University of Strathclyde
Glasgow

Development of Prediction Models for Postoperative Complications Following Cardiac Surgery

Linda Lapp

201787114

Department of Computer and Information Sciences
University of Strathclyde

Glasgow, 2022

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science.

Declaration of Authenticity

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Published Work

Research in this thesis has been published at the following venues:

1. Lapp, L., Young, D., Kavanagh, K., Bouamrane, M.-M., Schraag, S.: Using machine learning for predicting severe postoperative complications after cardiac surgery. European Association of Cardiothoracic Anaesthesiology (EACTA) Annual Congress 2018, Journal of Cardiothoracic and Vascular Anesthesia, 32 (S1), S84-S85 (2018).
2. Lapp, L., Bouamrane, M.-M., Kavanagh, K., Roper, M., Young, D., Schraag, S.: Evaluation of Random Forest and Ensemble Methods at Predicting Complications Following Cardiac Surgery. Conference on Artificial Intelligence in Medicine. AIME 2019. Lecture Notes in Computer Science. 11526, 376-385 (2019).
3. Lapp, L., Roper, M., Kavanagh, K., Schraag S.: Predicting the Onset of Delirium on an Hourly Basis in an Intensive Care Unit Following Cardiac Surgery. 2022 IEEE 35th International Symposium on Computer Based Medical Systems (CBMS), 234-239 (2022)

4. Lapp, L., Bouamrane, M.-M., Roper, M., Kavanagh, K., Schraag, S.: Definition and Classification of Postoperative Complications After Cardiac Surgery: A Pilot Delphi Study. JMIR Perioperative Medicine 5(1), e39907 (2022)
5. Lapp, L., Roper, M., Kavanagh, K., Schraag, S.: Performance of Machine Learning Algorithms to Predict Acute Kidney Injury Following Cardiac Surgery. Anesthesia & Analgesia. (Under peer-review, submitted on the 2nd of September 2022)
6. Lapp, L., Roper, M., Kavanagh, K., Bouamrane, M-M., Schraag, S.: Dynamic Prediction of Patient Outcomes in the Intensive Care Unit: A Scoping Review of the State-of-the-Art. Journal of Intensive Care Medicine. (Under peer-review, submitted on the 4th of November 2022)

The author of this thesis was the principal study designer and investigator in each of the listed articles. She also led the analysis and the write-up of the results. The first and second paper include the content discussed in Chapter 6, however, involves analysis of older data – hence the results are slightly different. The third paper discusses the findings from Chapter 8. The fourth paper discusses the study undertaken in Chapter 4. The fifth paper discusses the findings from Chapter 9.

Part of the results from Chapter 6 were presented at the AIME 2019 Conference in Poznan, Poland, where the second paper was accepted.

Some results from Chapter 6 were also presented at the EACTA Annual Congress 2018 in Manchester, UK. The abstract of the presented work was published in the Journal of Cardiothoracic and Vascular Anaesthesia (the first paper).

The results of Chapter 8 were presented at the CBMS 2022 conference in Shenzhen, China (online), where the third paper was accepted.

Acknowledgements

This PhD work is funded by the University of Strathclyde, Golden Jubilee National Hospital and Fresenius Kabi.

Firstly, I would like to thank my supervisors Dr Marc Roper, Dr Kimberley Kavanagh, Prof Stefan Schraag, Dr Matt-Mouley Bouamrane, and Dr David Young for excellent supervision that allowed me to explore my research interests, and for their patience to allow me to heal and overcome adversities.

I would also like to thank Dr Marilyn Lennon for feedback, support, and advice, and for trusting me with teaching opportunities that not only helped me to get inspired about my PhD again, but also benefitted me in terms of professional development.

Furthermore, I would like to thank Mr Wilson, Dr MacPherson, Dr Clark and all nurses and staff members at Beatson who took care of me during my cancer treatment. Without you, I would not have been able to write this thesis.

And above all, I would like to thank my family: emme, issi ja Martin, aitäh, et te mind alati toetate ja julgustate.

Table of Contents

List of Figures

List of Tables

Abbreviations

AKI – acute kidney injury
ASA - American Society of Anesthesiologists
AUC – Area under the receiver operating characteristic curve (used in Chapters 5 to 9)
AUROC – Area under the receiver operating characteristic curve (used in Chapter 2)
AUPRC – Area under the precision-recall curve
BARTm – Bayesian aggregated regression trees machine
BCART – bagging classification and regression trees
CABG – Coronary artery bypass graft
CAM-ICU – Confusion Assessment Method for the Intensive Care Unit
CaTHI – Cardiac and Thoracic Health Information database
CCI – comprehensive complication index
CI – confidence interval
EuroSCORE – European System for Cardiac Operative Risk Evaluation
GBM – gradient boosting model
GJNH – Golden Jubilee National Hospital
GLM – generalised linear model
ICU – Intensive care unit
IQR – inter-quartile range
LR – logistic regression
LV function – left ventricular function
MIMIC – Multiparameter Intelligent Monitoring in Intensive Care
NACSA – National Adult Cardiac Surgery Audit
NB – naïve bayes
NHS – National Health Service
NPV – negative predictive value
NYHA Grade – New York Heart Association Grade
OR – odds ratio
PCI – percutaneous coronary intervention
PPV – positive predictive value
RF – random forest
ROC – receiver operating characteristic curve
RQ – Research Question
SCTS – The Society for Cardiothoracic Surgery
SD – standard deviation
Sepsis-3 – Third International Consensus Definitions for Sepsis and Septic Shock Criteria
SIRS – systemic inflammatory response syndrome
SMOTE – Synthetic Minority Oversampling Technique
SVM – support vector machine
TRIPOD – Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

Abstract

While postoperative mortality in cardiac surgery has reduced in the past twenty years, due to changes in patient population undergoing open-heart surgery, postoperative complications are becoming more common. With the development of perioperative medicine, data-driven perioperative risk prediction models are now an integral component for decision-making about the type of treatment that is most suitable for the patient, for communicating risk of surgery, and for auditing purposes. However, the currently developed prediction models focus on mortality, rather than postoperative complications.

In this thesis, the problem of postoperative complications in cardiac surgery is investigated by analysing cardiac patient data in Golden Jubilee National Hospital to predict (1) severe postoperative complications, (2) acute kidney injury and (3) delirium. Furthermore, cardiac anaesthetists and surgeons were involved in explorative interviews about current challenges in cardiac surgery, and a study to define and classify postoperative complications in cardiac surgery.

Patients undergoing coronary artery bypass graft (CABG), valve and combined CABG and valve surgeries in Golden Jubilee National Hospital between 1st April 2012 and 31st December 2018 were analysed. The prevalence of severe complications, acute kidney injury and delirium for this patient population was 5.91%, 18.93% and 12.47%, respectively.

Two types of models were developed: (1) preoperative models using data that was available before surgery; and (2) hourly prediction models that used both preoperative data and laboratory results recorded in the intensive care unit.

Out of all preoperative modelling experiments (1), random forest predicting severe postoperative complications had the highest performance, with the area under the receiver operating characteristic curve (AUC) of 0.713, sensitivity of 0.562 and specificity of 0.748. When predicting the onset of acute kidney injury on an hourly basis in intensive care (2), BARTm achieved the highest mean AUC of 0.850 with sensitivity of 0.821 and specificity of 0.741. For hourly delirium prediction (3), support vector machine achieved the highest mean AUC of 0.941, sensitivity of 0.907 and specificity of 0.870.

This thesis shows that using routinely collected medical data can be used to develop both preoperative and hourly ICU predictive models for postoperative complications, such as acute kidney injury and delirium. Such prediction models could help with clinical decision making, communication about risk, research in complications and auditing.

Chapter 1. Introduction

According to the National Adult Cardiac Surgery Audit (NACSA), in the UK and Ireland, 34,000 cardiac surgeries were undertaken between 2016 and 2019 [1]. The most common open-heart surgeries in the UK are coronary artery bypass graft, (CABG), aortic valve replacement, and a combination of CABG and valve surgery.

CABG is a surgery where blood flow to the heart is improved for patients who have severe coronary heart disease. This disease can cause a waxy substance called plaque building up inside the coronary arteries which supply oxygen-rich blood to the heart. If the plaque hardens, it can narrow the arteries, which subsequently reduces the flow of oxygen-rich blood to the heart, causing chest pain or angina. CABG consists of connecting a healthy artery or vein from the body to the blocked coronary artery, creating a new path for oxygen-rich blood to flow to the heart [2].

The four valves of the heart open and close to regulate the blood flow through different parts of the heart, ensuring that it travels in one direction. If the valve does not open fully and obstructs blood flow, or if it does not close properly, allowing blood to leak backwards, a patient has a valve surgery [3].

On average, cardiac patients are reported to stay in the hospital for 7.8 days after surgery [1], however, the hospital stay can vary based on whether a patient has postoperative complications [4], [5]. While postoperative mortality in cardiac surgery has reduced in the past twenty years, and is remarkably low (below 3%) [1], [6], due to changes in patient population undergoing open-heart surgery, postoperative complications are becoming more common [7].

Currently, the focus in terms of cardiac surgery outcomes is on mortality[1], however, as suggested by the Society for Cardiothoracic Surgery (SCTS) *"Morbidity[2], rather than mortality, may be a better indicator of the quality of care"* [6]. Nevertheless, the SCTS and NACSA reports both state that morbidities, i.e., postoperative

[1] Mortality – the condition of being mortal or subject to death [406].

[2] Morbidity – the condition or state of being diseased, or being caused by disease; physical or mental illness [407]

complications, are recorded in electronic health records less accurately, and the data quality is lacking [1], [6]. Due to this, it is difficult to pinpoint, which complications are the most common and what their exact incidence in cardiac surgical population is. Furthermore, unlike for clinical diagnosis (i.e., the International Statistical Classification of Diseases, ICD-10[3]), postoperative complications do not have a classification system that enables high-quality reporting of complication diagnoses. However, it is known that complications can have a serious impact on patients' quality of life [8], [9], hospital length of stay [4], [5] and healthcare costs [10].

A way to mitigate adverse surgical outcomes, perioperative medicine practices, such as the development of preoperative clinics and services [11], involvement of multi-disciplinary teams [12] and shared decision-making with patients and their families [13] have been developed. With the development of perioperative medicine, data-driven perioperative risk prediction models are now an integral component for decision-making about the type of treatment that is most suitable for the patient, for communicating risk of surgery and for auditing purposes [14]. However, the current widely used prediction models focus on mortality, rather than postoperative complications [14]–[16].

1.1. Research Hypothesis and Research Questions

The work discussed in this thesis aims to highlight the opportunities of routinely collected medical data in developing clinical prediction models for predicting postoperative complications in cardiac surgery. Hence, the research hypothesis of this thesis is as follows: patient outcomes can accurately be predicted following cardiac surgery, using routinely collected preoperative and intensive care unit (ICU) data. This thesis aims to develop prediction models, both for preoperative and postoperative use, for predicting postoperative complications occurring after cardiac surgery. This aim is fulfilled by answering the following research questions:**RQ1:** What is the current

[3] https://icd.who.int/browse10/2019/en#/

landscape of dynamic prediction models in critical care in terms of predictive modelling methods?

RQ2: What are cardiac surgery experts' challenges in cardiac surgery and priorities for a new prediction model predicting patient outcomes?

RQ3: How can postoperative complications be classified using routinely collected medical data?

These research questions are answered through different studies presented in this thesis, specifics of which are shown in Table 1.1.

Table 1.1. Questions answered in each chapter of the thesis.

Chapter Number	Research Question	Specific Questions Answered
Chapter 2	RQ1	Which outcomes are predicted in critical care in a dynamic manner? Which methods are used to handle missing data when developing critical care prediction models? Which methods are used to deal with the imbalanced classification problem when predicting critical care outcomes? Which dynamic predictive modelling methods are used to predict patient outcomes in critical care?
Chapter 3	RQ2	What are the current challenges in cardiac surgery? What are the current processes to avoid adverse outcomes in cardiac surgery? What are the clinicians' priorities for clinical risk prediction models?
Chapter 4	RQ2	What are cardiac surgery experts' opinion on the usefulness of a definition and classification of surgical complications following cardiac surgery? How do cardiac surgery experts define what events constitute surgical complications following cardiac surgery? How do cardiac surgery experts classify surgical complications following cardiac surgery?
Chapter 6	RQ3	What is the optimal number of variables required to predict postoperative complications, using preoperatively available data? Would upsampling of benefit the predictive ability of models in case of an imbalanced classification problem? Which method performs best when predicting postoperative complications, using preoperatively available data?
Chapter 7 and 8	RQ3	How do models perform when predicting the onset of acute kidney injury and delirium hours in advance? How do models perform when using complete data, missing data and imputation methods? Which method performs best when predicting acute kidney injury and delirium on an hourly basis in an intensive care unit (ICU)?

1.2. Comparison to MPhil

This thesis is a follow-up work from the author’s MPhil project [17]. The Table 1.2 gives an overview of the differences between the author’s MPhil and PhD.

Table 1.2. Differences between the MPhil and the PhD.

Characteristic	**MPhil**	**PhD**
Literature Review	Commonly used preoperative prediction models predicting postoperative complications	Dynamic prediction models developed for ICU
Databases	CaTHI	CaTHI and Centricity^{TM} CIS
Dates of Procedures	April 2012 to March 2016	April 2012 to December 2018
Type of Data	Preoperatively available data and some surgical outcome data	Preoperatively available data, ICU laboratory variables and some surgical outcome data
Predicted Outcomes	Postoperative complications (Yes/No), level of postoperative complications (No/Mild/Moderate/Severe), severe postoperative complications (Yes/No or other)	Severe postoperative complications (Yes/No or other), postoperative acute kidney injury based on KDIGO criteria, postoperative delirium based on CAM-ICU assessment
Analysis	Classical statistical methods were used and risk factor analysis undertaken	Experiments with different data (preoperative and postoperative), experiments with number of variables in models, imbalanced classification problem approaches, missing data approaches.
Prediction Methods	Static models, only logistic regression	Static models and hourly dynamic prediction models. Various machine learning methods were used.
Stakeholder involvement	Clinical supervisor only	Clinical supervisor, cardiac anaesthetists and cardiac surgeons based in Scottish cardiac centres (Chapter 3), cardiac anaesthetists and intensivists largely based in the UK (Chapter 4)

The contributions of this thesis have addressed the main limitations of the MPhil project, listed below.

1. The PhD includes laboratory data from the intensive care unit, enabling more precise retrospective diagnosis of certain complications, as opposed to relying on only reported complications in the CaTHI database. The problems with the reporting of postoperative complications are addressed throughout this thesis.

2. The definition of “Severe” postoperative complication in the PhD are more objective due to being based on a Delphi study (Chapter 4).
3. The predicted outcomes in the PhD were chosen based on the identified needs of cardiac anaesthetists and cardiac surgeons (Chapter 3), and availability of widely used diagnostic criteria (Chapter 5).
4. The models predicting postoperative outcomes in the PhD include also more granular data, such as laboratory values recorded in the ICU, as opposed to only preoperatively available data (Chapters 7 and 8).
5. The MPhil, in general, is focusing on risk factor analysis and classical statistical analysis methods, such as logistic regression. The PhD, however, experiments with a number of machine learning methods to predict various clinical outcomes, also with the number of variables included in the prediction model, approaches for imbalanced classification methods and experiments with methods to handle missing data.

1.3. Contribution to Knowledge

This thesis presents several contributions to knowledge both in the field of computer and information sciences and cardiac surgery.

Computer and information sciences contributions:

1. The literature review conducted as part of this thesis is the first review to analyse the currently available prediction models developed to predict patient outcomes in critical care and ICU in real-time. The review identifies which patient outcomes are predicted, the methods used for model development, and the performance of the models.
2. In this thesis, two types of novel models were developed focusing on predicting “Severe” postoperative complications (acute kidney injury and delirium) using preoperative and intensive care unit data. Their performance was AUC = 0.713 for “Severe” complications, using preoperative data only, and mean AUC = 0.850 for the model predicting acute kidney injury between hour 0 and 25 at 1 hour intervals in the ICU, and mean AUC = 0.941 for the model predicting delirium between hour 0 and hour 13 at 1 hour intervals in the ICU, using

preoperative and ICU data. It was found that the optimal prediction time, based on performance measures, is achieved when acute kidney injury is predicted 1 hour in advance, using BARTm (AUC = 0.918), and delirium 13 hours in advance, also using BARTm (AUC = 0.997).

