section hoped to answer were:

- How would adding data for longer periods of time (i.e. adding more time bins into the possible predictors) affect the predictive power of the chosen models?

- Which static and longitudinal variables improved the predictive performance of models?

The initial approach to answering the questions was to specify models of increasing complexity. In particular, the first model fit was one that used none of the longitudinal variables, relying only on the static information. Then, a second model was fit that considered both the static variables, as well as all the longitudinal measurements in the first bin (first 4 hours of ICU admission). The third through fifth models would add the second, third, and fourth bin iteratively. At each iteration, all variables from the included bins are passed through the selection algorithm. This approach is referred to as the "cumulative approach", and the results are discussed in 3.2.1.

An alternative approach to the problems stated above was to verify if each individual bin had similar effects on the predictive power of the models produced. For example, it was checked whether a model that contained only static data and longitudinal measurements from the first bin would produce similar metrics as a model containing only static data and longitudinal measurements from the second bin. This approach is referred to as the "individual bin approach", and the results are discussed in 3.2.2.

Section 3.2.3 discusses a modification in how blood oxygen saturation (O2S) was specified in the data, as well as cumulative models built using the revised dataset.

The selection of which variables to include in each model was done using forward stepwise selection, with the Akaike Information Criteria (AIC) being the main metric of interest[6].

3.2.1 Cumulative Approach

Table 4 summarises the results from the model validations of those built using the cumulative approach. Note that in Table 4, the notation S + $B_1...n$ indicates the model using longitudinal measurements from bin 1, bin 2, ... , bin n, alongside all static variables.

	ROC		PR	
Model	**mean AUC**	**St. dev.**	**mean AUC**	**St. dev.**
Static Only (S)	0.5938	0.0465	0.0255	0.0269
S + Bin 1 (B_1)	0.7168	0.0415	0.0691	0.0416
S + $B_1...2$	0.7302	0.0519	0.0776	0.0455
S + $B_1...3$	0.7291	0.0417	0.0763	0.0389
S + $B_1...4$	0.7394	0.0610	0.0781	0.0361

Table 4: Ten-fold cross-validation metrics from the cumulative approach.

The results for each model, shown in Table 4, make use of ten-fold cross-validation, with figures for both the mean AUC of the ten folds, as well as the standard deviation (St. dev. in the table) of the obtained AUCs. While the mean AUC aimed to show the overall predictive ability of each model, the standard deviation quantifies how differently the models behave depending on the data partitions used to train and validate. An ideal model would have a high PR-AUC, indicating overall superior predictive power, as well as low standard deviation, indicating generalisability.

An immediate observation was that while the mean PR-AUCs of the models incorporating longitudinal measurements were higher than the model that included only static data, the results were still low, all being below 0.10. A possible explanation for this was the high imbalance present in the data. As shown in Section 2.2.3, a random classifier for the data used would have a PR-AUC of 0.0120. This indicated that while the predictive capabilities of all the models tested were quite weak, they remained improvements over the random classifier.

Another thing to note was that all the models that incorporated longitudinal measurements had mean PR-AUCs within one standard deviation of each other, suggesting that no substantial

[6]For more details on the method, visit `advstats.psychstat.org/book/mregression/selection.php`. For more details on the stepAIC() function used to perform the selection in R, see `cran.r-project.org/web/packages/MASS/MASS.pdf`.

predictive improvements had been obtained by using more time bins. Combined with the observation that the largest improvement in PR-AUC was when the first bin of data was added, this gave rise to the question of whether the large improvement was due to the information in the first bin or if it was due to the inclusion of any of the variables at all. This issue was explored by building and comparing models that used only one bin of longitudinal measurements at a time: the individual bin approach.

3.2.2 Individual Bin Approach

Table 5 summarises the results from the individual bin approach. An additional first row is included containing the validation metrics of the model built using only static data for reference.

	Ten-fold ROC		Ten-fold PR	
Model	**mean AUC**	**St. dev.**	**mean AUC**	**St. dev.**
S	0.5938	0.0465	0.0255	0.0269
S + B_1	0.7168	0.0415	0.0691	0.0416
S + B_2	0.7190	0.0615	0.0747	0.0414
S + B_3	0.7083	0.0490	0.0763	0.0479
S + B_4	0.7395	0.0677	0.0621	0.0263

Table 5: Ten-fold cross-validation metrics from the individual bin approach.

The inclusion of any time bin caused an improvement in the cross-validation PR-AUCs. The mean PR-AUCs of the models remained approximately the same. Similarly, the predictive power of the individual bin models were also within the same ranges as those from the cumulative approach. This indicated that it was the presence of any information on the longitudinal variables that caused the PR-AUC improvements, rather than the first bin specifically. Further, the results suggested that for the data and outcome used in this study, knowing a patient's physiological measurements for only the most recent bin was just as good as knowing all their measurements for the last few bins.

3.2.3 Categorical Blood Oxygen Saturation (O2S)

For both the initial and alternative approaches, all O2S variables were removed from consideration in the models by the scripts used to fit the data. A reasonable explanation for the exclusion of the variable is that due to most of its values being at or just around 100, all O2S variables exhibited high collinearity with the intercepts. To overcome the issue, models were built using data where the O2S variable was categorical rather than continuous. In particular, all the O2S measurements were classified as either 'OK' or 'LOW' depending on where the values fell relative to a threshold.

Literature indicates that a common threshold for detecting abnormal O2S is 95% [18]. Consultations with healthcare professionals suggested that values below 90%-93% were cause for alarm. For the study, a threshold of 90% was chosen, such that measurements <= 90 were categorised as 'LOW', while all others were categorised as 'OK'. 90% was chosen since, given the context that lower O2S was expected to be positively associated with the outcome, a lower threshold would produce fewer false positives than a higher threshold, and might exhibit more predictive power.

Table 6 summarises the results from models built using the categorised O2S in tandem with the cumulative approach. Again, a row containing the results from the static only model was included for reference. The notation S + $B_1...n$ is used identically as in Table 4.