3. Based on the results of the literature review, this thesis is the first to apply BARTm to critical care data that includes missing values when predicting patient outcomes. The findings of this thesis demonstrate that hourly BARTm models for acute kidney injury and delirium were robust at handling missing values, achieving high mean performance of AUC = 0.830 and AUC = 0.930 for acute kidney injury and delirium, respectively. These performance measures were achieved when the models were applied on data with 37.9% of missing data for acute kidney injury model and 3.2% of missing data for delirium model. It is worth noting that patients with more than 40% of missing data were removed from analysis. As missing data in electronic health records is common, being able to make a prediction for a patient who does not have all the necessary data available allows for clinicians to make a decision with the aid of prediction model for most patients.

Medical and cardiac surgery contributions:

1. The findings of the exploratory interviews identified that according to cardiac surgeons and anaesthetists, the main challenges of cardiac surgery are postoperative complications, changes in patient population and procedures. The study also found that for a new prediction model for patient outcomes, clinicians prioritise the prediction of postoperative complications to mortality.
2. A Delphi study reached a consensus on the definition for postoperative complications in cardiac surgery and classification of these are “Mild”, “Moderate”, “Severe” and “Death”. Consensus was reached on the characteristics for “Mild” and “Severe” postoperative complications. The results of this study allow to develop standardised way of identifying, recording and reporting of complications to help the development of future quality benchmarks, clinical audit, care quality assessment, risk management and research. For example, in this thesis, using the classification criteria found

from this study, “Severe” complications were predicted based on preoperatively available data.

3. The models developed in this thesis were able to predict acute kidney injury up to 24 hours in advance with sensitivity and specificity of 0.821 and 0.741, respectively, and delirium up to 13 hours in advance with sensitivity and specificity of 0.907 and 0.870, respectively.
4. It was found that creatinine, urea, daily fluid balance, urine output, lactate and hydrogen ion were the most important variables when predicting acute kidney injury. Lactate, urine output, potassium and hydrogen ion were the most important variables when predicting delirium.

1.4. Thesis Overview

As explained previously, the work in this thesis aimed to develop predictive models for postoperative complications in cardiac surgery that could be used preoperatively and postoperatively. Firstly, to understand the current landscape of dynamic ICU prediction models, Chapter 2 shows the findings of a literature review that analyses dynamic prediction models developed to predict patient outcomes in critical care and in an intensive care, which informed the predictive modelling methods and predicted outcomes in this thesis. To understand the requirements of potential users for such prediction models, Chapter 3 analyses exploratory interviews with cardiac anaesthetists and cardiac surgeons based in Scottish cardiac centres. Subsequently, to improve challenges that postoperative complications bring in cardiac surgery, a Delphi study was undertaken in Chapter 4 to find a consensus in definition and classification of postoperative complications in cardiac surgery. Chapter 5 explains the predictive modelling methods that were used in Chapters 6 to 8. Chapter 6 uses preoperatively available data to predict “Severe” postoperative complications, postoperative acute kidney injury and delirium. Chapter 7 predicts the onset of acute kidney injury in intensive care on an hourly basis. Chapter 8 predicts the onset of delirium in intensive care, also on an hourly basis. Finally, Chapter 9 discusses the overall findings in this thesis, strengths and limitations, and future work.

Chapter 2. Dynamic Prediction Models in the Critical Care: A Literature Review

2.1. Introduction and Background

2.1.1. Machine Learning and Deep Learning in Medicine

While this thesis focuses on the data analysis in the preoperative and ICU stage of cardiac surgery, the subject area of artificial intelligence (AI) and machine learning in medicine is vast. Numerous reviews have been undertaken to analyse different aspects of AI and dynamic prediction in medicine, including the current challenges in developing deep learning models in healthcare [18], [19], implementation and adoption of AI [20]–[23], and reporting standards of studies developing AI-based systems in healthcare [24], [25].

The vast amount of data the healthcare industry produces have resulted in expansive evolvement of machine learning techniques which are used to understand complex data in various areas of healthcare. Some of these include models for bioinformatics, speech recognition and medical image processing, and are used to develop decision support tools, medical devices, diagnostic tools and medical treatments [26].

The methods of understanding large amounts of data can be referred to as either artificial intelligence (AI), machine learning or deep learning, and which can be collectively defined as machines that mimic human intelligence and their ability to learn through automatic calculation, conceptualising, self-improvement, abstraction and creative thinking [27]. Machine learning is where an algorithm learns from a large dataset and responds to this specific dataset, often requiring human expertise identify relevant features. Deep learning is a further development of machine learning, which is based on very large artificial neural networks. Artificial neural networks are composed of inter-connected nodes that loosely resemble and mimic the brain's neuronal functions, and therefore do not require human expertise to identify relevant features (except through the provision of training data), but learn about the features from the aforementioned training data [26].

The main deep learning methods used to predict various outcomes in medicine are currently recurrent neural networks (RNN), autoencoders, convolutional neural networks (CNN) and transformer-based models. RNN is a method that can capture temporal aspects of longitudinal data, which makes it a very popular method in predictive modelling in medicine [18]. Since medical data is often vast, autoencoders are often used to reduce the dimension size of the data without losing essential properties of the data, such as structures and regular patterns [18]. CNNs are often used to label or classify clinical text due to their ability to analyse images, speech and videos [28]. Transformer-based models are used to learn computationally expensive tasks, such as gene expressions and medical imaging [29].

AI can be applied to a variety of decision support tools, including supporting patient self-management, automating triaging based on existing data sets, facilitating the interpretation of images and to help with medication adherence [23]. Some of the most successful endeavours of AI in medicine are currently in the fields of biomedical imaging and biomedical signal processing. Deep neural networks have been used to detect various cancers [30]–[34] from imaging technologies, such as magnetic resonance imaging (MRI), positron emission tomography (PET) and histopathology images. According to a recent systematic review by Kumar et al., convolutional neural networks are the most commonly used method to predict and diagnose cancer, based on image recognition [35]. However, various review articles have expressed the concern of limited evidence for the accuracy of deep learning in screening various cancers, such as breast cancer [36], predicting lymph node metastasis in colorectal cancer patients [37], and prostate cancer [38].

In terms of signal processing, machine learning and deep learning have been used to detect electrical signals from human body, including the prediction of atrial fibrillation from electrocardiogram (ECG) signals [39] and diagnosis of Parkinson’s disease, using electroencephalogram (EEG) and electromyography (EMG) [40].

2.1.1.1. Current Challenges in AI and Medicine

Developing AI in medicine can face a number of challenges when using electronic health record data. Firstly, lack of labels is an issue. A large amount of data is captured as free-text medical notes, making the extraction of meaningful clinical outcomes and patient status difficult [19], [41]. Furthermore, as medical notes are entered not necessarily in real-time, the time of when diagnoses were recorded might not be accurate [18]. Hence, efforts should be directed towards capturing diagnoses codes on real-time, based on patient information, such as laboratory results or vital signs. This issue is also connected to the problem of lack of appropriate IT infrastructure. The transition from paper-based records to digital standardised medical systems has been slow, which hinders the healthcare providers' readiness to embed new AI-based medical systems [19].

Another challenge of AI is interpretability as deep learning models are known to be of "black-box" nature due to not providing any explicit explanation of how the results were achieved [18], [19], [42]. As stated by van Smeden et al., with such "black-box" models, the predictions are difficult to scrutinise, which may reduce the trustworthiness of the AI prediction model for the user [25].

Ethics, privacy and confidentiality has also been stated as a concern when it comes to developing new AI-based medical prediction models [19]. As historically medicine-related studies have been randomised controlled trials, and not purely studies of patient data, there is a lack of infrastructure that enables safe and ethical exchange of patient data between medical institutions and researchers. To avoid this issue, open-source databases have been developed to help with innovation in AI-based prediction models [18]. An example of such widely used open-access database is the intensive care unit database Medical Information Mart for Intensive Care (MIMIC) [43] has been used by numerous studies to develop machine learning and deep learning models for patient outcomes such as mortality, sepsis and cardiac complications, as discussed in a systematic review by Syed et al. [44].

2.1.1.2. Implementation and Adoption of AI in Medicine

While there is a vast amount of AI-based prediction models in medicine, the information about implementation and adoption of AI is limited. As stated by

Seneviratne et al.: *"Very few of these algorithms ever make it to the bedside; and even the most technology-literate academic medical centres are not routinely using AI in clinical workflows"* [20]. A systematic review by Khanijahani et al. found that the main factors that influence the successful implementation of AI-based medical systems are perceived ease of use or usefulness, performance or effort expectancy, and social influence [21]. One of the main barriers to acceptance of AI by physicians was perceived threat to autonomy due to physicians worrying that AI will override or replace their judgement [21]. To address the shortcomings of current prediction models, Seneviratne et al. suggest shifting the focus from optimising performance metrics to practical aspects of model design, such as actionability, safety and utility, and consulting the potential users of the model [20].

While the current methods largely demonstrate theoretical and empirical benefits, all studies presented in this Section came to conclusion that before implementation of AI-based and dynamic prediction models, further investigation is needed in evaluating the performance and impact of the currently available models on clinical decision-making. It has also been found that reporting guidelines for developing AI prediction models have rarely been followed, which negatively affects the reproducibility and replicability of the findings [25]. To help with implementation and measuring effectiveness of the prediction models, specifically prospective testing and randomised controlled trials are needed [24], [44].

As claimed by Panch et al., "the inconvenient truth is that at present the algorithms that feature prominently in research literature are in fact not, for the most part, executable at the frontlines of clinical practice." [45] The existing ways of working do not allow to make room for AI innovations and the current healthcare data infrastructure that enables training algorithms in an optimal way is lacking [46].

2.1.2. Aims of the Literature Review

Critical care units (specifically intensive care units (ICU) and high dependency units (HDU)) continuously monitor patients and large volumes of both high and low frequency patient data is often captured and stored by patient monitoring systems. Studies have shown that the timeliness of health interventions has a significant effect on the clinical outcomes [47], [48]. The timeliness can be improved with accurate prognosis and early warning.

As stated by Huddar et al., *"Accurate knowledge of the aetiology of ICU complications is often lacking, leading to the inability of accurate identification of high-risk patients and prevention of complications."* This has resulted in current medical interventions needing to be reactive, with the adequate care being provided to patients after the complication has already been developed [49].

In this chapter various dynamic predictive modelling methods were explored for predicting patient outcomes, primarily in ICU or critical care unit, based on longitudinal time-series data, including laboratory results. What exactly is meant by "dynamic" models is explained in detail in Section 2.2.3.

This chapter aims to answer the following questions:

- Which outcomes are predicted in the critical care in a dynamic manner?
- Which methods are used to handle missing data when developing critical care prediction models?
- Which methods are used to deal with imbalanced classification problem when predicting critical care outcomes?
- Which dynamic predictive modelling methods are used to predict patient outcomes in critical care?

Through answering these questions, this review will inform this thesis on which methods to choose for developing a dynamic prediction model predicting postoperative complications following cardiac surgery based on ICU data. Furthermore, it will answer the second research question of this thesis: "What is the current landscape of dynamic prediction models in the ICU in terms of predictive modelling methods?".

2.2. Methods

2.2.1. Data Sources and Search Strategy

The search included the articles published between 1st January 2000 and 25th April 2022 and was undertaken using PubMed. In addition, references from found papers were screened, using PubMed.

The search was undertaken using the PubMed website for ((dynamic predict* [MeSH]) OR (real time predict* [MeSH])) AND ((patient outcome*[Title/Abstract]) OR (mortality[Title/Abstract]) OR (morbidity[Title/Abstract]) OR (complication*[Title/Abstract][MeSH])) AND ((critical care) OR (intensive care)) NOT (cancer) NOT (COVID-19) NOT (Paediatric) NOT (Pediatric) NOT (trauma).

The titles and abstracts of the found articles were screened, and the eligible articles were read in full to screen for eligibility. To make the process more systematic, the papers were downloaded and imported to NVivo [50] to undertake the full paper screening.

2.2.2. Citation Management

For the citation management and sorting the studies, Mendeley [51] was used. The Excel tables were created for managing the extracted data, including the first author of the study, year of publication, patients included in the study, predicted outcome, methods used to develop the predictive model, methods to pre-process the data and deal with missing values, types of features used in the model and performance measures used in the study.

2.2.3. Eligibility Criteria and Analysis

In Table 2.1 the inclusion and exclusion criteria are presented. In terms of study design, only papers about the development of the prediction model were included. Papers that only evaluated models or were review papers were excluded.

In terms of patient population, only adult critical care or ICU patients were included in the study. Prediction models developed specifically for cancer or trauma patients were excluded. While predicting ICU outcomes for cancer or trauma patients could help with managing unplanned ICU admissions [52], the additional confounding variables that cancer or trauma could add are most likely not relevant for predicting complications following cardiac surgery, which is the aim of this thesis. Since the COVID-19 pandemic from early 2020, many prediction models have been developed to predict COVID-19-related outcomes [53]. Because, this is a non-routine situation, studies with COVID-19 patients were excluded.

In terms of the setting, only adult critical care or adult ICU related studies were included. If the prediction model was developed in any other hospital setting that is not adult critical care or ICU, the study was excluded.

In terms of the predicted outcome, only studies including classification tasks were included. Studies developing regression models, or any other model that is not classification model, were excluded. This decision was made, because usually adverse clinical outcomes, such as mortality or complications, are defined as binary categorical outcomes, or are diagnosed based on a number of laboratory variables, as opposed to one numerical variable [54]. Therefore, as this review aims to understand the currently used methods to predict patient outcomes to develop clinical prediction models predicting complications, treating the outcomes as classification problems, as opposed to predicting the value of a certain laboratory variable is the chosen path in this thesis.

Papers developing prediction models for outcomes that are directly related to the patient health were included, including mortality, complications, and ICU stay. Studies that investigated other outcomes, such as bed planning or healthcare costs were excluded from this review.

Table 2.1. Inclusion and exclusion criteria for papers based on patients included in the study, variables used in analysis, outcome of the analysis, intervention, and study design.

Criterium	**Included**	**Excluded**
Study design	Primary study, i.e., study that develops a prediction model	Review article, validation study, commentary.
Patients	Adult critical care or intensive care patients, non-cancer patients, non-COVID-19 patients, non-trauma patients	Any other patient who is not admitted to critical or intensive care, cancer patients, COVID-19 patients, trauma patients.
Setting	Adult critical care or intensive care unit	Paediatric critical care or intensive care unit, emergency department, hospital wards, or any other hospital setting that is not adult critical care or intensive care unit.

Criterium	Included	Excluded
Type of a problem	Classification	Regression, or any other method that is not classification
Outcome	Patient outcomes: mortality, morbidity, postoperative complications, hospital length of stay, ICU length of stay or any other outcome that is directly related to patient's health	Outcomes that are not directly related to the patient (e.g., costs)
Variables	Includes laboratory data that were treated as dynamic variables	Includes only static variables (i.e., that are measured once) or variables that are not vital signs or laboratory data
Type of model	Must be a model predicting patient outcomes based on dynamic variables on a "real-time" basis.	Static prediction model
Comparator	Any model performance measure (e.g., AUC, sensitivity, specificity, accuracy, etc)	No model performance reported

Prediction models that were developed using dynamic laboratory test results were included in the study. The "dynamic variable" is defined as variables that are measured repeatedly as time changes. These could be laboratory results that are measured every hour or every day. Also, inclusion of vital signs that are measured every second or every minute was allowed. The models could include static variables (measured only once) to aid prediction, however, studies that used only static variables, were excluded.

Because the aim of this thesis is to develop prediction models that predicts postoperative outcomes using dynamic variables on an hourly basis, only papers that develop these kinds of models were included. The definition of "dynamic" here is flexible, where the prediction is made repeatedly as the time passes. The prediction could be made in every second, minute or hour, or even less often. The main idea is that the developed models make a prediction repeatedly as new information comes in, or as the predicted event gets closer in time.

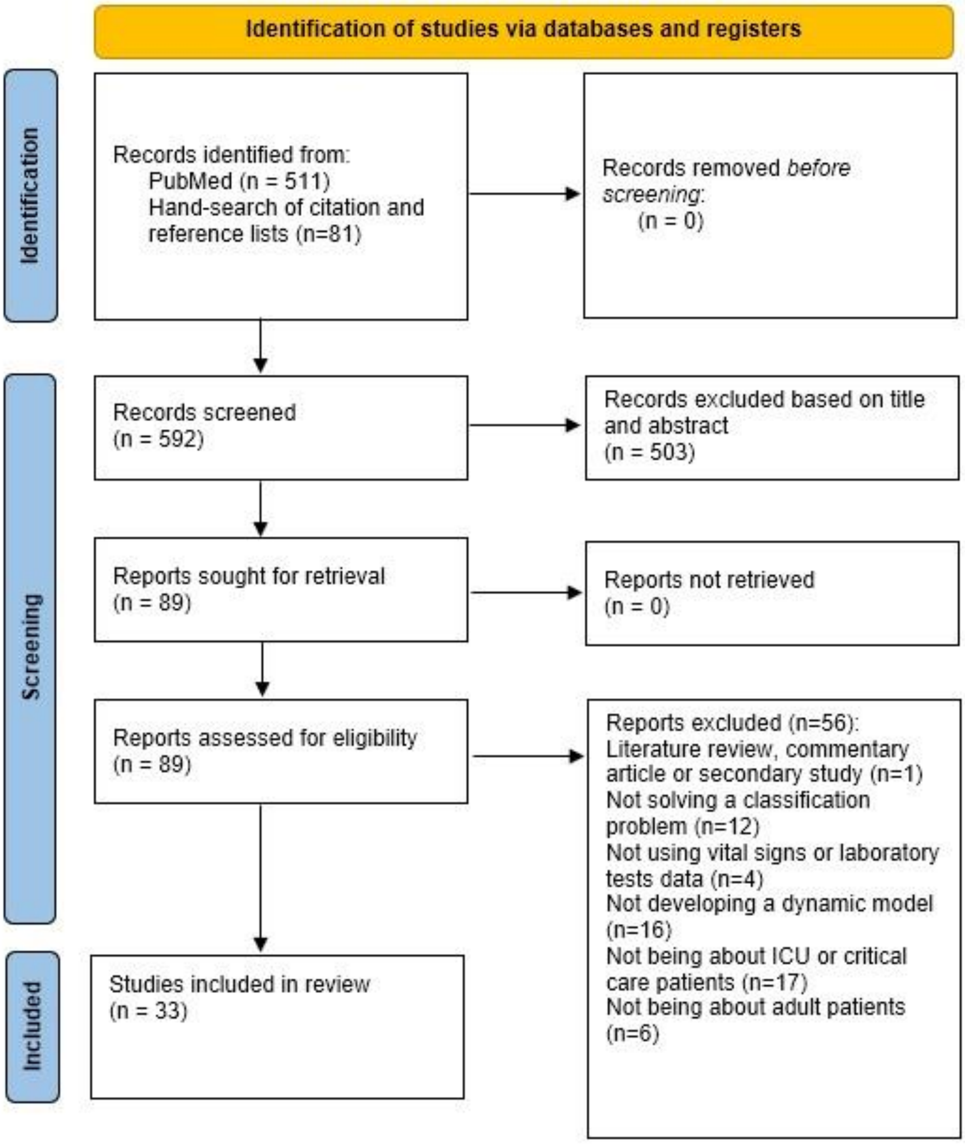
Finally, only studies that reported performance measures for their models were included in the review. If a study did not include performance measures, the study was excluded.

2.2.4. Study Selection

As seen in Figure 2.1, in the initial search, 511 articles on PubMed were listed. In addition, 81 papers were identified from hand-searching citations and reference lists of the qualifying papers. This resulted with 592 records that were screened based on the title and abstract. Based on title and abstract, 508 papers were excluded. Following reading the full text of the 89 articles, 1 paper was excluded due to being a secondary study, 12 papers were excluded due to not solving a classification problem, 4 papers were excluded due to not using dynamic vital signs and laboratory results data, 16 papers were excluded due to not developing a dynamic prediction, 17 papers were excluded due to not being about ICU or critical care patients, and 6 papers were excluded due to not being about adult patients.

Overall, 33 articles were included in the final review.

Figure 2.1. Flow diagram of study selection based on The PRISMA Statement [55].


Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
PubMed (n = 511)
Hand-search of citation and reference lists (n=81)
Records removed *before screening*:
(n = 0)
Records screened
(n = 592)
Records excluded based on title and abstract
(n = 503)
Reports sought for retrieval
(n = 89)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 89)
Reports excluded (n=56):
Literature review, commentary article or secondary study (n=1)
Not solving a classification problem (n=12)
Not using vital signs or laboratory tests data (n=4)
Not developing a dynamic model (n=16)
Not being about ICU or critical care patients (n=17)
Not being about adult patients (n=6)
Studies included in review
(n = 33)

2.3. Results

2.3.1. Brief Description of Studies

The studies included in the review are: [56], [57], [66]–[75], [58], [76]–[85], [59], [86]–[88], [60]–[65]. The information about the studies, including the first author, year, country, number of patients, types of patients, predicted outcome, data types and types of variables in studies can be found from the data extraction table (Table 2.2).

The majority of the studies were conducted in the USA (19 studies), four studies were conducted in China, and two in India. Other countries where studies were undertaken were Australia, Germany, Finland, the Netherlands, Portugal, South Korea, Thailand and the UK.

In terms of study size, five studies used more than 30,000 patient records in the development of their models [58], [67], [75], [79], [89], Johnson et al. using the largest number of patients of 50,488 [67]. Nine studies used between less than 30,000, but more than 10,000 patient records [57], [63], [66], [74], [77], [82], [83], [85], [88]. Nine studies use considerably small datasets of less than 1000 patient records [49], [62], [64], [70], [71], [73], [80], [81], [84], the smallest study population being in the study by Shashikumar et al. (242 patient records) [81].

External validation was carried out by three studies [74], [77], [88]. When other studies were single-centre studies, it is worth noting that Silva et al. used data from 42 ICUs from 9 European Union countries [82].

In terms of the data used, out of 33 studies included in this review, 19 studies developed their models, using a version of the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database [90]. Eleven studies used the MIMIC-II [56], [57], [73], [61], [63], [65], [66], [68]–[71], six studies used the MIMIC-III [58], [60], [67], [85]–[87], and two studies used the MIMIC-IV [79], [88]. Two studies used the MIMIC-III for validating their models externally [74], [77]. Of publicly available datasets, two studies also used the eICU database [91] – one for development of the model [79] and one for external validation [88].

In the next sections, the studies were discussed based on the outcomes they were predicting, how the studies approached missing data and imbalanced classification problems, the predictive modelling methods they used and the performance of the models.

Table 2.2. Data extraction table for the included studies.

First Author	Year	Country	Number of patients	Patient Population	Outcome predicted	MIMIC	eICU	Data types	Types of Variables
Bhattacharya	2018	India	4,547 patients	All ICU adult patients	Acute hypotensive episode	MIMIC-II	No	Numerical	5 vital signs
Caballero	2015	USA	11,648 patients	All ICU patients	Mortality	MIMIC-II	No	Mixed	Vital signs, laboratory results, medical notes
Deasy	2020	UK	46,476 patients	All adult patients admitted to critical care who stayed in hospital for > 48 hours	Mortality	MIMIC-III	No	Mixed	Patient demographics, vital signs, laboratory tests
Dummitt	2018	USA	7,819 patients	All ICU patients	Septic shock	No	No	Mixed	Patient demographics and vital signs
Feng	2021	China	5,653 patients	Adult patients with non-invasive ventilation for over 48 hours in ICU	Non-invasive ventilation failure	MIMIC-III	No	Mixed	Patient demographics, vital signs, laboratory tests
Ghosh	2017	Australia	1,310 patients	ICU patients	Septic Shock	MIMIC-II	No	Numerical	Mean arterial pressure, heart rate and respiratory rate
Gultepe	2014	USA	741 patients	ICU Patients with Systemic Inflammatory Response Syndrome	Mortality, Lactate Level (Sepsis)	No	No	Numerical	Vital signs and laboratory results

First Author	**Year**	**Country**	**Number of patients**	**Patient Population**	**Outcome predicted**	**MIMIC**	**eICU**	**Data types**	**Types of Variables**
Henry	2015	USA	16,025 patients	All ICU patients, any surgery	Septic shock	MIMIC-II	No	Mixed	Patient demographics, vital signs, and laboratory results
Hernandez	2021	USA	406 patients	Adult patients in ICU recovering from surgery	Haemodynamic instability	No	No	Mixed	Patient demographics, vital signs, and ECG data
Huddar	2016	India	775 patients	ICU patients	Acute respiratory Failure	MIMIC-II	No	Mixed	Clinical Notes + vital signs
Hug	2009	USA	10,066 patients	Adult ICU patients	Mortality	MIMIC-II	No	Mixed	Patient demographics, vital signs, and laboratory results
Johnson	2017	USA	50,488 patients	ICU patients corresponding to adults for surgical, medical, neurological and coronary critical illness	Mortality	MIMIC-III	No	Mixed	Patient demographics, vital signs, and laboratory results
Joshi	2012	USA	10,000 patients	Adult patients in ICU	Mortality	MIMIC-II	No	Mixed	Vital signs and laboratory results
Lee	2010	USA	1,311 patients	ICU patients	Hypotensive episodes	MIMIC-II	No	Mixed	Vital signs
Lehman	2015	USA	453 patients	ICU patients	Mortality	MIMIC-II	No	Numerical	Vital signs

First Author	**Year**	**Country**	**Number of patients**	**Patient Population**	**Outcome predicted**	**MIMIC**	**eICU**	**Data types**	**Types of Variables**
Lehman	2013	USA	337 patients	ICU patients with day 1 SAPS-I scores and at least 18 jours of blood pressure data since 24h from ICU admission	Mortality	MIMIC-II	No	Numerical	Vital signs
Ma	2019	USA	3,763 patients	Medical ICU patients	Mortality	No	No	Mixed	Patient demographics, vital signs, laboratory tests
Mao	2012	USA	772 patients	ICU patients	Mortality	MIMIC-II	No	Numerical	Vital signs
Meyer	2018	Germany	11,492 patients (development), 5,898 (external validation)	All ICU patients, Any surgery	Bleeding, mortality and renal failure	MIMIC-III (external validation)	No	Mixed	Patient demographics, vital signs, and laboratory results
Misra	2021	USA	45,425 patients	Adult ICU patients	Septic shock	No	No	Mixed	Patient demographics, vital signs and laboratory results
Mohammed	2020	USA	5,958 patients	Adult medical ICU patients	Sepsis	No	No	Numerical	Vital signs
Nemati	2018	USA	27,527 (development), 42,411 (external validation)	All ICU patients, regardless of reason being there	Sepsis	MIMIC-III (external validation)	No	Mixed	Patient demographics, vital signs and laboratory results

First Author	Year	Country	Number of patients	Patient Population	Outcome predicted	MIMIC	eICU	Data types	Types of Variables
Park	2020	South Korea	36,023 patients	Adult ICU and Ward patients	Bacteremia (Septic complication)	No	No	Mixed	Patient demographics, vital signs and laboratory data
Pattalung	2021	Thailand	18,353 MIMIC-III patients, 18,134 MIMIC-IV patients, 36,283 eICU patients	Adult ICU patients staying in ICU for > 48h	Mortality	MIMIC III and MIMIC-IV	Yes	Numerical	Vital signs and laboratory variables
Raj	2019	Finland	472 patients	Adult traumatic brain injury patients	Mortality	No	No	Numerical	Patient demographics, vital signs, and laboratory results
Shashikumar	2017	USA	242 patients	Adult ICU patients	Sepsis	No	No	Mixed	Patient demographics and vital signs
Silva	2006	Portugal, 42 ICUs of 9 EU countries (list of countries unavailable)	13,164 patients	Adult patients in ICU that did not have burns or had a bypass surgery	Mortality	No	No	Numerical	17 variables collected within the first 24h of admission
Thoral	2021	The Netherlands	14,105 admissions	Adult ICU patients	ICU readmission and mortality as a composite outcome	No	No	Mixed	Patient demographics, vital signs and laboratory results
van Wyk	2019	USA	754 patients	All ICU adult patients	Sepsis	No	No	Numerical	Vital signs and laboratory results

First Author	Year	Country	Number of patients	Patient Population	Outcome predicted	MIMIC	eICU	Data types	Types of Variables
Xia	2019	China	18,415 patients	Adult ICU patients with length of stay >10 days	Mortality	MIMIC-III	No	Numerical	Vital signs and laboratory results
Yee	2019	USA	9,165 patients	All ICU patients	Septic shock	MIMIC-III	No	Mixed	Patient demographics and laboratory results
Yijing	2022	China	1,860 patients	Adult ICU patients	Cardiac arrest	MIMIC-III	No	Numerical	ECG, and vital signs
Zhao	2021	China	11,362 patients (development), 35,252 (external validation)	Adult ICU patients who stayed in ICU >24h	Sepsis-induced coagulopathy	MIMIC-IV (development)	Yes (external validation)	Mixed	Patient demographics, vital signs, and laboratory results

2.3.2. Specific Outcomes Predicted by the Models

The Table 2.3 shows the outcomes predicted by the studies included in this review.

Table 2.3. Outcomes predicted by included studies.

Outcome	Number of papers	Author and Year	Prevalence of the Outcome
Mortality	16	Caballero 2015	Not reported
		Deasy 2020	13.0%
		Gultepe 2014	35.0%
		Hug 2009	Not reported
		Johnson 2017	Not reported
		Joshi 2012	12.0%
		Lehman 2013	14.0%
		Lehman 2015	15.0% - 19.0%*
		Ma 2019	1.2% - 17.0%*
		Mao 2012	2.3%
		Meyer 2018	6.2%
		Pattalung 2021	8.0% - 14.1%*
		Raj 2019	19.5%
		Silva 2006	Not reported
		Thoral 2021	5.3%
		Xia 2019	11.7%
Septic complications	12	Dummitt 2018	2.3%
		Ghosh 2017	15.9%
		Gultepe 2014	20.3%
		Henry 2015	14.1%
		Misra 2021	12.7%
		Mohammed 2021	10.35%
		Nemati 2018	8.6%
		Park 2020	1.9%-2.3%*
		Shashikumar 2017	22.0%
		van Wyk 2019	32.5%
		Yee 2019	1.9%
		Zhao 2021	59.0%
Cardiac complications	4	Bhattacharya 2018	28.5%
		Hernandez 2021	35.0%
		Lee 2010	24.2%-25.4%*
		Yijing 2022	9.1%
Respiratory complications	2	Feng 2021	46.7%
		Huddar 2016	11.7%
Bleeding	1	Meyer 2018	4.9%
Renal complications	1	Meyer 2018	1.0%

* If a range is reported, the authors carried out different experiments with different datasets, where prevalence of outcome varied.

When comparing the models amongst papers included in this review, it is important to keep in mind that the predicted outcomes can have varying definitions. In addition, the data pre-processing can play a role in how well the models perform.

2.3.2.1. Mortality

Most models predicting mortality had the outcome defined as mortality that happened any time in the ICU [57], [58], [62], [67], [68], [70], [79], [82], [85], [92]. Hug et al. however, added mortality within 30 days from ICU discharge to ICU mortality [66]. Lehman et al., Mao et al. and Meyer et al. predicted in-hospital mortality [71], [73], [74]. Finally, Raj et al. predicted 30-day mortality [80] and Thoral et al. predicted a composite outcome of in-hospital mortality and ICU readmission within 7 days of ICU discharge [83].

The prevalence of mortality in the studies ranged vastly, specifically between 1.2% and 35.0%. This is because the studies analysed different types of patient populations, where in some the mortality is more prevalent than in others. For example, Ma et al. found that amongst medical ICU patients, 1.2% of patients died within 6 hours since ICU admission [72]. Gultepe et al., however predicted mortality amongst sepsis patients, for which high mortality rate (35.0%) was expected [62].

2.3.2.2. Septic Complications

Septic complications are incredibly serious complications. A large study investigating sepsis-related mortality in English ICUs found that sepsis can affect a quarter of adult ICU patients in England, and can kill one in four ICU patients affected [93]. Sepsis occurs when an infection in the body results in the systemic inflammatory response syndrome and is defined to be severe if sepsis causes organ dysfunction [94]. Septic complications can have a significant impact on patient due to being associated with increased mortality and life-long complications, such as permanent organ damage, cognitive impairment, and physical disability [94].

Even though septic complications were predicted by 12 studies, the definition of the outcome varied substantially. Four papers [76], [77], [81], [88] used the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) criteria [54] to predict the onset of sepsis in general. Two studies [59], [75] used the Systemic inflammatory Response Syndrome (SIRS) criteria [95] to predict specifically septic shock. It is worth mentioning that Misra et al. treated septic shock patients as the cases and patients with other septic complications as controls [75]. The studies by Ghosh et al. [61] and Henry et al. [63] defined septic shock as the outcome in a similar way as SIRS criteria, however, they did not specifically state that they were using this widely used, agreed upon criteria. Van Wyk et al. [84] predicted the onset of sepsis by following the Sepsis-2 criteria [94], even though a new criteria (Sepsis-3) had already been published three years prior the van Wyk et al.'s study. This is a limitation to van Wyk et al.'s study as later published papers showed that the definition of sepsis by these two criteria were very different [96], and hence the developed models can misclassify patients to have different levels of sepsis.