Compared to all the previous models built, those that used a categorical O2S showed marginally better PR-AUCs. However, all the figures were still well within the standard deviations of the previous models, so definitive improvement could not be concluded. Nonetheless, the results seem to suggest that the inclusion of O2S as a predictive variable was able to improve

	Ten-fold ROC		Ten-fold PR	
Model	**mean AUC**	**St. dev.**	**mean AUC**	**St. dev.**
S	0.5938	0.0465	0.0255	0.0269
S + B_1	0.7287	0.0381	0.0825	0.0461
S + B_1...2	0.7410	0.0591	0.0866	0.0479
S + B_1...3	0.7411	0.0457	0.1021	0.0612
S + B_1...4	0.7524	0.0640	0.1014	0.0567

Table 6: Ten-fold cross-validation metrics from the categorised O2S models.

the discriminatory ability of the models. It must be noted, however, that these improvements may be artificial since the threshold put in place to divide the samples into 'LOW' and 'OK' O2S was chosen as 90% in the interest of being able to determine subjects who suffered from significantly deteriorated O2S. A higher specified threshold might not have produced similar results.

3.3 Final Logistic Model

Selection of a 'best' logistic model in terms of predictive power was difficult due to the fact that all models tested exhibited similar mean PR-AUCs, which were all low. In such a case, whichever model was smallest would be preferred. Of all the models fitted, the models which used only bins 2, 3, and 4 (individually) were the smallest, each using 5 variables. Of those models, the individual bin model using only the third bin of data had the highest mean PR-AUC. The model is summarised in Table 7. The coefficients presented were fitted using 80% of the cohort; each model fitted during the cross validation may have had slightly different coefficients.

Variable	**Estimate**	**Std. Error**	P-value
Intercept	5.4539	2.9369	0.0633
Age	0.0231	0.0059	0.0001
Service Category			
Medical	0.4282	0.3362	0.2029
Cardiac Medical	-0.1276	0.3955	0.7469
Neuro. Medical	1.3635	0.4669	0.0035
Neuro. Surgical	1.3924	0.4380	0.0015
Cardiac Surgical	-0.9617	0.5457	0.0780
Surgery	0.4586	0.4169	0.2713
Trauma	1.4314	0.4029	0.0004
Longitudinal Variables			
HR	0.0308	0.0057	0.0000
RR	-0.0913	0.0292	0.0018
SBP	-0.0366	0.0056	0.0000

Table 7: Summary of logistic model incorporating static patient data and longitudinal variables measured within 8 to 12 hours of ICU admission.

The first column of Table 7 lists the variables that were included in the model by the stepwise selection algorithm. The column named 'Estimate' contains the mean parameter estimates of the fitted model in the log-odds scale, while the entries on the column named 'Std. Error' show the standard deviations of the estimated parameters. The last column, entitled 'P-value' contains p-values computed using the standard normal distribution, testing the significance of the estimated coefficients against a null hypothesis that the real parameter coefficient is 0.

For the model specified in Table 7, the baseline individual was described as a 64-year old white female that had been admitted to the ICU for general medical purposes.

Assuming a conventional significance level of 5%, the static variables that were found to have been significant were age, service-neuro. medical, service-neuro. surgical, and service-trauma. In particular, admission for neuro. medical, neuro. surgical, and trauma purposes were each associated with approximately 400% increased odds of in-ICU fatality. Age was also positively related with risk of in-ICU death, with each additional year of age above the cohort average of 64 being associated with a 2.34% increase in odds.

Among the longitudinal variables, only heart rate was found to be positively associated with risk of in-ICU fatality, with a unit increase in heart rate increasing the odds by 3.12%. RR and SBP were negatively associated with the risk of in-ICU death, with a unit increase in RR and SBP resulting in an 8.63% and 3.6% decrease in the odds ratio respectively.

Exponentiating the intercept coefficient of Table 7 gives the risk for the baseline individual described above. Such a baseline was found to have an extremely high odds ratio in favour of in-ICU fatality, but any interpretation of this result would be problematic since it would describe an individual that is not physiologically viable. Specifically, outside of the demographic information (64-years old, white, female, general medical services), the baseline individual was also assumed to have HR, SBP, and RR measurements of 0.

4 Joint Modelling

4.1 Formulation

The joint model is a framework that incorporates the feature of survival models and mixed-effects models to combine longitudinal and time-to-event information. Prior to discussing the specifics of the joint model, a brief overview of its components, the survival model and the linear mixed-effects model, is presented. The discussion and notation used are based on Rizopoulos, 2012 [13].

4.1.1 Linear Mixed-Effect Model

At its core, the linear mixed-effect model uses the idea that for a given longitudinal variable, the measurements for a subject will have some unique average which is different from those of other subjects. Let subject i have longitudinal measurements y_i. Further, let X_i and Z_i be design matrices containing information on designated fixed and random predictors respectively. Let β and b_i denote the fixed-effect and random-effect regression coefficients respectively. The random-effect regression coefficient b_i is defined as following a multivariate normal distribution with mean 0 and covariance matrix D. Let ϵ_i be some random error following the multivariate normal distribution with mean 0 and covariance matrix $\sigma^2 I_{n_i}$, where I_{n_i} is the identity matrix of dimension n_i. The linear mixed-effects model is defined as

$$\begin{cases} y_i = X_i\beta + Z_i b_i + \epsilon_i \\ b_i \sim \mathcal{N}(0, D) \\ \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}) \end{cases} \tag{4.1}$$

4.1.2 Survival Model

Survival models are focused on using covariates to predict the time until some specified event takes place. A particularly popular class of survival models is the proportional hazard model, which describes covariates as having multiplicative effects on the baseline hazard rate for the event of interest [2]. Let $h_i(t|w_i)$ denote the hazard for an event at time t for a subject with the covariate measurements contained in w_i. The proportional hazard model is defined as

$$h_i(t\,|w_i) = h_0(t)\exp(\gamma^\intercal w_i) \tag{4.2}$$

where γ is the vector of regression coefficients, and the function $h_0(t)$ is the baseline risk function which indicates the risk hazard at time t of a subject for whom $\gamma^\intercal w_i = 0$.

One of the main strengths of survival models is their ability to take into consideration the fact that the event time may not be observed for all subjects. The mechanism that survival models use to do this is called censoring, and it allows for incomplete data in the sense that not all subjects are required to have information for the same lengths of time, unlike in logistic regression. There are different kinds of censoring based on when the unobserved event is known to have taken place. In this study, only right censoring is relevant, where the event is known only to take place after a certain time point, such as for subjects who had not died in the ICU after 72 hours of admission.

The utilisation of censored data in survival models is made possible through specifying distributional assumptions for the event times. The information of this distribution is the major component of the baseline risk function h_0, which takes on different functional forms depending on the distribution specified.