Avoiding the conflicting sepsis definition criteria, Gultepe et al. predicted high lactate levels (≥ 4mmol/l vs <4mmol/l), which is considered to be a sign of possible sepsis [62]. Yee et al., however, made their own criteria for septic shock [86]. This is considered to be a limitation to the study as, even though the currently available criteria for diagnosis of septic shock are not perfect [97], they are still based on consensus and are heavily validated [98], [99]. Finally, Park et al. predicted bacteraemia [78], which is a septic complication, and hence was included under this category in this review.

Similarly to the models predicting mortality, for septic complications, the prevalence of outcomes also varied substantially: 1.9%-59.0%. Yee et al. predicted septic shock in the whole ICU population, resulting with a very low prevalence of septic shock of 1.9% [86]. Park et al. predicted bacteraemia also in the general ICU, resulting with low prevalence (between 1.9% and 2.3%, depending on experiment) [78]. Zhao et al., however, analysed sepsis patients only, and predicted sepsis-induced coagulopathy, which turned out to be very prevalent (59.0%) amongst sepsis patients [88].

2.3.2.3. Cardiac Complications

Two models predicted hypotensive episodes [56], [69]. Acute hypotensive episode is a sudden onset of a period of sustained low blood pressure [56]. Bhattacharya et al. defined hypotensive episode as a period of 30 minutes where at least 90% of mean arterial pressure measurements were no greater than 60mmHg. A long-lasting hypotension can result in dangerously decreased tissue blood flow with consequent end-organ damage. Treating hypotension appropriately can be effective to avoid severe sepsis [100], shock [101] and acute coronary syndrome [102].

Hernandez et al. predicted haemodynamic instability, which is related to arrhythmia, respiratory failure and hypotension [64]. In their paper they did not clearly define, however, what they considered haemodynamic instability to be.

Yijing et al. predicted cardiac arrest in critically ill patients. The cardiac arrest was defined as the start time of the first occurrence of the specified abnormal events, however the abnormal events were not described in the paper [87]. This is a limitation to the study as it makes the prediction model difficult to reproduce.

When looking at studies predicting various cardiac complications, the prevalence was also very variable: between 9.1% and 35.0%. This is because hypotensive episodes and haemodynamic instability are more common complications, especially in cardiac patients, who were included in Hernandez et al.'s, Bhattacharya's and Lee's datasets, resulting in high number of patients with the predicted outcomes [56], [64], [69]. Cardiac arrest, however, is a less common complication, especially if all ICU patients are included in the dataset, not only cardiac surgery patients [103]. Hence, Yijing et al. predicted an outcome that had prevalence of 9.1% in their study population [87].

2.3.2.4. Other Complications

Feng et al. predicted late non-invasive ventilation failure in ICU. They defined the outcome as death during or intubation after non-invasive ventilation [60]. Interestingly, in Feng et al.'s patient cohort, the prevalence of late non-invasive ventilation failure was very high (46.7%). This could be because they included patients who received non-invasive ventilation as a primary treatment following ICU admission [60].

Huddar et al. predicted acute respiratory failure, which occurs when the respiratory system fails in oxygenation and/or CO_2 elimination from the lungs [49]. It is considered to be the end point of respiratory complications, such as pneumonia or atelectasis. There are various factors than can be associated with acute respiratory failure: patient-related factors, including age, pre-existing chronic obstructive pulmonary disease, congestive heart failure and arrhythmia; and procedure-related variables, including emergency surgery, prolonged surgery and surgical site [104]. Compared to the other studies, Huddar et al. reported the common incidence of acute respiratory failure ranging between 0.2% to 3.4%, however, in Huddar et al.'s patient population, the incidence of acute respiratory failure was 11.7%. This might be because Huddar et al. retrospectively diagnosed the complication based on a specific criterion that followed the vital signs recorded automatically in the ICU [49], whereas studies in the literature are using different definition of what constitutes for a patient to have respiratory failure [105]. This shows that some complications that are reported without specific criteria based on laboratory results or vital signs can be under-reported in the electronic health records.

In addition to mortality, Meyer et al. also predicted postoperative bleeding and renal failure requiring renal replacement therapy [74]. The renal failure was defined using Kidney Disease: Improving Global Outcomes (KDIGO) criteria [106]. Acute kidney injury, formerly called acute renal failure, is a sudden decline in glomerular filtration rate [107]. Glomeruli are tiny filters in the kidneys that filter waste from the blood. This rate estimates how much blood passes through the glomeruli each minute. Acute kidney injury is usually caused by an event that leads to kidney malfunction, such as dehydration, blood loss from major surgery or injury, or the use of medicines [108].

Even though acute renal failure in cardiac patients is often considerably low [109], Meyer et al.'s prevalence for renal failure was very low (1.0%). This might be due to different studies defining acute renal failure differently. Meyer et al., however, used the KDIGO criteria, which is an internationally recognised criteria for diagnosing renal complications, including renal failure [106].

2.3.3. Methods Used to Develop Dynamic Models Predicting Patient Outcomes

2.3.3.1. Missing Data Approaches

Missing data in ICUs, being an incredibly data-rich environment, is inevitable [110]. The large number of missing values in medical databases represent a challenge as a patient's health state needs to be examined even when no observations are available [57]. Furthermore, data collection errors are common in clinical practice [111], and hence using various approaches to treat missing data can be beneficial when dealing with these errors.

Table 2.4 Missing data approaches used by each included paper.

Missing Data Approach	Number of Papers	Author and Year	Amount of missing data
Imputation	9	Caballero 2015	34% estimated
		Dummitt 2018	0-94.7%
		Lehman 2015	Not reported
		Ma 2019	0-93.21%
		Misra 2021	Not reported
		Park 2020	Not reported
		Pattalung 2021	Not reported
		Shashikumar 2017	Not reported
		Zhao 2021	Not reported
Carrying Forward/Interpolation	7	Dummitt 2018	0-94.7%
		Joshi 2012	Not reported
		Meyer 2018	Not reported
		Park 2020	Not reported
		Pattalung 2021	Not reported
		Yijing 2022	Not reported
		Xia 2019	Not reported
Informative Missingness	2	Deasy 2020	Not reported
		Huddar 2016	Not reported
Removal of variables	3	Dummitt 2018	0-94.7%
		Feng 2021	Not reported
		Misra 20021	Not reported
Removal of entries	3	Lehman 2015	Not reported
		Raj 2019	Mean = 70 values for intracranial pressure, 78 values for mean arterial pressure and 70 for cerebral perfusion pressure
		Silva 2006	4 entries
Model "handles"	1	Zhao 2021	Not reported
Unclear	4	Bhattacharya 2018	Not reported
		Gultepe 2014	Not reported
		Thoral 2021	Not reported
		Yee 2019	Not reported
None reported	11	Ghosh 2017	Not reported
		Henry 2015	Not reported
		Hernandez 2021	Not reported
		Hug 2009	Not reported
		Johnson 2017	Not reported
		Lee 2010	Not reported
		Lehman 2013	Not reported
		Mao 2012	Not reported
		Mohammed 2020	Not reported
		Nemati 2018	Not reported
		van Wyk 2019	Not reported

The common methods for handling missing data found in this review are:

- Removal of variables with large numbers of missing data
- Removal of records with large numbers of missing data
- For time-series data, if some data are available for a patient, carrying forward or interpolation methods
- If no time-series data available, imputation methods
- Using models that are robust when handling missing data

How the studies included in this review dealt with missing data are shown in Table 2.4. Nine studies used an imputation method, seven papers reported carrying forward the previously reported values, three papers removed variables with a certain number of missing values, three papers removed records with a certain number of missing values, two papers used informative missingness (i.e., where missing data can give information itself, e.g., tests not taken due to patient being healthy), and one paper used a prediction method that took missing values into account. In four papers, missing data approaches were discussed, however it was unclear what exactly was done about it, and for eleven papers the missing data approaches were not mentioned at all.

Imputation Methods

Imputation methods are used to fill in missing values with another, probable value. Using imputation methods can be beneficial as this allows including patients who can have relevant features for analysis but could be otherwise be excluded from analysis due to data collection or recording errors [112]. Imputation methods were used by almost a third of the studies (10 papers).

Shashikumar et al. [81] and Park et al. [78] used mean imputation to replace missing values. Mean imputation is a single imputation method where the missing values are replaced by the mean of the observed values of the variable with missing data [112]. Even though mean imputation is a very straight-forward method, it can cause severely biased estimates due to changing the variance of the data [113]. That being said, if the missingness of data is low (<10%) and the variables with missing values are not highly correlated with the predicted outcome, the effects on the reliability of predicted outcome are marginal [114].

Dummitt et al. [59] and Zhao et al. [88] used the population median to populate missing values. This was also done for clear outliers that were removed from dataset and replaced with new values based on the population median [59]. Median imputation is very similar to mean imputation, but instead of replacing the missing values with the population mean, the median is used. Similar to mean imputation, median imputation is also a very straight-forward approach, however, takes into account that real-life data are not always normally distributed [115]. It has been shown to perform similarly well as more sophisticated imputation methods [116].

Interestingly, Pattalung et al. replaced the missing values with the value "-1" [79]. In their paper they do not explain further why this decision was made and what assumptions they had when undertaking this approach. In practice, replacing missing values with a certain agreed-upon value would simplify the usage of a prediction model if a clinician was faced with missing values. However, this can significantly alter the probability for a patient to have the predicted outcome. In Pattalung et al.'s paper it was found that their model had slight differences in the variable importance values when trained on two different databases [79], which could be a result by the missing data approach. This makes the developed prediction model less usable if applied to other institutions.

Caballero et al. used Regularised Expectation Maximization (EM) to fill out the missing values [57]. EM is based on iterated analyses of linear regressions of variables with missing values on variables with available values, with regression coefficients estimated by ridge regression, a regularised regression method in which a continuous regularisation parameter controls the filtering of the noise in the data. The regularisation parameter is determined by generalised cross-validation, such as to minimise, approximately, the expected mean-squared error of the imputed values. The regularised EM algorithm has been shown to be able to estimate missing values for various types of missing data problems [117].

Lehman et al. used Gaussian noise imputation for time-series data to replace missing or invalid values in their dataset [71]. In the literature there is a wealth of information about Gaussian processes handling missing values, however Lehman et al. used logistic regression (Section 2.3.3.3) for their prediction model. Since they did not provide a reference to what exactly they mean by using Gaussian noise to "fill in the missing or invalid values” [71], it is unclear what processes they carried out to handle missing data.

Ma et al. used a tree-based estimation algorithm to replace missing values [72]. To do that, they replaced the missing values with the value “-1000” as this value was very different from the non-missing values. This was done for the estimation algorithm to treat these missing values differently. However, Ma et al. do not provide a reference to what kind of tree-based estimation algorithm they used for the imputation [72]. There are various tree-based algorithms that have been developed to impute missing data, including methods developed by D’Ambrosio et al. [118], Vateekul et al. [119], and Rahman et al. [120]. Not specifying which tree-based method was used for missing data imputation makes the developed models not reproducible and reduces the chance for the models to be implemented in clinical practice due to lack of transparency [121].

Misra et al. used random forest imputation to replace missing values in their data [75]. Random forest imputation, i.e., missForest is a pattern-based method, which can be applied to any kind of data (numerical or categorical). It requires no tuning of parameters or assumptions about the data distribution [122]. MissForest has been shown to outperform most other methods of missing data imputation, showing low imputation error and maintaining predictive ability in clinical prediction models [123], [124].

Carrying Forward and Interpolation

A few studies used the “carrying forward” method, where the patients’ most recent reading from earlier in the database is used, if available. This method was used by Dummitt et al. [59], Joshi et al. [68], Meyer et al. [74], Park et al. [78], Pattalung et al. [79], and Yijing et al. [87].

An alternative to the carrying forward method is linear interpolation, used by Xia et al. [85]. In one-dimensional data sequence, linear interpolation estimates the missing value based on the two data points adjacent to the points that has a missing value [125]. Therein lies the difference between interpolation and carrying forward: for carrying forward, the missing values are replaced with the previously recorded value; for interpolation, however, the missing values are replaced with a value that has been calculated based on also the next available value. The primary assumption of carrying forward method is that the value did not change from the previously recorded value, which is likely incorrect [126]. The interpolation method can be more reflective of the changes in patient's health as it takes into account the next recorded value but also makes an assumption that the trajectory between the two data points is linear [125].

Removal Methods

Three papers excluded variables from the analysis that had a high level of missing data. Dummitt et al., for example, removed variables with >89% of missing data, however kept other variables in the analysis that were deemed essential for their prediction task [59]. Feng et al. and Misra et al. removed variables with missing rate over 40% [60], [75].

Three papers also excluded records with a high number of missing values. Silva et al. removed entries due to missing values, however the total number of remaining records was very high (13,164 records) [82]. Lehman et al. excluded patients with more than 15% of missing or invalid samples [71]. Raj et al., however, excluded the patient only if the missing values fell in a specific time window [80].

Informative Missingness

Two papers used informative missingness to approach missing data. Deasy et al., for example, explained that the missing data included in their recurrent neural networks model was treated as separate discrete events where models used these as "informative missingness" [58]. Informative missingness, as Huddar et al. argue, means that the data are missing because clinicians deemed the test unnecessary. This provides information on the patient's status by showing that the patient was too healthy to need a test or receive medication. The missingness can be incorporated, for example, by creating a separate category for a variable that has missing values [127].

Incorporating informative missingness needs to be decided upon based on how the developed prediction model will be used in practice. If a system is incorporated in an electronic health system, and calculates probabilities for predicted outcomes automatically, the effect of informative missingness can be hidden from the clinician. However, if risk prediction is done by hand using a scoring system, the clinician is able to make an informed decision by also including informative missingness [127].

Since Deasy et al. developed a prediction model that incorporates all data from the electronic health record [58], and Huddar et al. developed a model that is incorporated with the electronic health record [49], it is difficult to know how much the developed models actually take this informative missingness into account.

Model-Handled Methods

Zhao et al. explain that the boosting machine learning methods they were using in their analysis can use missing data when making the prediction. These models were CatBoost, light Gradient Boosting, Extreme Gradient Boosting (XGBoost), and Gradient Boosting Machine (GBM) [88].

Even though Zhao et al. stated that no entries with missing data were removed from analysis, they did not report what the rate of missingness in the data was. It has been shown that the amount and the distribution of missing data plays an important role when developing predictive models in terms of variability and bias of the results [113], [114], [128].

Other Approaches

Four papers did not explain their methods for handling missing data very clearly. Bhattacharya et al. mentioned that data were “cleaned”, but no further information was given [56].

Gultepe et al. stated that no missing data were included in the study, meaning that assumably only records with complete data were included in the study [62].

Thoral et al. mentioned that values with no biological plausibility were removed from analysis, however, did not report what was done with these missing values in the data as they did not mention removing variables with these values or patient records with these values [83].

Yee et al. reported interpreting missing values as “not measured”, but they did not explain how these “not measured” values were handled in their analysis [86].

In total, eleven studies did not report any action taken regarding missing values [61], [63], [84], [64], [66], [67], [69], [70], [73], [76], [77].

2.3.3.2. Imbalanced Classification Problem Approaches

In total, 14 papers mentioned facing an imbalanced classification problem in their analysis and reported how they approached this problem (Table 2.5).

Table 2.5. Approaches for imbalanced classification problem used by papers.

Imbalanced Classification	Number of Papers	Author and Year
Balanced by method	7	Huddar 2016
		Lee 2010
		Mao 2012
		Misra 2021
		Mohammed 2020
		Pattalung 2021
		Silva 2006
Performance Measures	3	Johnson 2017
		Ma 2019
		Thoral 2021
Data selection	2	Meyer 2018
		van Wyk 2019
Modelling method approach	2	Caballero 2015
		Dummitt 2018

Balanced by Method

Seven papers reported using a specific method to manipulate the sample to achieve a balanced dataset. Two papers used Synthetic Minority Oversampling Technique (SMOTE) on their training sets [49], [75]. Misra et al. also reported using upsampling [129], however, it is unclear why both SMOTE [130] and upsampling were used, and how these two methods were used at the same time [75]. Over-sampling was also used by Pattalung et al. on their training data [79].

Lee et al. reported using subsampling [131] in their training data to achieve a balanced data set [69]. Undersampling [132] was used by Mao et al. [73], and Silva et al. [82]. Mohammed et al. used a Bayesian bootstrap method [133] to balance controls with cases [76].

Most papers carried out these sampling methods on training sets only, and left the testing sets as original, as is recommended [134], however, Mao et al. [73] and Mohammed et al. [76] carried out the balancing methods on both training and testing data. This approach is not recommended [134] as this means that their models were tested on balanced datasets which do not reflect the proportion of cases and controls as it does in practice. In addition, this means that the predicted probabilities can be incorrect and not be applied in a real-world situation [135].

Data Selection

Two studies approached the imbalanced classification problem by sampling their data as equally sized case and control groups [74], [84]. This means that both of these studies worked with balanced training and testing datasets, which makes the predicted probabilities by models not applicable in a real-world setting [135].

Choosing Appropriate Modelling Method

Caballero et al.[57] and Dummitt et al. [59] approached the imbalanced classification problem by choosing predictive modelling methods that have been shown to be robust when handling imbalanced datasets. Caballero et al. report that the naïve Bayes classifier they were using for text classification had shown good predictive performance for unbalanced classes [57]. Dummitt et al. checked that the number of events per variable was kept above the recommended thresholds for the classification methods they used to ensure the model coefficients would not be biased by the case balance [59]. It was, however, not explained how this was achieved.

Choosing Appropriate Performance Measures

Three papers reported that due to the imbalanced classification problem, instead of reporting only the accuracy of the model, they also reported area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) [67], [72], [83]. These performance measures have been shown to give a better reflection of how the model recognises both positive and negative classes [136], [137].

2.3.3.3. Classification Methods Used by Studies to Predict Patient Outcomes in a Dynamic Manner

Table 2.6 summarises which classification method each paper used. The most used methods were logistic regression (18 papers), random forest (11 papers), support vector machines (10 papers), and neural networks (10 papers). Other more commonly used methods included gradient boosting machines (5 papers), and naïve Bayes (4 papers).

Table 2.6. Classification methods used by papers to predict patient outcomes dynamically.

Method	Number of studies	First Author and Year
Logistic regression (all versions)	18	Caballero 2015
		Dummitt 2018
		Feng 2021
		Huddar 2016
		Hug 2009
		Johnson 2017
		Joshi 2012
		Lehman 2013
		Lehman 2015
		Mao 2012
		Misra 2021
		Raj 2019
		Shashikumar 2017
		Silva 2006
		Thoral 2021
		van Wyk 2019
		Zhao 2021
Random Forest	11	Caballero 2015
		Dummitt 2018
		Feng 2021
		Hernandez 2021
		Huddar 2016
		Ma 2019
		Misra 2021
		Mohammed 2020
		Thoral 2021
		van Wyk 2019
		Zhao 2021
Support Vector Machines	10	Ghosh 2017
		Gultepe 2014
		Hernandez 2021
		Huddar 2016
		Mao 2012
		Misra 2021
		Mohammed 2020
		Thoral 2021
		van Wyk 2019
		Zhao 2021
Neural Networks (any kind)	10	Deasy 2020
		Meyer 2018
		Park 2020
		Pattalung 2021
		van Wyk 2019
		Feng 2021
		Lee 2010
		Silva 2006
		Feng 2021
		Xia 2019
Gradient Boosting Machine (all versions)	5	Feng 2021
		Johnson 2017
		Thoral 2021

Method	Number of studies	First Author and Year
		Yijing 2022
		Zhao 2021
Naïve Bayes	4	Caballero 2015
		Gultepe 2014
		Hernandez 2021
		Zhao 2021
Cox proportional hazards	3	Dummitt 2018
		Henry 2015
		Nemati 2018
Decision Trees	3	Huddar 2016
		Misra 2021
		Zhao 2021
AdaBoost	2	Hernandez 2021
		Huddar 2016
Bayesian Networks	2	Gultepe 2014
		Yee 2019
Hidden Markov Models	2	Ghosh 2017
		Gultepe 2014
C5.0	1	Misra 2021
CatBoost	1	Zhao 2021
Dual boundary classifier	1	Bhattacharya 2018
Gaussian Mixture Model	1	Gultepe 2014
LASSO	1	Johnson 2017
LUCCK (Learning Using Concave and Convex Kernels)	1	Hernandez 2021

The outcomes were predicted in varying frequencies. The closest to "real-time" models were those that updated their prediction every time new measurements were entered into the system. Eleven studies followed this prediction frequency [49], [56], [87], [61], [63], [64], [73], [74], [82], [85], [86]. Eight studies developed models to predict outcomes on an hourly basis [57], [58], [67], [69]–[71], [79], [83].

Twelve studies predicted the outcomes less often [59], [60], [84], [88], [62], [72], [75]–[78], [80], [81]. More specifically, Ma et al. predicted mortality in every 6 hours [72]. Nemati et al. predicted sepsis 12, 8, 6 and 4 hours before the onset [77]. Park et al. predicted bacteraemia 8, 16 and 24 hours in advance [78]. Raj et al.'s model made new predictions of mortality every 8 hours [80]. For Shashikumar et al.'s model, sepsis was predicted 4 hours in advance [81]. Dummitt et al. made the prediction of septic shock 4, 8 and 24 hours beforehand [59]. Feng et al.'s model predicted late non-invasive ventilation failure in 8, 16, 24, 36 and 48 hours after the start of non-invasive ventilation [60]. Gultepe et al. predicted mortality and high lactate levels in 6, 12 and 24 hours [62]. Misra et al. predicted septic shock within 1, 3, and 6 hours before the onset [75]. Mohammed et al. predicted sepsis at around 18 hours beforehand [76]. Van Wyk et al. predicted sepsis 3 and 6 hours in advance [84]. Zhao et al. predicted sepsis-induced coagulopathy on a daily basis [88].

For two studies it was unclear how often their dynamic models predicted the outcome [66], [68].

2.3.4. Performance of the Models

By far the most reported performance measure was area under the receiver operating characteristic curve (AUROC), reported by 29 papers. It was common to also report sensitivity (18 papers) and specificity (17 papers). Less commonly reported performance measures included accuracy (10 papers), positive (7 papers) and negative predictive values (5 papers), area under the precision-recall curve (AUPRC) (4 papers), and F1 score (5 papers).

A number of papers tested various methods to predict patient outcomes (see Section 2.3.3.3), however, Tables 2.7 to 2.9 show the highest performing models and their respective performance measures for the papers.

When looking at how the models performed based on predicting mortality (Table 2.7), Meyer et al. had the highest AUROC of 0.950 when predicting mortality, achieved with recurrent deep neural network [74]. The second-best performance was achieved by Johnson et al. with the AUROC of 0.920 (gradient boosting machine) [67], followed by Pattalung et al. (AUROC = 0.910, recurrent neural network) [79] and Ma et al. (AUROC = 0.905, random forest) [72].

In terms of sensitivity, the model by Gultepe et al. has by far the highest sensitivity of 0.949, achieved with support vector machine [62]. The model developed by Mao et al. has the highest specificity of 0.950 (support vector machine) [73]. Based on accuracy, Meyer et al. had the highest performance of 0.880 (recurrent deep neural network), and they also achieved very high positive and negative predictive values of 0.900 and 0.860, respectively [74]. Only four papers reported AUPRC when predicting mortality, Johnson et al. with the highest of 0.665 (gradient boosting machine) [67], and out of the two papers that reported the F1 score, Deasy et al. achieved the highest of 0.821 with recurrent neural network [58].

Looking at the papers that predicted septic complications (Table 2.8), Park et al. achieved very high AUROC of 0.960 (recurrent neural network) when predicting bacteraemia, which is a septic complication [78]. Misra et al. also achieved a high performance (AUROC = 0.948, random forest) when predicting septic shock [75]. Based on sensitivity, Park et al. also had the highest performance (Sens = 0.940) [78], and Misra et al. had the highest specificity of 0.796 [75].

The papers developing models to predict some other complications achieved considerably high AUROC, sensitivity and specificity (Table 2.9). Interestingly, Meyer et al., when predicting renal complications, achieved very high accuracy (0.900, recurrent neural network), AUROC (0.960), sensitivity (0.940), specificity (0.860), and positive and negative predictive values (0.870 and 0.940, respectively).

Notably, the models by Meyer et al. achieved high performance measures for the predicted outcomes. However, as explained in Section 2.3.3.2, Meyer et al. used a balanced dataset for both training and testing data, meaning their model performance is not necessarily reflective of the real-world situation [135]. In their patient demographics, mortality was present in 6.2% of patients, bleeding in 4.9% and renal failure in 1.0%. These proportions show highly imbalanced data, meaning that the models tested on a testing set where 50% of the patients experienced renal failure reach AUROC of 0.960 is not applicable on a real-world situation where renal failure occurs in only 1% of patients.

Table 2.7. Best-performing classification method and their respective highest reported performance of papers predicting mortality.

Mortality									
Author and Year	**Classification Method**	**Accuracy**	**AUROC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUPRC**	**F1 score**
Caballero 2015	Logistic Regression		0.866	0.789	0.791				
Deasy 2020	Recurrent Neural Network		0.770						
Gultepe 2014	Support Vector Machine	0.728	0.726	0.949	0.308				0.821
Hug 2009	Logistic Regression		0.885						
Johnson 2017	Gradient Boosting Machine		0.920					0.665	
Joshi 2012	Logistic Regression		0.890						
Lehman 2013	Logistic Regression		0.800						
Lehman 2015	Logistic Regression		0.700						
Ma 2019	Random Forest		0.905					0.381	
Mao 2012*	Support Vector Machine		0.633	0.143	0.950	0.415	0.791		
Meyer 2018*	Recurrent Deep Neural Network	0.880	0.950	0.850	0.910	0.900	0.860		
Pattalung 2021	Recurrent Neural Network		0.910	0.810	0.860	0.850	0.820		
Raj 2019	Logistic Regression		0.840						
Silva 2006	Artificial Neural Network	0.792	0.871	0.781	0.795				
Thoral 2021	Gradient Boosting Machine		0.789					0.202	
Xia 2019	Long-Short Term Memory	0.753	0.845	0.776	0.750	0.294		0.486	0.426

Table 2.8. Best-performing classification method and their respective highest reported performance of papers predicting septic complications.

Septic Complications									
Author, Year	**Classification Method**	**Accuracy**	**AUROC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUPRC**	**F1 score**
Dummitt 2018	Generalised Linear Model via Penalised Maximum Likelihood		0.860						
Ghosh 2017	Coupled Hidden Markov Models	0.871							
Gultepe 2014	Gaussian Mixture Model	0.843	0.849	0.928	0.500				0.905
Henry 2015	Cox Proportional Hazards		0.830	0.850	0.670				
Misra 2021	Random Forest		0.948	0.839	0.881				
Mohammed 2021*	Random Forest	0.768		0.739	0.796	0.788			0.760
Nemati 2018	Weilbull-Cox Proportional Hazards	0.670	0.850		0.670				
Park 2020	Recurrent Neural Network		0.960	0.940					
Shashikumar 2017	Elastic Net Logistic Classifier		0.780	0.850	0.550				
van Wyk 2019*	Random Forest			0.800					0.680
Yee 2019	Bayesian Network		0.810	0.790	0.660	0.460	0.900		
Zhao 2021	Categorical Boosting		0.869	0.820	0.757				

Table 2.9. Best-performing classification method and their respective highest reported performance of papers predicting respiratory, cardiac, bleeding and renal complications.

Author and Year	Classification Method	Accuracy	AUROC	Sensitivity	Specificity	PPV	NPV	AUPRC	F1 score
Respiratory complications									
Feng 2021	Time Updated Light Gradient Boosting Machine		0.912						
Huddar 2016	Support Vector Machine		0.873						
Cardiac complications									
Bhattacharya 2018	Dual Boundary Classifier	0.870		0.830	0.900				
Hernandez 2021	Random Forest		0.890						
Lee 2010	Artificial Neural Network	0.758	0.819	0.748	0.746	0.665	0.833		
Yijing 2022	Extreme Gradient Boosting	0.960	0.940	0.860	0.850				
Bleeding									
Meyer 2018*	Recurrent Deep Neural Network	0.800	0.870	0.740	0.860	0.840	0.770		
Renal complications									
Meyer 2018*	Recurrent Deep Neural Network	0.900	0.960	0.940	0.860	0.870	0.940		

2.4. Discussion

In this literature review, studies developing dynamic prediction models predicting patient outcomes in critical care were reviewed. The studies were discussed based on the outcomes they predicted, how they approached missing data and imbalanced classification problems, the methods they used to develop the prediction models, and the performance their models achieved.

2.4.1. Predicted Outcomes

By far, the most predicted outcome by papers included in this review was mortality. There are various reasons why predicting mortality is so common.

Firstly, mortality is very straight-forward to define, and is a binary outcome: “dead” or “alive”. Having a clearly defined and binary outcome is a lot easier to predict as opposed to more complex multi-level outcome that has a more varying definition (e.g., when predicting morbidity). Secondly, mortality is the ultimate negative outcome, which should be avoided. Thirdly, historically, mortality has always been a way to audit and measure the performance of healthcare centres [6].

However, nowadays, mortality rates are getting lower [138], and become less relevant when looking at the ways to sustain healthcare systems. With the aging population, morbidity, on the other hand, is becoming more prevalent and is the reason why healthcare systems around the world are struggling to sustain their current model [4], [139].

The definition of the predicted outcome can be what makes or breaks a prediction model: because the definition of mortality is clear, there is no bias in the recorded outcome. As was seen in the papers predicting septic complications, the studies had various ways to define sepsis. These definitions included internationally approved definitions and classifications of sepsis, such as SIRS, Sepsis-2 and Sepsis-3, however, these agreed-upon definitions and classifications are not perfect [97], and are constantly evolving [54]. Even though sepsis is a widely researched complication, as evidenced by the large number of papers predicting septic complications in this review, sepsis patients are still often identified too late [140]. The problem of varying definitions of sepsis outcomes might be also explain the lack of prediction of critical care complications in general. For example, acute kidney injury is a relatively common complication [141], and is now easily identified using the Kidney Disease: Improving Global Outcomes (KDIGO) criteria [106], which hopefully enables the development of more prediction models for this complication.

Even though electronic health records have come a long way, databases still do not take into account the current consensus definitions of various complications, such as acute kidney injury, sepsis or the definition of complications in general, which lead to the prediction models being unusable in practice [142]. Both the sepsis and kidney disease criteria can be calculated once necessary laboratory measurements are taken. This is also the case for other complications that have an agreed criteria for diagnosis, such as liver failure [143]. This means that the time of the onset of the predicted complication can be compromised and shows that further effort in defining complications to enable timely and accurate diagnosis for these outcomes is required.

2.4.1. Missing Data

Surprisingly, a third of the included studies did not report how missing data were handled in their research. This is a clear limitation of these studies as missing data in electronic health records is highly prevalent [110]. Reporting how missing data were handled when developing a clinical prediction model is a critical step for transparency, as also required by the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) reporting guidance [121].

As stated by Tsvetanova et al., currently there is no clear guidance on how to handle missing data when developing, validating and implementing clinical prediction models [144]. As seen in this review, various imputation methods, specifically mean and median imputation, are very popular ways of handling missing data. Single imputation methods, such as mean and median imputation, and even replacing missing values with a specific value, like Pattalung et al. did, can introduce some bias to the results. However, alternative methods, such as missForest imputation and k-nearest neighbours imputation, are more computationally intensive and therefore potentially incompatible when producing rapid, deployable prediction models for serious complications, like septic shock, in real-time [59], [114], [144].

In general, the description of methods used to approach missing data were lacking in the found papers. In total, 4 papers did not state their methods clearly to understand what exactly was done. Eleven papers did not report missing data at all. Besides, the papers that did report their methods were not entirely clear whether the methods were applied on the full data or only to training or testing data.

In addition to the methods, the rate of missing data in papers was not very well documented. The TRIPOD adherence assessment requires transparency on (1) whether there is missing data, (2) the method for handling missing data, (3) details of the software used to handle missing data, and (4) description of which variables were affected by the missing data methods [121]. Only five papers stated clearly how much missing data there was in their datasets [57], [59], [72], [80], [82]. This means that for the majority of papers, it is unknown how many variables or patient records were removed from their data, how much of the data were replaced with imputation methods, carry-forward or interpolation methods, and how much data had to be handled by the prediction models. This is a serious limitation to the studies in general as it reduces the transparency of the model development, and hence makes it less clear how applicable the models are to use in practice [121].