4.1.3 Joint Model

The joint model is specified in two parts: a longitudinal submodel, and a survival submodel. The longitudinal submodel is a linear mixed effect model formulated as

$$\begin{cases} y_i(t) = m_i(t) + \epsilon_i(t) \\ m_i(t) = x_i^\intercal(t)\beta + z_i^\intercal(t)b_i \\ b_i \sim \mathcal{N}(0, D), \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \end{cases} \tag{4.3}$$

In equation 4.3, $y_i(t)$ denotes the observed measurement for some longitudinal outcome that was recorded at time t with error $\epsilon_i(t)$ following a normal distribution with mean 0 and variance σ^2 , $m_i(t)$ denotes the true value of the longitudinal outcome at time t, expressed as a mixed effect model with fixed and random design vectors $x_i(t)$ and $z_i(t)$. β and b_i are the associated fixed and random effects, with b_i following a multinomial normal distribution with mean 0 and covariance matrix D.

The longitudinal submodel assumes that the error terms $\epsilon_i(t)$ are independent both among themselves, as well as of the random effects b_i.

Expectedly, the longitudinal submodel specified in 4.3 is very similar to the specification of 4.1 except for a few key differenced. First, the joint modelling framework assumes that instead of having static design matrices X_i and Z_i, there exist time-dependent design vectors $x_i(t)$ and $z_i(t)$. Second, the model explicitly specifies the component $m_i(t)$, used to denote the true correct value of the longitudinal outcome. It is this $m_i(t)$ that is used as one of the inputs in the relative risk submodel, formulated as

$$h_i(t \,|\mathcal{M}_i(t), w_i) = h_0(t) \exp[\gamma^\intercal w_i + \alpha m_i(t)], \quad t > 0 \tag{4.4}$$

In equation 4.4, $h_0(t)$ denotes the baseline risk function, $\mathcal{M}_i(t)$ contains all the true unobserved values for the longitudinal variables until time t, w_i is a vector containing information on the non-longitudinal covariates, used with regression coefficient γ, and α is the regression coefficient associated with the longitudinal outcome described by the longitudinal submodel. Note that covariates used need not be exclusive between the submodels. Variables can appear both in the design vectors $x_i(t)$ of the longitudinal process, as well as in w_i from the survival process.

Taking the natural logarithm of equation 4.4 gives a form that allows straightforward interpretations. In particular,

$$\log h_i(t \,|\mathcal{M}_i(t), w_i) = \log h_0(t) + \gamma^\intercal w_i + \alpha m_i(t) \tag{4.5}$$

Specifically, for some covariate w_j with coefficient γ_j, $\exp(\gamma_j)$ is the proportion of the hazard when w_j increases by one unit over the hazard rate at its current value. The same is true for the longitudinal outcome m_i with coefficient α.

While Rizopoulos [13] suggests that it is often preferable to leave the baseline hazard function completely unspecified when building relative risk models, he advises against it in the framework of joint modelling, citing that the resulting models could suffer from underestimating the standard errors of the parameters. As a solution, two simple yet flexible parametric specifications are suggested for the baseline risk function: piece-wise constant, and regression-splines. While both are valid choices, the study chose to use the regression splines approach, as it was more consistent with the assumptions made on the behaviour of the longitudinal measurements (i.e. they behave continuously).

The regression splines approach expands the log baseline risk function into cubic splines, specified as

$$\log h_0(t) = \kappa_0 + \sum_{d=1}^{m} \kappa_d B_d(t, q)$$

where B is the collection of B-splines basis functions of degree q, $m = \ddot{m} + q - 1$, where $\ddot{m}$ is the number of interior knots by which the splines are fit, and where $\kappa^\intercal$ is the $(m+1)$-dimensional vector containing the spline coefficients (including κ_0). The number of knots described the number of splines used to fit the function, and is often kept generally low to avoid over-fitting.

Under a frequentist framework, the estimation of the joint model is performed through the maximisation of log-likelihood functions, making use purely of the data. Under a Bayesian framework, the distributions of the joint model parameters are estimated using a combination of the likelihood from the data, as well as some specified priors. For this study, the JMBayes[7] package was used to fit the Bayesian joint models. The default priors utilised by the package were $\mathcal{N}(0, 1000)$ for fixed effects, and $\Gamma^{-1}(0.1, 0.1)$ for scale parameters[8].

After having fitted all the parameters to a joint model, it is possible to predict a new patient's survival up to some time $u > t$ conditional on surviving up to time t. In particular, the conditional survival probability $\pi(u\,|t)$ can be expressed as

$$\pi_i(u\,|t) = Pr(T_i^* \geq u\,|T_i^* > t, Y_i(t), w_i, X_n; \theta^*), \tag{4.6}$$

where T_i^* denotes the true event time, $Y_i(t)$ is the set of longitudinal measurements for the new patient up until time t, w_i is the new patient's non-longitudinal covariate measurements, X_n is the information from the sample used to fit the joint model with true parameter values θ^*. In this study, as the main outcome was 72-hour in-ICU death, the t of interest is $t = 72$.

Under the assumption that both the longitudinal and survival processes are dependent only on previous information rather than future longitudinal measurements, the approximation of π_i is performed either using Monte Carlo methods or, when computational limitations are tight, first-order estimates. From these predicted survival probabilities, ROC and PR curves can be built using on 1-π_i.

4.2 Results

This section of the analyses aims to answer the following questions:

- How does the predictive power of joint models change as more longitudinal information is included?
- Taken either individually or in groups, which longitudinal variable(s) exhibited predictive capabilities for the outcome of interest?

In Section 4.2.1, frequentist joint models are fit on the data for the entire cohort using the R package JM[9]. Since subjects no longer needed to have complete sets of observations in the joint modelling framework, all observations up to 72 hours after ICU admission were used in the model building. However, measurements were still aggregated into 4-hour bins using the mean to avoid biases that could arise due to some patients having more measurements than others in small periods of time. The software used provided support only for single longitudinal outcome models, so these were what the analyses focused on. All cross-validations performed used only five folds due to the heavy computational demands of the software used.