Finally, only one paper used models that are robust enough to handle missing data [88]. There are now various methods available that have been developed to take missing data into account, including CatBoost [88], C5.0 [145] and BARTm [146].

2.4.2. The Imbalanced Classification Problem

A third of the studies were dealing with highly imbalanced classification problems, where the prevalence of the predicted outcome was <10%. In total, 14 studies reported taking an action to help with the imbalanced classification problem. Four main methods were found to be used to approach the class imbalance: (1) using balancing methods, such as upsampling, downsampling or SMOTE; (2) selecting equal case and control groups; (3) using performance measures reflective of class imbalance; and (4) using a modelling method that has been shown to perform well with imbalanced outcome.

There were some limitations in the studies found in this review in terms of the imbalanced classification problem. Surprisingly, four studies did not report the prevalence of the outcome they were predicting. Knowing the prevalence and the methods to pre-process the data (e.g., missing data and balancing methods) helps to understand how applicable the models are in clinical practice with the real-world data. In addition, when most studies applied the balancing methods correctly, the studies by Mao et al. [73] and Mohammed et al. [76] applied the balancing method on the full dataset, before partitioning the training and testing dataset. This is not a recommended approach [134]. In addition, the study by Meyer et al. [74] and van Wyk et al. [84] chose equal case and control cohorts for the full dataset, again, before dividing the data into training and testing dataset. For these four papers, this means that their models were evaluated on the testing dataset that was balanced, which does not reflect the real-world data.

In general, it is surprising that so many studies used balancing methods to solve the imbalanced classification problem, especially, if so many predictive modelling methods, including common ones, like logistic regression and random forest, have been shown to handle imbalanced classes well [147]–[149].

It is known that balancing methods or developing models on training sets that have a balanced outcome can lead to poor calibration, where the probability of the predicted outcome is overestimated. As said by van den Goorbergh et al., *"Outcome imbalance is not a problem in itself"* and *"imbalance correction may even worsen model performance"* [135]. This can also explain that Meyer et al.'s results of AUROC of 0.95, 0.87 and 0.96 for mortality, bleeding and kidney failure, respectively, may be an overestimation of how well the outcomes were predicted, especially because in their original data, the prevalence of these complications was 6.2% for mortality, 4.9% for bleeding and 1.0% for renal failure.

The models by Mao et al., Mohammed et al., and van Wyk et al. achieved only moderate performance measures. This is surprising as they tested their models on balanced datasets.

2.4.3. Classification Methods and Prediction Frequency

The most common classification method to predict clinical outcomes was logistic regression. This is not surprising as logistic regression has been shown to have very competitive performance compared to more complex machine learning methods [150]–[152]. In addition, logistic regression is a highly interpretable model, showing which variables are associated with the predicted outcome with easily interpretable odds ratios. Understanding why a prediction model predicts a certain level of probability for a patient to have an outcome is important in practice, so that clinicians know which factors need to be paid attention to.

The majority of the studies predicted outcomes in a certain frequency. Even though all studies in this review developed dynamic, “real-time” models, in reality the outcomes were predicted less frequently than on a real-time basis. The reason for this is simple: when vital signs are collected very often (e.g., every few minutes) [84], then laboratory results are collected less frequently. Some laboratory results could be collected every few hours, and some daily [58], [79]. This makes a fully real-time prediction impossible.

Often when predicting the outcome every time when new information is entered into the system, not all variables are updated, which means that in reality the variable values with no new information were carried forward from the previous timestamp, as done by a number of studies in this review. As stated by Haukoos et al. [126], this assumes that the patient state in terms of the carried forward variable stays the same, while in reality this might not be the case.

2.4.4. MIMIC Databases

The MIMIC databases were commonly used in studies included in this review. While using publicly available databases to develop clinical prediction models helps with the transparency and reproducibility of the models [43], there are a few limitations to using certain MIMIC databases. Namely, a third of the papers used the MIMIC-II database, which includes ICU patients' data collected between 2001 and 2008 [153]. Even though this database was the only one available during the time when nine of the studies were published, for two studies, the newer version – MIMIC-III – was already available for almost two years [56], [61].

The MIMIC-III database was first released in 2015 and includes ICU patients' data collected between 2001 and 2012 [90]. MIMIC-IV database was first released in 2020 and includes patient data collected between 2008 and 2019. It also includes clinical data prior to ICU admission [154].

Understandably, there was a substantial gap between the release of MIMIC-III and MIMIC-IV, and hence the papers were using data that were up to a decade old (e.g., Yijing et al.'s paper was published in 2022 and used MIMIC-III [87]). This is a limitation to the studies as patient population is ever-changing [138], [155], [156], and clinical interventions, practice and policies change constantly [157]. In addition, with more studies investigating electronic health records, the data quality in clinical systems are improving [158]. Hence, using a data that was recorded many years ago might make the developed clinical prediction models not usable in current patient population.

Meyer et al. and Nemati et al. validated their locally developed prediction models on the MIMIC-III databases. Meyer et al.'s original data were collected between 2000 and 2016, which does include the years of when MIMIC-III data were collected, making it hopefully more relevant to validate on [74]. However, Nemati et al. developed their model using the data collected between 2013 and 2015, and hence testing their model on an older database is counterintuitive. This might be also the reason of the moderate performance Nemati et al.'s model achieved [77]. It is difficult to comment on the Meyer et al.'s models' performance due to their approach regarding imbalanced classification problem (as discussed in Sections 2.3.3.2 and 2.4.3).

An alternative database to the MIMIC is the eICU database, released in 2018. The eICU database includes ICU data collected between 2014 and 2015 [91]. Even though the dataset is newer, only two studies used this dataset [79], [88]. The lack of usage of eICU might be that the MIMIC databases have been widely used in the literature for over a decade, whereas the eICU has been available for four years only.

Another limitation of using the MIMIC and eICU is that they are both US-based databases. Although, eICU consists of data from 208 US hospitals, the MIMIC databases consist of patient data only from the Beth Israel Deaconess Medical Center. Even though a third of the studies were based in the USA and used the MIMIC databases, ten studies were conducted outside of the USA and still used the MIMIC databases. This means that the majority (21 out of 33) of the studies have developed US-centric prediction models which might not necessarily be applicable in other countries, or even in the general US patient population.

Overall, the availability of open-source large ICU databases brings a lot of opportunities for clinical data analytics innovation. These databases are great sandpits to test and develop new methodologies and approaches to improve clinical outcomes [43]. However, to be able to apply models in practice, more recent and diverse data should be used to ensure the applicability of the models in a current, up-to-date patient population.

2.4.5. Limitations

In general, from the studies found, it was difficult to determine whether the dynamic prediction models found in this review have been put into practice: as found by van Smeden et al., studies analysing how prediction models fit in the existing clinical workflow are rare [25], and hence scientific literature regarding implementation of such prediction models is lacking. This problem occurs in studies developing AI-based prediction models in general, as these models are rarely validated, as also explained in Section 2.1. Hence, in the future, more studies should be conducted that analyse the impact of the prediction models on patient outcomes and clinical workflow.

As this thesis focuses on analysing cardiac patient population, in this review, the study population was kept considerably narrow: only adult critical care or ICU patients were included. While excluding settings, such as emergency department and remote monitoring, could cause missing out on interesting dynamic prediction modelling approaches in these other settings, the data collected in the emergency department and remote monitoring can be quite different from cardiac ICU and general ICU data, where laboratory values and vital signs are collected in a similar manner. Furthermore, the length of stay in these settings can be considerably different, causing the data collection patterns to be different in these patient populations also. Because of the differences in the data, the opportunities for an objective comparison between the found models can be limited.

However, to understand the current landscape of AI and dynamic predictive modelling in medicine in general, numerous systematic reviews have been undertaken to look at different aspects of these models, as further explained in Section 2.1.1. At the time of writing, this chapter, however, is the only known review concerning dynamic prediction models predicting patient outcomes in critical care and ICU.

2.5. Conclusion

This review analysed published papers that predicted patient outcomes in ICU in a dynamic manner. The real-time prediction models are able to warn the clinicians earlier, and therefore provide the ward staff with sufficient time to intervene in order to prevent clinical deterioration [159]. The found studies show that there is a strong interest in developing dynamic prediction models for various ICU patient outcomes, however, the models developed so far have limitations. Based on these limitations identified in this chapter, the following steps will be taken in this thesis:

1. With many studies predicting mortality, more focus should be directed towards the prediction of complications. Hence, in this thesis acute kidney injury and delirium will be predicted both preoperatively and on an hourly basis in the ICU.
2. Predictive modelling methods that have shown to be robust when dealing with missing data should be further explored. In this thesis, in addition to other widely used predictive modelling methods, two methods that are shown to be appropriate to use with missing data - C5.0 and BARTm - will be experimented with.
3. More emphasis should be placed on testing the models in local databases that are appropriate for the patient population that is the potential demographic

where the prediction model is used. The analysis in this thesis will be carried out in the largest Scottish cardiac centre Golden Jubilee National Hospital.

Chapter 3. Exploratory Interviews with Cardiac Surgeons and Anaesthetists

3.1. Introduction

Due to healthcare being a data rich domain, and the evolving technology, various clinical decision support tools have been developed to help in managing a patient's journey through treatments. Data-driven clinical decision support tools are now an integral part of hospitals, especially in cardiac surgery, where various risk stratification tools have been developed to be used in perioperative assessment, focusing commonly on mortality [14], [16]. However, the currently used preoperative risk prediction models are widely criticised due to overestimating the risk of postoperative mortality [160] and for not being updated to fit the current cardiac care due to the changing cardiac patient population and evolving surgical procedures [161], [162].

While there are numerous risk prediction models for various outcomes in cardiac surgery, as shown in Chapter 2, the main risk prediction models used in UK cardiac centres are reported to be logistic European System for Cardiac Operative Risk Evaluation (EuroSCORE) [6] and Acute Physiology and Chronic Health Evaluation (APACHE) II [163], both of which were developed to predict mortality [164], [165].

As per Medical Research Council, involving stakeholders into the development phase when developing a new intervention is a crucial step in building a meaningful and usable clinical system [166]. As in thesis, a prediction model predicting postoperative complications that can be put into practice for cardiac surgery is developed, further understanding was required about the current challenges in cardiac surgery, clinicians' current views on and expectation for risk prediction models, and their opinion about postoperative complications in cardiac patients.

Semi-structured interviews were conducted with cardiac surgeons (N=3) and cardiac anaesthetists (N=9) in three Scottish cardiac centres: Golden Jubilee National Hospital, Royal Infirmary Edinburgh, and Aberdeen Royal Infirmary.

This study was an exploratory qualitative scoping study that was not intended to be in-depth qualitative work, however, was essential to set the scene and rationale for the studies presented in this thesis. In order to develop usable clinical prediction models, this study aimed to understand the following:

- What are the current challenges in cardiac surgery?
- What are the current processes to avoid adverse outcomes in cardiac surgery?
- What are the clinicians' priorities for newly developed clinical risk prediction models?

Through previous work in MPhil (see Chapter 1, Section 1.2) there was some understanding of the commonly used preoperative cardiac risk prediction models and complications in cardiac surgery population. Hence, a deductive approach was used for this scoping phase, following the Ritchie and Spencer framework [167].

Since the main contribution of this thesis is the development of models to predict postoperative complications, a top-level exploratory thematic analysis was used to understand the clinicians' views on clinical prediction models and the problem of postoperative complications in cardiac surgery.

3.2. Methods

This study was approved by the Department of Computer and Information Sciences Ethics Committee at the University of Strathclyde (ID: 837).

The study was conducted through N=11 semi-structured interviews with cardiac surgeons, cardiac anaesthetists, and cardiac intensivists. Open-ended questions were used to encourage exploratory and reflective discourse, placing an emphasis on the participant's perspective (Appendix 3.1). This also allowed clinicians to discuss what they themselves considered important and noteworthy in relation to challenges in cardiac surgery, adverse postoperative outcomes and perioperative risk prediction models. The questions were collated to reflect these aspects based on the knowledge that all UK cardiac institutions use logistic EuroSCORE and APACHE scores to predict mortality preoperatively and in the ICU [6], [163]. However, the problem of complications is rarely discussed in official auditing reports [6], which makes it difficult to understand the size of the problem. Furthermore, there is a lack of qualitative work including cardiac surgeons and anaesthetists to understand their perspective regarding postoperative outcomes and also how they use currently available risk scoring systems. The questions were also discussed with the clinical supervisor Prof Stefan Schraag who is a consultant cardiac anaesthetist at the Golden Jubilee National Hospital.

Based on the questions stated in Section 3.1, *a priori* themes were created, as per the framework approach developed by Ritchie and Spencer [167]. These *a priori* themes were *challenges in cardiac surgery*, *current processes to avoid adverse outcomes in cardiac surgery*, and *clinicians' priorities for new clinical risk prediction models*.

Since the PhD project was in collaboration with Golden Jubilee National Hospital (GJNH), most of the interviewees were from GJNH, which is the largest cardiac centre in Scotland. To understand the bigger picture of cardiac surgery in Scotland, participants from Royal Infirmary Edinburgh (RIE) and Aberdeen Royal Infirmary (ARI) were also interviewed. Invitations for interviews were sent to 64 potential participants, 28 of them were cardiac surgeons (18 from GJNH, 6 from RIE, and 3 from ARI), and 37 were cardiac anaesthetists (16 from GJNH, 13 from RIE, and 8 from ARI). In total, invitations were sent three times.

Table 3.1. Themes collated through top-level thematic analysis of the interviews.

Main Theme	**Sub-Theme**
Challenges in cardiac surgery	Postoperative complications and changes in cardiac surgery
	Communication between clinical professionals and patients
	Data collection on adverse postoperative outcomes and audit
Current processes to avoid adverse outcomes in cardiac surgery	Perioperative management processes
	Clinicians' perceptions on clinical risk prediction tools
Clinicians' priorities for new clinical risk prediction models	

Following the interviews, as recommended by Braun and Clarke [168], the interviews were transcribed by the author, then reviewed, and participants' quotes were highlighted and coded based on the topics of what was said. In addition, sections of two transcripts were coded separately by two other PhD students who have extensive experience conducting qualitative research (Dr Diane Morrow and Ramsay Meiklem) to ensure higher level of objectivity of the analysis. Discrepancies were discussed at supervisory meetings and resolved with the guidance of Dr Matt-Mouley Bouamrane. This process resulted in 13 codes, which were collated under the *a priori* themes. If codes were more appropriate for a theme that was not pre-selected, a new theme was created. The final themes were then refined and reported as main themes (*a priori* themes) and sub-themes[4], which were collated through the thematic analysis [168], shown in Table 3.1.

3.3. Findings

3.3.1. Participant Characteristics

Twelve participants took part in the interviews. Nine participants were from Golden Jubilee National Hospital, two participants were from Aberdeen Royal Infirmary, and one was from Royal Infirmary Edinburgh. Nine participants were cardiac anaesthetists (eight consultants) and three were cardiac surgeons (all consultants). The transcripts from eleven interviews were included in the analysis due to the corruption of the recording, resulting with nine participants from Golden Jubilee National Hospital and two participants from Aberdeen Royal Infirmary. The mean time of the interviews was 25 minutes and 34 seconds with standard deviation of 10 minutes and 30 seconds. The median time of interviews was 23 minutes and 2 seconds. The shortest interview lasted for 14 minutes and 4 seconds, and the longest interview lasted for 48 minutes and 30 seconds.

In terms of involvement in patient pathway from preoperative assessment to surgery itself and postoperative recovery, surgeons were always involved in preoperative assessment, and so were anaesthetists in most cases. In general, surgeons were more

[4] The codes with their respective sub-themes and main themes can be found from DOI: 10.15129/fe89d27b-fa1e-4698-865e-5e4a2697b8ee.

involved in preoperative assessment than anaesthetists, however, anaesthetists were more involved in postoperative care, especially in the intensive care unit (ICU).

When asked to describe their normal work week, a consultant cardiac anaesthetist said the following:

> *"Well, depends on exactly what I'm doing, if I'm working in intensive care, then we generally look after the patients after they've undergone cardiac or thoracic surgery. And that's either patients who are relatively straightforward and go through sort of a fast-stream ICU - we have very little input to those patients. Generally, the nursing staff will look after them, but on the other side in the long-term ICU we have the patients, who obviously develop complications and have a longer stay in intensive care. They require a lot more input, and that's really where I spend more of my time, perhaps." (Participant 3)*

A consultant cardiac surgeon said the following:

> *"So, it's more normal week, so I have two days operating a week and one day for cardiac surgery, which is mostly aortic valve replacements, coronary bypass grafting, occasional mitral type procedures and there's another day operating in thoracic surgery, which is predominantly lung dissections, etc. I'm in outpatient clinic once a week, I have two MDTs [multi-disciplinary team meetings] a week, which can have outpatient clinics attached to them as well." (Participant 5)*

3.3.2. Current Challenges in Cardiac Surgery

3.3.2.1. Postoperative Complications, Changes in Patient Population and Procedures

There were various challenges mentioned by the interviewees, the most common ones included adverse outcomes in cardiac surgery, changing patient population, and changing procedures in cardiac surgery.

> *"We see a lot of patients who survive, but have a very rough road, lots of complications, long hospital stays and fully expect not to have their presurgical function." (Participant 1)*

There were various adverse outcomes referred to by participants, the most frequently mentioned ones being atrial fibrillation, bleeding, delirium, infections, low blood pressure, multiorgan failure, renal complications, respiratory complications, sepsis and stroke. It was evident that some complications especially, such as stroke and delirium, have become more prevalent based on the experts' opinion.

> *"And probably the biggest one we now see is delirium. You know, delirium is a massive problem in cardiac intensive care. Taking an elderly population, giving them lots of medication, taking them outside their normal routine and you know, polypharmacy, put all of those things into a melting pot and then their brain is scrambled at the end. That's [delirium] a really big problem, that's the single biggest change I've probably seen in the last five years, is the increase of delirium." (Participant 3)*

Some complications, such as atrial fibrillation, can be considerably straightforward.

> *"Cardiac problems such as atrial fibrillation, about 40% of the patients have it, and that's just a bit of a standard risk. And I've been involved with some work trying to minimise atrial fibrillation, there's a whole number of interventions that you can do, some of which are minor differences, but most of it is just, half the patients get it, half the patients won't. As you get it, you treat it appropriately." (Participant 11)*

> *"You know, the easy one is atrial fibrillation, but I don't really see it as a complication. It happens to a quarter to a third of people. So it's too common to be a complication as such." (Participant 2)*

However, some complications can have a lasting impact on a patient.

> *"But if you hit the bump or they develop complications or something goes wrong at some point along the process then suddenly being in their older years, the comorbidities catch up on them. And it becomes a much bigger problem then, they don't bounce back in the way that they would do." (Participant 1)*

> *"Even minor or moderate complication can have an impact. But it's more the severe complications that are most definitely making a mark on pretty much every patient, regardless of who they are." (Participant 4)*

One challenge that often drives the prevalence of adverse outcomes, according to the interviewees' opinion is the aging population.

> *"Because surgical complications, once you're elderly, your tissues are frailer and the big blood vessels often have calcium in them, tissue tears more easily. You can end up with surgical complications that don't happen when you're younger, dealing with younger, fitter patients. And there's also calcification of arteries, changes that can lead*

> *to strokes, gut, infarctions, kidney problems, lots of different complications" (Participant 9)*

> *"Perhaps, the patient population has gotten older, they have more comorbidity, more medical problems in the background." (Participant 3)*

In addition to aging, patients who undergo open-heart surgery also have more comorbidities, which is a challenge for cardiac surgeons and anaesthetists.

> *"First there are clinical challenges that I deal with – patients that have variable and sometimes unpredictable level of morbidity." (Participant 4)*

> *"Most complications happen not because anybody has done anything wrong, but it's the nature of doing surgery in patients that are 80 years old and have lots of vascular disease, not only in their coronaries, but also clotted arteries supplying their brain and all that kind of stuff, so... Quarter of them are diabetics, so that increases the risk of having sternal wound infection, etc." (Participant 5)*

In addition to the changes in population, also the types of procedures patients receive have changed. Namely, as opposed to having coronary artery bypass graft, patients are having more valve surgeries.

> *"I've been a cardiac anaesthetist for about over 20 years, so when I started, the patients who I was seeing were primarily requiring coronary artery surgery with a few valve patients. So, what has happened with cardiac surgeries is that you need to take on aging population and slightly different population. So, when I started 20 odd years ago, it was very rare to see a patient in their 80's presenting for cardiac surgery. And now the majority of patients we have are coming for valve surgery and many of whom are in their 80's. So, the population that we have for surgery, an awful lot of valve patients and an awful lot of elderly valve patients" (Participant 11)*

In addition to more valve surgeries, patients with more straightforward cases tend to have minimally invasive surgeries, such as transcatheter aortic valve implantation (TAVI) and percutaneous coronary intervention (PCI), meaning open-heart surgeries are now common for more complex patients.

> *"More and more risky patients for me. People expect to live longer. 10 years back we weren't even looking at those patients, we weren't even considering them. But the main thing is the changing patients, so getting worse patients, who we previously declined and now are offering surgery, or they are coming for TAVI." (Participant 7)*

> *"Now, because of the changing techniques in cardiology, particularly PCI and the cardiologists taking on more complex PCI cases is that what before was our core population group, they all go now to cardiology." (Participant 11)*

It is common that these three challenges are linked to one another, as explained by Participant 3:

> *"My impression is that, perhaps the instance of postop stroke has gone up slightly, but again that can be partially explained on the ageing population. You know, patients that are over 75 have an increased risk of stroke, patients with coronary artery disease for more than 3 years have an increased risk of stroke. Patients who are having valves have an increased risk of stroke. And that's the population that we are usually dealing with now, so that also includes kidney problems. I suspect the kidney problems have fallen a bit, you know, we seem to be, we see patients taking a little renal hit, but I suspect... the thing is we're a bit more aggressive on sending patients on filter than we used to be. We used to sit and wait for a few days, whereas now when we suspect, we put them under filter a tad quicker, so we aggressively manage them going into renal failure rather than waiting for them to establish renal failure." (Participant 3)*

And also agreed by Participant 11:

> *"I think stroke has probably gone up, but that probably reflects different patient groups, different operations, different technology." (Participant 11)*

Due to the changes in cardiac population within the past decade, and the complexity of open-heart surgery patient cases, mentioned by the experts, complications can be a significant challenge in cardiac surgery.

> *"Very few of them [patients] have just one single complication." (Participant 1)*

> *"Sometimes I think unrealistically, and we are often operating people who are in their eighties who have little realistic chance of getting back to full functional lives." (Participant 9)*

> *"Patients come through that initial operation, but they won't get out of the system. We get them extubated on day one. Then they have an atrial fibrillation, then they go into cardiac failure, then they get reintubated. then they get a chest infection, then they get tracheostomy, then they go into gradual decline, and you know, two-three months later we meet them back at the ward, then they've got poor physiological reserve, they end up developing respiratory failure. At the end of the day, you know, 3-4 months down the line they just die in slow undignified death, because basically, although we fixed the*

> *mechanical problem, we never fixed them physiologically, so we never get out of that system." (Participant 11)*

3.3.2.2. Communication Between the Clinical Professionals and Patients

Other challenges also included communication both with other professionals in cardiac care and with patients and their families. When it comes to communication of risk and understanding of the seriousness of open-heart surgery, there appears to be a disconnect between the patient and the clinician.

> *"I think the expectations of the healthcare have increased enormously over the last decade even. The patients expect to get more treatment or we're able to fix most problems and sometimes I think there's a mismatch in terms of the expectations and what we can deliver." (Participant 3)*

The interviewed clinicians seem to have a same view that being able to communicate risk to patients and their families can be challenging.

> *"It's not just about medical side of things and I think if you consider that, cause you're not just treating patients, it's also about being their psychologist. And there's a journey that you have to take to help them to make sure they are in the same place as me." (Participant 7)*

In addition to the communication between the clinicians and patients, there are also challenges of communication and understanding between the clinicians working directly with patients (such as surgeons and anaesthetists) and hospital management.

> *"The type of procedures we do in cardiac surgery in particular have become more complex over the last 10-20 years, but theatre allocation hasn't adapted to that new need." (Participant 5)*

It was agreed that more communication is needed to ensure the common aim to provide excellent patient care.

> *"I guess you just need mutual respect between teams and there needs to be a buy-in in all areas about what the calls are of care." (Participant 6)*

3.3.2.3. Data Collection on Adverse Postoperative Outcomes and Audit

Even though it was evident from the interviews with the cardiac surgeons and cardiac anaesthetists that complications following cardiac surgery are a challenge, there is also a lack of cooperation between data collection, evidence-based medicine, and auditing. Because the emphasis is mostly on mortality, the clinicians have doubts about how well complications are recorded. Because the data about complications can be of low quality, it is difficult to know how often certain complications actually occur in the patient records.

> *"I doubt it's all robustly recorded at the moment. We just don't know what the incidence of everything is." (Participant 1)*

One important barrier to data collection that came out from the conversations with the clinicians was staffing. Regardless of whether the staff member is a surgeon, anaesthetist, nurse or a medical student, their priority is patient care, as opposed to data entry. This can cause some variability in the quality of the data entered the electronic health records.

> *"I put data in for our patients. It's relatively accurate, but I wouldn't say it's with most comorbidities. We just leave out. Because we can't be bothered, and we don't think it makes much difference. But I'm not a data inputter, I'm a consultant and I take my time to do it for all of my patients, but even that as accurate of a consultant I am and I understand it." (Participant 6)*

As currently the auditing of cardiac surgery is based on postoperative mortality, and as long as the auditing processes do not include complications, the data entered about complications in medical systems can remain low in quality.

> *"Mortality is what everybody benchmarks and that's what goes forward to the government. Having more time or meeting more regularly and better data systems - these are the main barriers at the moment [to discuss morbidity or complications]." (Participant 3)*

The interviewees were also asked about the mortality and morbidity meetings and the discussion about adverse postoperative events there. Mortality and Morbidity (M&M) meetings are an organised discussion, where adverse events from surgery are investigated and presented. These were put into practice with an aim to improve patient outcomes, quality of care, patient safety and education of clinical staff [169].

All participants agreed that most of the time, only mortality is discussed, and morbidity is discussed rarely. This was mainly expressed to be due to lack of time, and also due to aforementioned data collection barriers. If the incidence of complications was known, the seriousness of the issues would be clearer to the M&M meeting attendees.

> *"We don't talk about morbidity at all, do we? There's no time. And you struggle to get the mortality presented, let alone morbidity. Some of the mortalities... it's difficult to know if there's any learning from it, but I think it's good that they all get presented whether there's learning or not. It's just a level, minimum requirement, which I think is healthy." (Participant 6)*

Because the number of cases for mortality is considerably low, it can be difficult to know for certain what went wrong, which brings out the flaws of the M&M meeting mortality discussion as well.

> *"M&Ms are more the sort of clinical dissection of what happened to that patient and can we learn from that. The problem with these is, it's good practice to do so, but it's sentinel event, it's quite often not possible to draw a general conclusion to your practice, unless it's supported by general evidence. And that's mostly not the case." (Participant 4)*

3.3.3. Current Processes to Avoid Adverse Outcomes in Cardiac Surgery

There are various ways that cardiac surgeons and cardiac anaesthetists are managing patient risk to avoid adverse postoperative outcomes. In general, however, risk is inevitable.

> *"The only way you can limit surgical risk is to not operate, which is not always possible." (Participant 5)*

Hence, various steps have been created to minimise the risk of severe adverse outcomes, including death after surgery. These steps include having multi-disciplinary team (MDT) meetings to decide the appropriate treatment plan for the patient. The preoperative clinics help to decide what the risk of various adverse outcomes for the patient is. The patient flow at the Golden Jubilee National Hospital from one stage to another can be seen in Chapter 5.

3.3.3.1. Perioperative Management Processes

The ultimate way to minimise risk is patient selection, which is discussed in multi-disciplinary teams, however the final decision is made by the surgeon.

> *"The traditional model is that we see the patients preoperatively and that is usually, you know, the night before the surgery. The clinic means that they're seen in advance of this and helps to try and, you know, spot problems, but if I wasn't happy with the patient the night before or the day before then I still would cancel surgery. So, I think it's getting the balance of which patients should be operated on versus who shouldn't." (Participant 9)*

Even though it was clear from the interviews that the anaesthetists are less involved with preoperative clinics, unless further discussion was needed for deciding whether the patient is suitable for surgery, all anaesthetists in this study found it necessary to see the patient at least the night before surgery.

> *"I'm still, even if they've been seen in the clinic by somebody else, I still want to see them myself with my own eyes and speak to them myself before surgery. I find it very difficult to miss that step out." (Participant 9)*

A few anaesthetists were also as part of the preoperative team, where they assessed patient's readiness and appropriateness for the surgery.

> *"So often we see patients where we think it's not appropriate that they go for surgery and they should really go for some other type of intervention, whether that's in cardiology or they would have no surgery at all, because their risk is too big. I think that's the way to decrease complications in ICU, and that is to be more selective about the patients that we take for surgery. So, we look at some known factors, we look at the diabetic control, whether they're anaemic, you know, their mobility, muscle mass, and also just put that all together and sort of how feel about the patient. There isn't really a number as such we can put on it just yet." (Participant 3)*

Another important part of deciding appropriate treatment plan is shared decision-making. Shared decision-making is a process in which a healthcare professional and a patient discuss treatment options to decide the best care option for them together. This involves discussing possible benefits and risks of various treatments and making sure the patient understands the consequences of different options [170].

> *"We had a discussion with an anaesthetist to discuss whether one patient would be appropriate to have a double valve replacement, considering the comorbidities and her age. We sit down with the patient in that situation to see what she wants. So, we make collaborative decisions too. Doing these things is tough and making these decisions... There is no right or wrong at times. [...] you don't get any prizes for getting last week's lottery number, you know. Sometimes our decisions prove to be wrong, but not generally we don't make a grossly wrong decisions and if we have doubt, we always take somebody's - colleague's opinion to make that decision if we have the time." (Participant 7)*

Conversely to the preoperative stage where discussions about risk are had and various treatment options are weighed, during surgery the number of possibilities for minimising risk are low.

> *"You can modify your operative approach and things to try to deal with this, try to minimise this, and you'll never get a 0% mortality or morbidity. There always has to be a finite number and the lower it gets the less predictable it gets. Well, you might pay more attention to things like blood pressure or bypass and things like that. Which you do anyway, but you get into the habit of expecting a confusionist to know what they're doing. A lot of the damage has been done before they get to intensive care. So, there's a concept called ICU proofing that surgeons employ. So, things that you have to make sure is that they're not bleeding, that they're warmed up properly and you know, the cardiac support drugs are at minimum. And if you do those three things then variably, they'll be in and out of there in no time." (Participant 2)*

Overall, the majority of the risk reduction and decision-making is done at the preoperative stage when there is a possibility to carry out various tests and have discussions with the multi-disciplinary team, the patients and their families.

> *"So sometimes things don't go as planned. You don't walk into work thinking you're going to kill somebody. We treat people with our best intensions and intension to treat. To get a better outcome and we do have to make certain decisions, sort of prognostics, however accepting life is sacred, we think if somebody has complete brain injury, we discuss with the family, that this is not compatible with life and so sometimes we're lucky if the patient just ends up passing away or if they are brain dead too and the family is willing we let them make the decision, which can be quite significant part of the job." (Participant 7)*

3.3.3.2. Clinicians’ Perceptions on Clinical Risk Prediction Tools

It was evident that all participants had used some clinical risk prediction tools in one way or another, however, the routine usage was lacking. All interviewees mentioned logistic EuroSCORE (European System for Cardiac Operative Risk Evaluation), which is a preoperative risk prediction model for postoperative mortality [171]. It was also common, that the logistic EuroSCORE was used more by surgeons than anaesthetists. The interviewees agreed that there are various risk prediction models in cardiac surgery, however, these have their limitations due to which the tools were not often used.

> *“Things like EuroSCORE... They're useful to some degree. One of the things is that probably most people are... They do two things; they allow you to risk stratify the patient and it probably gives you some idea on how you're performing or the unit performing or the individual surgeon.” (Participant 11)*

Interestingly, there were various approaches in how the logistic EuroSCORE was used. A number of interviewees used logistic EuroSCORE to both understand the risk of patient outcome and to explain that risk to the patient. However, for some, these risk prediction models should be strictly for the clinician’s usage.