In Section 4.2.2, a Bayesian joint model was fit on the two most predictive variables, as determined in Section 4.2.1 using the R package JMBayes. However, JMBayes, along with the similar library rstanarm[10] were unable to process the data for the cohort used in all previous models due to its size. As such, models were built using only a subsample of the data. While there are many advanced methods presented in literature for undersampling [4], a simple random sample method was utilised, building a balanced subsample that utilised all 169 cases as well as 169 randomly selected controls. In the spirit of cross-validation, five groups of controls were sampled and used to build five subsamples of the cohort.

[7] For more details, visit `cran.r-project.org/web/packages/JMbayes/JMbayes.pdf`.
[8] For details on default priors used, see `drizopoulos.com/vignettes/multivariate%20joint%20models`.
[9] For more details, visit `cran.r-project.org/web/packages/JM/JM.pdf`.
[10] For more details, visit `cran.r-project.org/web/packages/rstanarm/rstanarm.pdf`.

4.2.1 Frequentist Models

Table 8 presents a summary of the cross-validation metrics from the joint models built using all the static variables as well as a single longitudinal outcome each. The joint models were fit under the specification that the longitudinal variables were dependent only on time, and the longitudinal outcomes were not transformed.

	Joint Models		Logistic Models	
Variable	**mean PR-AUC**	**St. Dev.**	**mean PR-AUC**	**St. Dev.**
HR	0.1319	0.0433	0.0413	0.0273
SBP	0.1283	0.0441	0.0620	0.0312
DBP	0.0819	0.0306	0.0306	0.0443
RR	0.1138	0.0299	0.0211	0.0199
O2S	0.0829	0.0163	0.0211	0.0199

Table 8: Cross-validation metrics on single longitudinal outcome joint models, compared with logistic models using only one of the longitudinal variables.

Aside from the column indicating the longitudinal outcome included in each model, the first two columns of Table 8 contain the mean PR-AUC and the standard deviation of the AUCs obtained in the five-fold cross-validation performed. The last two columns contained the same metrics, except they were obtained from ten-fold cross-validations performed on logistic models using only the static information as well as the first four binds of longitudinal measurements for one variable at a time. While the resulting figures from the two different techniques are were not directly comparable due to differences in cross-validation partitions and amount of information used, a side by side look gives insight into how predictive powers can differ based on how much longitudinal information was used. In particular, recall that due to logistic regression requiring complete data, only the first 16 hours of information were incorporated, whereas the single longitudinal joint models were able to make use of any and all information gathered during the first 72 hours of admission.

With respect to comparing the logistic models with the joint models, the results are quite clear. For all longitudinal variables, the joint models exhibited higher PR-AUCs than their logistic counterpart. It is notable that the conclusion is still true even when the intervals built using mean $\pm$ st. dev. are considered. While the difference between the PR-AUCs of the two methods may not be statistically significant if tested formally[11], such tests would be unreliable due to the very small sample size (5 sample AUCs) in question.

Comparison between the joint models, however, was not as straightforward. Even though all the joint models were improvements over their logistic counterparts, their overall predictive power was still very low, all being below 0.15. Further, the three variables that exhibited the highest mean AUCs were all within one standard deviation of each other. Since each model used the same number of variables (all static + one longitudinal), no parsimony argument could be made to help choose a model. Thus, the model with the highest PR-AUC (HR) was chosen as the model to focus and build extensions on.

The joint model using static information and heart rate measurements is summarised in Table 9. Note that the coefficients estimated were based on an 80% subset of the cohort; slight differences are expected in the coefficients for the models built during the cross-validations.

The first few rows contain information on the longitudinal submodel used to fit the heart rate data, while the rest of the rows contain information on the survival submodel.

Note that in the longitudinal submodel, only the intercept was significant based on p-value. In fact, while the p-value of time was much larger than the conventional confidence level of 0.05, its effect size was also very small at -0.0011. This suggested that, on the whole, there was not much trend over time in the heart rate measurements observed, which agrees with observations in 2.1.2 that discussed the relative flatness of the data.

[11]For example, constructing a basic 95% confidence interval with the factor 1.96 on the standard deviation would result in the intervals built for the joint model metrics and logistic model metrics to intersect.

Longitudinal Process			
Variable	**Estimate**	**Std. Error**	**P-Value**
Intercept	86.8439	0.1264	< 0.0001
Time	-0.0011	0.0010	0.2950
Survival Process			
Variable	**Estimate**	**Std. Error**	**P-Value**
Age	0.0241	0.0058	< 0.0001
Heart Rate	0.0448	0.0068	< 0.0001

Table 9: Summary of joint model using static information and heart rate as the only longitudinal outcome.

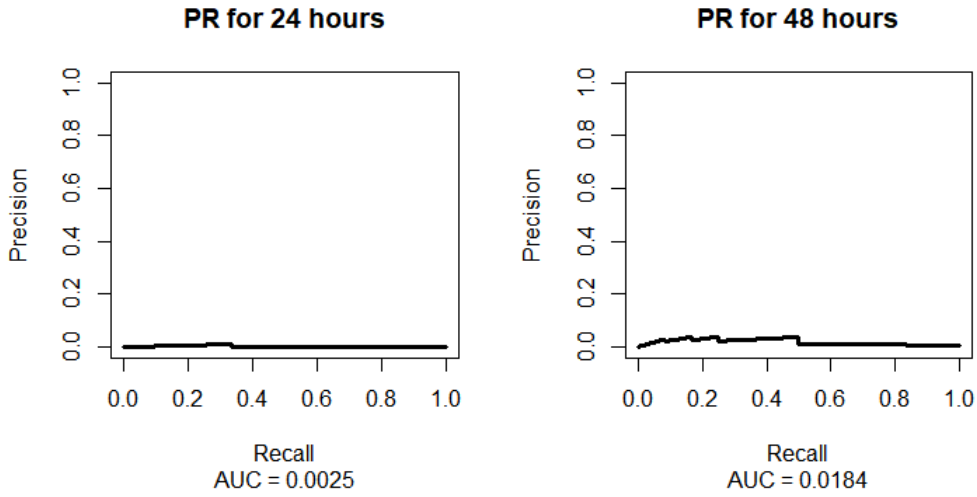

Figure 5: PR curves of the joint model using static data and heart rate to predict in-ICU fatality within 24 and 48 hours of admission. The curves shown are from the first fold only.

In the survival process, both age and heart rate are positively associated with hazard. In particular, a unit increase in age (above the cohort mean of 64) was associated with a 1.02-fold increase in the risk of in-ICU death, while every unit increase in heart rate was associated with a 1.05-fold increase.

Extension: Predictions over time

An advantage of joint models over logistic models is their ability to model survival probabilities as evolving over time. Whereas predictions in logistic models can only specify whether or not a subject would be a case over the entire period of interest, a joint model is able to make predictions about a patient's survival up to different time points. Figure 5 shows the PR-curves of the model using only HR as a longitudinal outcome when it attempts to predict in-ICU death within 24 and 48 hours of admission.

The probable reason for the AUCs being very low was the extremely low case proportion. For the testing sample (20% of the data, 2,820 subjects), only 3 died within 24 hours of admission, and only 12 for the first 48.

Another way to take advantage of the ability of joint modelling to produce predictions that change over time is by seeing how survival trajectories change as more longitudinal data is collected. A subject is chosen arbitrarily, and Figure 6 shows plots of the subject's survival trajectory given all heart rate measurements up to 0, 16, 24, and 48 hours of admission (0 hours denotes the first 4 hours of admission into the ICU).

The subject was chosen since they showed some amount of an increasing trend compared to the rest of the cohort which had more oscillating measurements. However, as was the case with

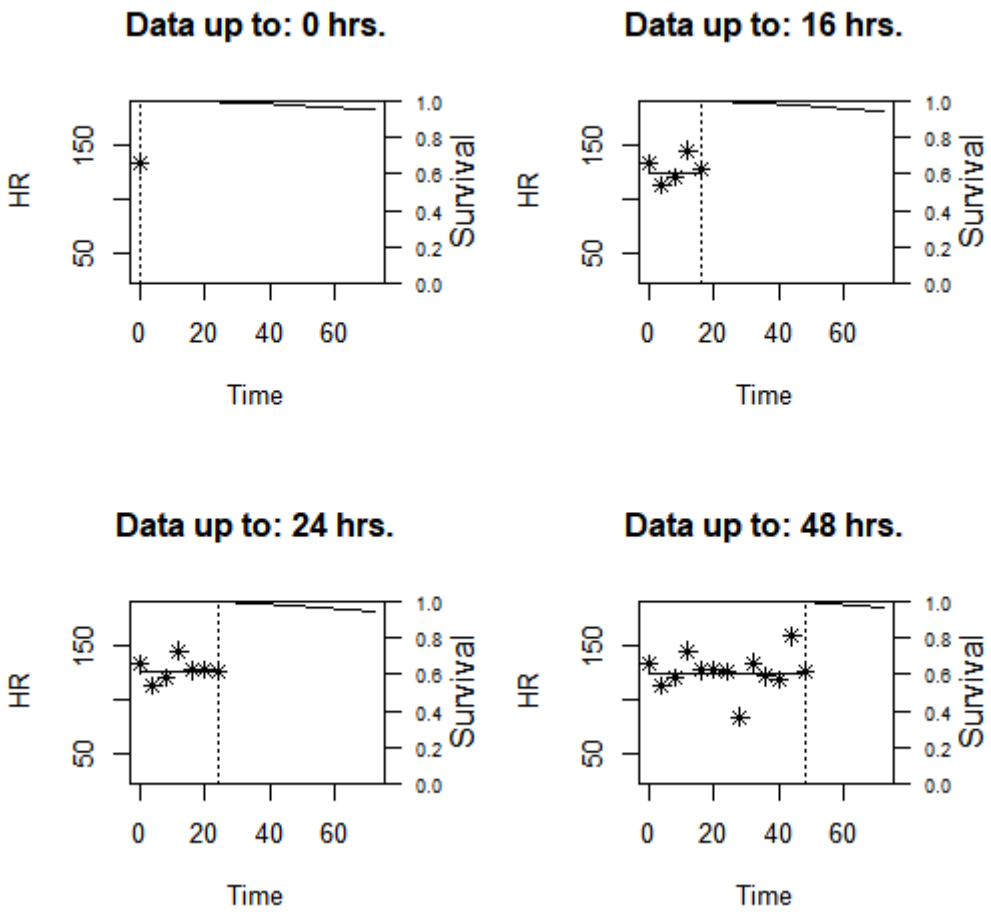

Figure 6: Dynamic survival probabilities for patient identified by icustay_id 295467 at different points in their ICU stay. Left-hand portion of the plots, before the vertical line, show the evolution of the longitudinal outcome (HR), while the right-hand side plots their survival trajectories.

most of the data, the trend remained minimal and the measurements for the subject remained relatively flat. The consequences of the flatness can be seen in the plots in Figure 6, where no significant changes can be seen in the survival trajectory as more longitudinal measurements were taken.

Alternative longitudinal model specifications

A natural modification to the basic joint model was to specify a linear mixed effect model that described the longitudinal outcome as dependent not only on time, but also on age as part of the fixed effects. The model was fit and is summarised in Table 10.

Longitudinal Process			
Variable	**Estimate**	**Std. Error**	**P-Value**
Intercept	86.8374	0.1229	< 0.0001
Time	-0.0010	0.0010	0.3056
Age	-0.1930	0.0075	< 0.0001
Survival Process			
Variable	**Estimate**	**Std. Error**	**P-Value**
Age	0.0240	0.0058	< 0.0001
Heart Rate	0.0452	0.0068	< 0.0001

Table 10: Summary of joint model using age as a covariate for the longitudinal process.

The only notable result from the model described in Table 10 was that for the longitudinal process, age did turn out to be a significant variable, albeit with a small effect size. However, when compared with the results summarised in Table 9, all the other entries were statistically identical, with differences generally being within margins of ± 0.01. This observation suggested that while the longitudinal process could have been slightly more accurate, the overall predictive power would not have changed.

Another extension to the basic joint model is that of stratification. In particular, assume that the baseline hazard rate differs between men and women. A formal procedure that could be used to check if stratification improved model performance was by using Wald tests comparing the likelihoods of the stratified and unstratified models. Once performed, the test yielded a p-value of 0.9929, preventing the rejection of the null hypothesis that the baseline risk functions were the same across strata[12].

Joint model of a subsample

A final exploration was performed by seeing how a joint model specified on a subsample of the cohort would perform in comparison to the one fitted using the entire cohort as discussed in the previous sections. A subsample was generated by taking all 169 subjects that were cases and merging them with a random sample of 169 controls, creating a balanced subsample of the data. A joint model was fit on this subsample, and the resulting model is summarised in Table 11. Note that following the results of the extensions presented, age was specified as a covariate of heart rate in the joint model specified.

Figure 7 shows plots for the ROC and PR curves when the model fit on 80% of the subsamples attempts to classify the remaining 20%. Note that since the subsample is balanced, the class imbalance issues which prevented the use of ROC for evaluation no longer apply.

As seen in Table 11, the model has a number of key differences compared to the one fitted on the entire cohort (see Table 10). First, in spite of the fact that age was specified as a fixed effect for the longitudinal process, it was found to not be significant at the 0.05 confidence level. Second, while age and heart rate remained significant in the survival process, service-trauma was

[12]This was performed using the wald.strata() function included in the JM package.

Longitudinal Process			
Variable	**Estimate**	**Std. Error**	**P-Value**
Intercept	89.8386	0.9660	< 0.0001
Time	0.0018	0.0088	0.8376
Age	-0.1011	0.0573	0.0776
Survival Process			
Variable	**Estimate**	**Std. Error**	**P-Value**
Age	0.0143	0.0060	0.0167
Service - Trauma	1.6551	0.4503	0.0002
Heart Rate	0.0393	0.0061	< 0.0001

Table 11: Summary of joint model using a balanced subsample.

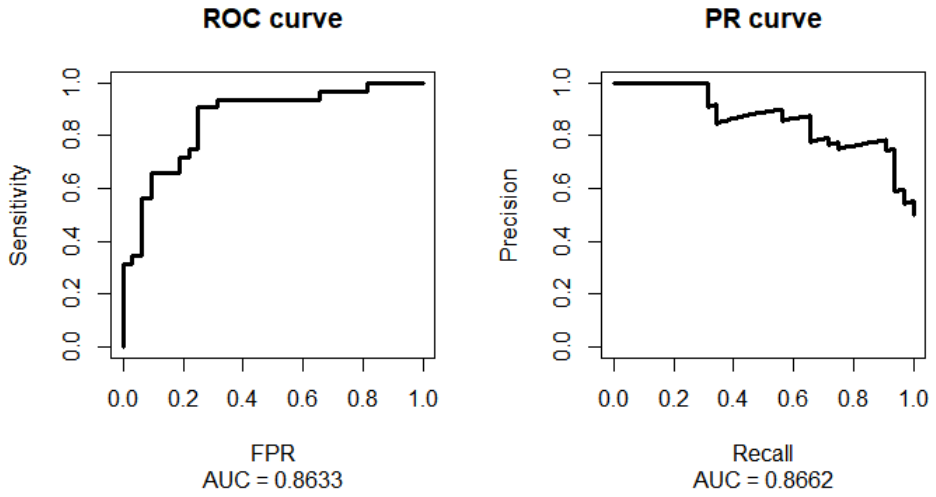

Figure 7: ROC and PR curves for the joint mode fit using an 80% partition of a balanced subsample of the cohort.

also found to be significant and positively associated with risk of in-ICU death. In particular, an individual that was admitted to the ICU for services related to trauma suffered a 5.23-fold increase in risk. Finally, aside from the fact that more spline knots were specified indicating finer control of the baseline risk, the coefficients were also relatively higher for the last few knots, indicating a higher baseline risk of death as time passed.

Based on the results in Figure 7, the predictive power of the model was also reasonably high, with an ROC-AUC of 0.8633 and a PR-AUC of 0.8662. This result suggests that the low AUCs obtained in all the models fitted on the full cohort may be due to the heavy imbalance present in the data rather than the quality of the variables being considered. However, conclusions to this end are difficult to ascertain since the prevalence of the dataset used plays a big part in how well the AUC metrics perform.

Model Diagnostics

Model diagnostics were performed on the basic HR joint model (summarised in Table 9) to check if the assumptions of the framework were followed. On one hand, the longitudinal process assumed a linear relationship between heart rate and observation time, as well as normal errors. On the other hand, the survival submodel assumed a B-spline baseline risk function, as well as controlling for all static variables and the true HR measurements.

Figure 8 shows diagnostic plots for the longitudinal process. The left-side graph plots the residuals of the process against the fitted values, while the right-side graph is a Q-Q plot of the

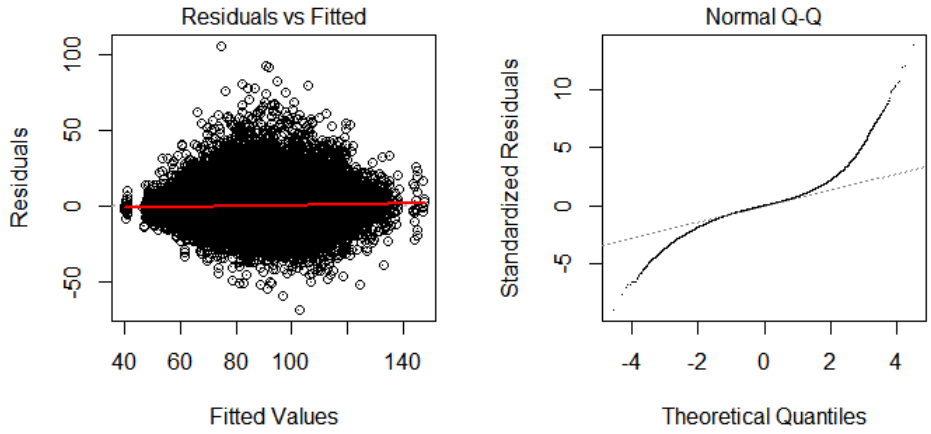

Figure 8: Diagnostic plots for the joint model using only HR as a longitudinal outcome, as fitted using the entire cohort. Left: Residuals vs. fitted values of the longitudinal process. Right: Q-Q plot of the subject-specific residuals of the longitudinal process.

residuals.

An appropriate fit would have residuals that are homoscedastic and approximately normal. In particular, when plotting the residuals against the fitted values, the residuals would all ideally scatter around 0, with no discernible patterns or shapes. For the normal Q-Q plot, the empirical cumulative distribution of the residuals would ideally follow the dotted line indicating the cumulative distribution of the normal distribution.

Figure 8 indicates that the model in question failed on both accounts. First, the residuals vs. fitted values plot shows that the error terms were greater around the mean of the longitudinal outcomes. Second, the Q-Q plot was not following the normal CDF at both tails. These observations suggested that the specified longitudinal model was inappropriate. A common solution that could be utilised would be to specify the longitudinal process using the log of HR rather than HR on its own.

For the survival model, the assumption to be checked was that the specified relationship between the longitudinal outcome and the hazard rate was appropriate. In all the models fitted, the association can generally be expressed as

$$h_i(t) = h_0(t) \exp[\gamma^\intercal w_i + \alpha(HR)] \tag{4.7}$$

One way of checking if the specified relationship is appropriate is by evaluating the martingale residuals. Conceptually, the subject-specific martingale residual can be thought of as the difference between the number of observed events for a subject over some time t and the number of events the model predicts the subject to have experienced in the same time span [13]. When plotting martingale residuals versus fitted longitudinal outcomes, an appropriate hazard model specification would have most of the residuals at and around 0. Figure 9 is the plot of the martingale of the simple HR joint model against the fitted HR values.

Figure 9 shows that the specified relationship for the hazard model was problematic. As the fitted HR values increased, the magnitude of the residuals increased as well, and the residuals were generally in the negative region, indicating that the model was overestimating the number of events. It was possible that the high skew of the data was causing these issues. For the sake of comparison, Figure 10 shows the martingale plots for the HR joint model fitted using a balanced subsample of the cohort.

While the magnitude of the residuals were not as drastic for the subsample as they had been for the full cohort, the behaviour was still evident - the residuals got more negative as the fitted heart rate increased. Similar to the proposed solution for the longitudinal process, using the log

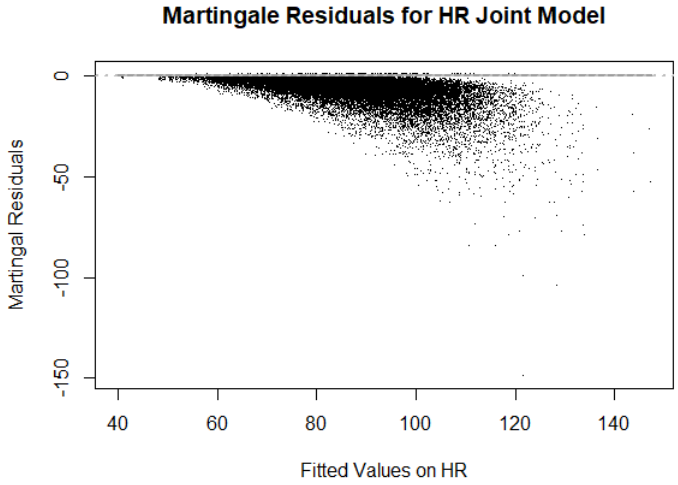

Figure 9: Subject-specific martingale residuals vs. subject-specific fitted values on HR for the simple HR joint model.

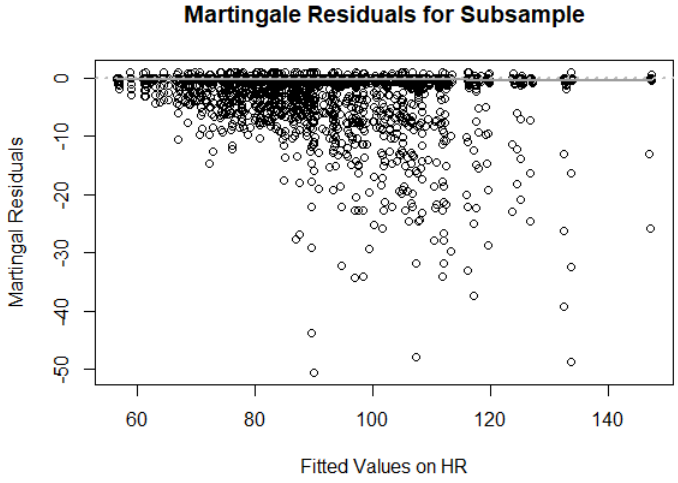

Figure 10: Subject-specific martingale residuals vs. subject-specific fitted values on HR for the HR joint model fitted on a balanced subsample.

of HR in the survival process may help the model adhere to the assumptions.

4.2.2 Bayesian Models

Based on the results summarised in Table 8, the two variables that exhibited the highest average predictive power were HR and SBP. Simple joint models were fit on five balanced subsamples of the cohort, and the model results are summarised in Table 12. Table 13 summarises the Bayesian joint model fit for one of the subsamples in terms of the coefficients found to be significant at a 0.05 confidence level. The models were fit using package default priors; in particular, independent normal priors with mean 0 and variance 1000 were used for both the parameters of the longitudinal outcomes and of the survival model.

An immediate observation from Table 13 was that HR was not found to be a significant variable in the joint model utilising both SBP and HR. This result suggests that, at least in the context of the subsample used, it was possible that HR was highly correlated with one of the other variables included in the model such that it became obsolete. Alternatively, for the subsample used, heart rate might not have had enough associative power with the outcome to

	Mean AUC	St. Dev.
ROC	0.6910	0.0659
PR	0.7190	0.0575

Table 12: Validation metrics for the Bayesian joint model using HR and SBP as longitudinal outcomes.

Longitudinal Process : HR			
Variable	**Estimate**	**Std. Dev.**	**P-Value**
Intercept	89.1624	1.0001	0.0000
Longitudinal Process : SBP			
Variable	**Estimate**	**Std. Dev.**	**P-Value**
Intercept	115.2454	1.2785	0.0000
Survival Process			
Variable	**Estimate**	**Std. Dev.**	**P-Value**
Service - C. Surg.	-1.3746	0.5907	0.0140
Service - Trauma	1.1591	0.4716	0.0100
SBP	-0.0387	0.0091	0.0000

Table 13: Summary of significant coefficients in the Bayesian joint model fit using a subsample of the cohort.

be of use to the model.

Of the two hypotheses, the latter seems more probable given that on the whole, the validation metrics presented in Table 12 were unimpressive compared to the expected performance of a random classifier given the same data. The results also compare unfavourably to those presented in Table 11, which used all available heart rate data to build a joint model on a similar subsample. This suggests that, in terms of the data being used, predictive power was not based primarily on introducing more variables into the model, but that the power was improved by having more measurements over longer periods of time.

5 Conclusion

5.1 Summary of findings

This study utilised the publicly available MIMIC-III dataset to explore the dynamic prediction of in-ICU fatality using logistic regression and joint longitudinal and survival modelling techniques. The main tools used to assess the predictive performance of models were the ROC-AUC and PR-AUC.

The logistic regressions performed in Section 3 all showed relatively poor predictive performance. Further, they suggested that adding information for more periods of time within the first 16 hours did little to improve the predictive performance of models. Among the logistic models shown in this study, the best model (high AUC and parsimonious) included heart rate (HR), respiration rate (RR), and systolic blood pressure (SBP) for only the third time bin as predictors for 72-hour in-ICU mortality (see Table 7). The model indicated that HR was positively associated with being a case, while respiration RR and SBP were negatively associated. However, despite the statistical significance of these results, the actual effect sizes were small, each changing the odds ratio in increments of less than 10%.

The joint models fitted in Section 4 also performed poorly in terms of prediction, in spite of having been able to incorporate data gathered over longer periods of time. Of the single outcome longitudinal models that were fit, it was found that heart rate and systolic blood pressure had the highest AUCs, somewhat coinciding with the fact that the logistic models found those two variables significant as well. However, the effect sizes of these variables still remained small, and survival trajectories did not seem to change upon the introduction of additional measurements (see Figure 6). Further, diagnostics suggested that the specification of the models were inappropriate. This was seen both in how the residuals of the longitudinal process were neither normal nor homoscedastic (see Figure 8), and the very large martingale residuals of the survival process (see Figure 9).

One result of particular interest was when a joint model using heart rate as a longitudinal outcome was fit to a balanced subsample of the data, where the number of cases and controls were equal. The model performed exceptionally well (see Figure 7), at least in comparison to all other models presented in the study. This suggested that the reason for the low predictive power of all the models resided not in the techniques used, but in the highly imbalanced nature of the dataset itself.

Multivariate Bayesian joint models were fit on five balanced subsamples of the data, making use of the first 16 hours of SBP and HR measurements as longitudinal outcomes. The model performed poorly (see Table 12) compared to the single outcome joint model fit on a balanced subsample. This suggested that it was perhaps the data gathered past the first 16 hours, closer to the actual event occurrence times, that had true predictive power.

For both techniques, while there were some variables that stood out in terms of relative predictive power (HR and SBP), the actual discriminatory ability of the models presented were still very low. This coincides with what was expected in Section 2.1.2, where it was found that the longitudinal variables tended to be flat, and the high variance of measurements both in cases and controls prevented the variables from making clear classifications.

5.2 Discussion and Future Prospects

EHRs present a unique avenue through which statistics, data science, and machine learning can be utilised to life-saving effect, and publicly available datasets such as MIMIC-III are invaluable to the development of this industry. However, the quality of these datasets must be evaluated if they are to be useful in a research setting. The MIMIC-III dataset suffered from high amounts of missing static data (e.g. BMI, height), forcing researchers to either exclude patients or exclude potentially useful variables from their study. Due to the data being gathered during routine care rather than at designated periods, data availability for longitudinal variables

also suffered (e.g. pH, temperature), resulting in the exclusion of these variables, and in the case of those that were not excluded, the reduction of sample size to accommodate the irregular intervals at which data was recorded.

Class imbalance was an issue touched upon multiple times in this study. However, class imbalance is a direct result of how cases and controls are defined. The study considered as cases only those who died within 72 hours of their ICU admission, and before they were discharged. This definition might be seen as too restrictive, ignoring patients who died shortly after their discharge. Future studies might consider relaxing how they define their cases to lessen the class imbalance present in the MIMIC-III dataset.

Aside from evaluating the quality of data, technological limitations must also be faced. The software used to perform joint modelling in this study (JM and JMBayes) were unable to easily handle the datasets used, even though in the greater scope, the cohort used here was still quite limited in size. Medical predictions are very subject-specific - each person will tend to have different physiological contexts, and will react differently to various scenarios, making it of prime importance that such pieces of software are extended to accommodate very the large sample sizes that are characteristic of EHR studies.

References

[1] M. Capan, J. S. Ivy, J. R. Wilson, and J. M. Huddleston. A stochastic model of acute-care decisions based on patient and provider heterogeneity. *Health Care Management Science*, 20(2):187206, 2015.

[2] D. R. Cox and D. Oakes. *Analysis of survival data.* Chapman and Hall/CRC, 1998.

[3] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning - ICML 06*, 2006.

[4] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, and J. Ye. Analysis of sampling techniques for imbalanced data: An n=648 adni study. *NeuroImage*, 87:220241, 2014.

[5] M. Ghassemi, L. Celi, and D. J. Stone. State of the art review: the data revolution in critical care. *Critical Care*, 19(1):118, Mar 2015.

[6] D. W. Hosmer and S. Lemeshow. *Applied logistic regression.* John Wiley, 2nd edition, 2000.

[7] 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 2014*, pages 157–160, Sept 2014.

[8] A. E. W. Johnson, T. J. Pollard, and R. G. Mark. Reproducibility in critical care: a mortality prediction case study. In *MLHC*, 2017.

[9] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035 EP –, May 2016. Data Descriptor.

[10] G. K. Lighthall and C. Vazquez-Guillamet. Understanding decision making in critical care. *Clinical Medicine & Research*, 13(3-4):156168, 2015.

[11] C. D. Manning, P. Raghavan, S. Hinrich, and u. u. undefined. *8. Evaluation in information retrieval.* Cambridge University Press, 2009.

[12] G. A. Ospina-Tascn, G. L. Bchele, and J.-L. Vincent. Multicenter, randomized, controlled trials evaluating mortality in intensive care: Doomed to fail? *Critical Care Medicine*, 36(4):13111322, 2008.

[13] D. Rizopoulos. *Joint models for longitudinal and time-to-event data: with applications in R.* Chapman & Hall/CRC, 2012.

[14] D. Rizopoulos. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7), Aug 2016.

[15] L. N. Sanchez-Pinto, Y. Luo, and M. M. Churpek. Big data and data science in critical care. *Chest*, 2018.

[16] E. Steyerberg. Stepwise selection in small data sets a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 52(10):935942, 1999.

[17] E. Vittinghoff and C. E. Mcculloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, 165(6):710718, Dec 2007.

[18] M. L. Vold, U. Aaseb, T. Wilsgaard, and H. Melbye. Low oxygen saturation and mortality in an adult cohort: the troms study. *BMC Pulmonary Medicine*, 15(1), Dec 2015.

[19] N. G. Weiskopf and C. Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144151, Jan 2013.