> *“As far as the patient is concerned, when we speak to them, we give them pretty well the same numbers. So could be a complicated operation, but the mortality won't be that much higher than an uncomplicated, straightforward operation. So I won't get my phone out and tell them if they are going to die or not.” (Participant 2)*

When asked about why the clinical risk prediction models are used so little, the interviewees listed limitations, including only involving preoperative data, not making personalised prediction, not giving guidance on what to do with the result of the prediction, and the models being outdated.

> *“One of the problems with the EuroSCORE, for example, is that it seems outdated. EuroSCORE gives you a rough percentage of a chance of mortality, but most units, a number of years now, if not decades have outperformed the EuroSCORE. So, if the risk of death on EuroSCORE is 8%, people would find... which is a high EuroSCORE, relatively, people would have mortality rates of 2%. So, it's almost fallen behind. That's one reason why people don't look at it quite so much.” (Participant 9)*

However, the main usage of the logistic EuroSCORE was told to be for documentation to support certain decisions.

> *"Risk scores are useful, at least you have something to go by. If it comes in defending yourself in doubts of malpractice, I think these scores have a place if they are well validated. There are lots of criteria in medicine which are useful like CHILD's criteria, APACHE score, Thoracoscore. If you want to decline somebody for surgery, it makes our job very easy. But then you offer somebody a surgery with very high Thoracoscore, then people can go back with the rationale that it isn't very much validated. So, people use it to their own advantage in times. What we have to do, when we're making our decision, our thinking process has to be explained and put into black and white on the paper. When 10 years later if there's an investigation of, or even within two weeks' time there's a mortality incident meeting, that process is explained, and I think that is the problem." (Participant 7)*

3.3.4. Clinicians' Priorities for Improving Clinical Risk Prediction Models

The interviewed clinicians thought it would be beneficial for a risk prediction model to be more personalised to a patient, as opposed to using a population-level prediction, like EuroSCORE models do.

In terms of outcome predicted, one criticism towards currently available risk prediction models was that they mostly predict mortality.

> *"I think a lot of emphasis is often placed on mortality, which I think is in a wider picture obviously devastating for the patient and their family, but actually the morbidity is what happens more frequently and has a greater long-term impact on patients, their family and the wider care and cost of care for the whole community. So, I think being able to, but again I think that's even more difficult, is to be able to predict morbidity as opposed to mortality. And I think that's why there aren't many scoring systems around that can do it." (Participant 8)*

Various options were suggested to be predicted, including intensive care unit stay, delirium, stroke, bleeding, infections, respiratory complications, and renal complications. However, it was mentioned that before predicting these complications, the data collection on the incidence of these complications should be improved. In addition, being able to predict a combination of complications, as opposed to individual complications was suggested.

> *“A lot of the time, you know we can look at patients and if we do the operation, we get them in the HDU without complications and just push them through the system, it should be OK. But you know if you have other problems, it gets into a cycle, of you know, a bit of chest infection, bit of failure, bit of reintubation. Then it starts spiralling downwards. And that's probably the one area that you could almost argue you need a sort of a better prognostic tool.” (Participant 11)*

Above all, validation of the prediction models appeared to be important for all interviewees.

> *“Depends on who you speak to, we have different perceptions of the validation quality. For example, even the EuroSCORE, although it's been used everywhere, some people believe it's not quite as validated in our population, or as the population has changed the validation studies are not applicable anymore to a degree they used to.” (Participant 4)*

3.4. Discussion

In this chapter, clinicians were interviewed to understand the rationale behind developing a clinical prediction model predicting postoperative complications. This was done to understand the current context of which clinical prediction models are used in, and what requirements the clinicals have regarding a prediction model predicting postoperative complications in cardiac surgery. As per Medical Research Council guidance, involving stakeholders in the development phase of a clinical system is a necessary step to develop a usable, fit-for-purpose and implementable prediction model [166]. Eleven interviews with cardiac anaesthetists and cardiac surgeons in Scottish cardiac centres were analysed, using top-level exploratory thematic analysis.

While clinicians working in cardiac patient pathway have been involved in some projects to do with digital innovation in cardiac surgery [172], [173], studies involving potential users of prediction models for outcomes in cardiac surgery are rare. There are hundreds of published papers (as shown in Chapter 2) about developing prediction models for postoperative outcomes, however, the evidence of the requirements of potential users of these models is lacking. The study presented in this chapter takes the first step to involve stakeholders' opinion in the model development.

3.4.1. Current Challenges in Cardiac Surgery

The main challenges in cardiac surgery currently were found to be adverse outcomes, changing patient population and changing procedures, all of which are interconnected.

Regarding adverse outcomes, various complications were mentioned to be an issue that can make patients stay in the ICU for longer and be dependent on the healthcare services for a very long time. The commonly mentioned outcomes included bleeding, infections, renal complications, delirium, respiratory complications, and stroke. These adverse outcomes were explained to be connected to the changing population undergoing open-heart surgery. One commonly mentioned change was aging - older patients who were previously rejected surgery have now open-heart surgery more often. This has also been shown by the Society for Cardiothoracic Surgery reports: more than 40% of open heart surgery patients are older than 70 years, and the proportion of patients aged older than 80 years has increased from 4% to 11% since 2008 [6]. Higher age has been shown to be associated with various adverse postoperative outcomes, including mortality [174], and complications, such as bleeding [175], infections [176], pulmonary complications [177], renal complications [178], and delirium [179]. Age itself is usually not considered to be the cause of these outcomes, however it can be a marker of other risk factors, such as frailty, hypertension, diabetes, and increased atherosclerotic burden, which is associated with stroke [180], [181]. Atherosclerosis is thickening or hardening of the arteries, which is caused by the build-up of plaque in the inner lining of an artery.

Another change in population was explained to be patients who have more pre-existing conditions, which can add to the risk of developing adverse postoperative outcomes. In the literature it has been shown that patients with preoperatively undetected renal problems can increase risk of chronic kidney disease after cardiac surgery [182]. In addition, frailty, being common in cardiac patients, can add to the risk of mortality and postoperative complications [183]. In addition, the number of insulin-dependent patients is increasing [184], which is a risk factor for postoperative complications after cardiac surgery [185].

In terms of changing procedures, it was found that more straight-forward cases undergo minimally invasive surgeries, and patients with higher risk profile tend to go for open-heart surgery, such as CABG, valve and combined CABG and valve procedures. Furthermore, within the past decade, more patients tended to undergo valve procedures, which can be more complicated than CABG. When CABG surgery intends to restore circulation to the coronary arteries, which are on the heart surface, valve surgeries intend to replace valves, which are located inside the chambers of the heart. Valve surgery patients tend to have more advanced heart disease, making valve surgeries more complicated, which can be associated with higher risk of mortality and morbidity for valve patients than CABG patients [186].

3.4.2. Processes to Avoid Adverse Outcomes in Cardiac Surgery

3.4.2.1. General Processes

It was evident from the interviews that most of the decision-making and risk management happens in preoperative stages, where the appropriate treatment with the multi-disciplinary team is chosen for the patient. In addition, in preoperative clinics, where different laboratory tests can be taken, the risk of mortality and complications is assessed. Preoperative examination has shown to reduce adverse outcomes following cardiac surgery by helping to identify pre-existing conditions and help mitigate possible postoperative risks of unwanted outcomes [187].

Patient selection was the ultimate way to reduce patient's risk. If a patient was eligible for a minimally invasive surgery, that option was preferred. However, shared decision-making processes were used, where patient and their family were also involved. Shared decision-making together with a patient and a multi-disciplinary team has been shown to help provide care that is more consistent with patient’s expectations [188].

Interestingly, even with the establishment of preoperative clinics and perioperative pathway [189], the surgeons ultimately make the decision whether the patient is eligible for surgery or not. This is because historically surgeons are held accountable in case of adverse surgical outcomes, especially in case of mortality, with the report of the Bristol Royal Infirmary Inquiry in 2001 [190]. This report prompted The Society for Cardiothoracic Surgery to collect data on surgical outcomes and the responsible surgeons [6]. This had a positive effect on being able to audit individual cardiac centres in terms of quality [6], however, it placed full responsibility on the surgeon, not on the full multi-disciplinary team that have an important role in decision-making and caring for the patient. That being said, this is an illustration, where a policy change can improve data collection on certain hospital outcomes, meaning this could also be done with postoperative complications.

3.4.2.2. Risk Prediction Tool Usage

In general, the participants have used some risk prediction tools, however, the main usage for these was for documentation purposes to justify the decisions made. The commonly mentioned risk score was logistic EuroSCORE, which is used in Golden Jubilee National Hospital for auditing purposes and is reported on a national scale to The Society for Cardiothoracic Surgery in the UK [6].

The main reason why risk scores, such as EuroSCORE was not used for regular decision-making was the perception of EuroSCORE not being validated enough. The first version of EuroSCORE was published in 1999, and was developed in 132 cardiac centres in 8 countries, using 20,014 patient records [171]. The original EuroSCORE has been validated in various studies [191]. That being said, because of the changing patient population in cardiac surgery, there is criticism that the logistic EuroSCORE overestimates observed mortality [192]. This is also shown in The Society for Cardiothoracic Surgery report, where the average mortality predicted by logistic EuroSCORE was reported to increase from 5.6% in 2008 to 8.5% in 2016, however the crude mean in-hospital rate was reported to have fallen from 4.0% in 2008 to 2.8% in 2016 [6]. Interestingly, the newer version of EuroSCORE (EuroSCORE II) [193], published more than a decade ago, still is not used to audit mortality in UK cardiac centres [6].

3.4.3. Priorities to Improve Risk Prediction Models

The main suggestion for improving clinical risk prediction models was that the predicted outcome should be different from mortality, and the focus should be more on outcomes that have a potentially long-term negative effect on patients, such as complications and ICU stay. As expected, the complications suggested were same as the ones explained to be current challenges, which were most commonly bleeding, delirium, infections, renal and respiratory complications, and stroke. Currently developed perioperative risk stratification models mostly predict mortality or specific complications [14], [16]. However, the interviewees expressed the need for a model that captures several complications at the same time. Currently existing clinical prediction models for both preoperative and postoperative prediction will be further discussed in Chapters 4, 6, 7 and 8.

Predicting other outcomes than mortality has some barriers, that were also identified by the interviewed clinicians. The main problem the clinicians agreed about was the fact that due to historical focus on mortality in auditing, complications are not very well recorded in clinical databases, which is part of the problem of data quality in electronic health records in general [194]. A great barrier to being able to predict complications is that it is unknown what the incidence of the outcomes is, which makes knowing the magnitude of the problem difficult. Currently, at the mortality and morbidity (M&M) meetings, according to the participants, morbidity is almost never discussed due to lack of time, but also not prioritising complications. In order to improve the data quality regarding postoperative complications, efforts on auditing other outcomes than mortality, and investment of time and resources should be directed towards data quality improvement [195].

However, beyond the requirements of predicted outcomes, the main facilitator for clinicians to actually use a tool that would predict adverse outcomes is appropriate and thorough validation of such tool. A number of steps are required for a clinical prediction model to be validated and to be ready for use in practice. Firstly, the model should be developed, using the TRIPOD guidelines [121] to follow reporting requirements. Secondly, qualitative work is required to understand the acceptability and applicability of the prediction model [196]. Once the prediction model is shown to improve patient outcomes, its cost-effectiveness should be evaluated. Finally, implementation and dissemination strategy should be identified to help put the prediction model into practice [197].

3.4.4. Limitations and Areas for Further Research

As this study was a top-level exploratory analysis to understand the general view of cardiac anaesthetists and surgeons regarding the challenges in cardiac surgery, how adverse outcomes are avoided and the priorities for new clinical prediction models are, the study has a number of limitations and provides a number of avenues for future work.

In terms of participants in this study, even though invitations were sent to all three cardiac centres in Scotland, majority of the participants were from the Golden Jubilee National Hospital, which reflects the opinions and processes of cardiac patient care for that hospital mainly. The reason why most participants were from GJNH might be because of the project being affiliated with this hospital, and the author's regular presence and local connections. Even though the majority of the cardiac surgeries in Scotland are held in GJNH, which should reflect the cardiac surgical processes in Scotland as a whole, the information gathered does not take into account the variance in quality and processes that other two cardiac centres might offer. Hence, to understand the challenges, care processes and needs for clinical prediction tools for Royal Infirmary Edinburgh and Aberdeen Royal Infirmary, further research is needed. However, it can be assumed that the clinicians taking part of this study have worked in various other cardiac centres around Scotland and the UK in general – hence the participating clinicians' opinions are likely to be based on their experiences throughout their careers in different cardiac centres.

In terms of study results regarding the priorities for improving current clinical prediction models, this was not an in-depth analysis of system requirements for a clinical prediction model that could be put into practice. Because the nature of the responses was speculative, without showing any prototype of such clinical prediction models, a usability study should be undertaken where the prediction models can be presented to the clinicians to gather feedback on the developed models. Furthermore, to aid seamless integration into practice, a contextual inquiry should be undertaken to understand how clinicians interact with clinical prediction models and to understand how such tools could fit into daily practice.

3.5. Conclusions

This chapter presented preparatory work that shows the need for a clinical prediction model for postoperative complications that could potentially help to mitigate the identified current challenges in cardiac surgery. As the study presented in this chapter was not an in-depth analysis of clinicians' requirements for a clinical prediction model, further research is needed to understand the particular requirements for such model, and to understand how the model could fit into practical context. While there are numerous prediction models in cardiac surgery, the involvement of potential users of such models at the development process is rare. Hence, this study provides a contribution by presenting the cardiac surgeons' and anaesthetists' priorities fora clinical prediction model and highlights the need for a model predicting postoperative complications, as opposed to mortality. The findings, regarding which complications to predict, were used to decide upon which exact complications are predicted in this thesis.

Chapter 4. Definition and Classification of Postoperative Complications Following Cardiac Surgery: A Delphi Study

4.1. Introduction

This chapter is investigating how experts working with cardiac surgery patients would define and classify complications following cardiac surgeries, such as coronary artery bypass graft (CABG), aortic valve, and combined CABG and valve surgeries.

Currently, the focus of outcome prediction in cardiac surgery is on mortality. However, with the number of cardiac surgery patients steadily increasing and more patients presenting myocardial infarction before CABG surgery, postoperative complications are becoming more prevalent [6]. According to the Society for Cardiothoracic Surgery (SCTS), *"[In the UK] the mortality rates across all [cardiac surgery] groups are some of the lowest in the world despite increasing age, risk profile and frailty of patients"* [6]. Complications after surgery, however, are common [17], and can have a debilitating impact on patients' quality of life [8], [9]. Depending on the severity of complications, they can also increase hospital length of stay [4], [139], and hence increase healthcare costs [10]. It is therefore essential that "*efforts would be directed to further reducing morbidity and length of stay"* [6] and that adequate systems are developed within clinical care in order to better plan and mitigate these severe complications. To achieve lower levels of morbidity and reduced length of stay following surgery, SCTS identifies the following required changes [6]:

- More data needs to be collected on morbidity and long-term mortality, as opposed to 30-day or in-hospital mortality.
- Long-term outcomes need to be monitored more.

- The quality of reporting by adding patient-reported outcomes, specifically developed for cardiac surgery, in the care pathway needs to be improved [6].

From Chapter 3 it was found that prediction models for postoperative complications are needed, however, at present, a major obstacle in analysing and predicting morbidity is the lack of agreed definition and classification of postoperative complications [198]. Due to this, when comparing different research studies within this field, all studies have a different definition for "morbidity", which includes a different set of combined complications [199]. The reporting of different complication outcomes in the scientific literature therefore prevents the objective comparison of the performance of currently developed risk models predicting morbidity.

This study aimed to address these issues by using the Delphi method [200] in order to answer the following questions:

1. What is cardiac surgery experts' opinion on the usefulness of a definition and classification of surgical complications following cardiac surgery?
2. How do cardiac surgery experts define what events constitute surgical complications following cardiac surgery?
3. How do cardiac surgery experts classify surgical complications following cardiac surgery?

4.2. Related Work and Rationale

4.2.1. The Clavien-Dindo Complications Classification System

The first and also the most well-known proposal for classification of postoperative complications in any surgery was published in 1992 by Clavien et al. [201]. They created general principles to classify surgical complications based on severity. Dindo et al. modified this system in 2004 [202], to include life-threatening complications requiring intensive care, and complications involving the central nervous system. To show the usefulness of the Clavien-Dindo system, the authors also validated the updated system in 6336 general surgery patients. The Clavien-Dindo system assumes a patient to have one complication only, however this is unrealistic. To solve that problem, Slankamenac et al. [203] created the Comprehensive Complication Index (CCI), allowing the addition of weights for each complication grade, creating the CCI score. The CCI score ranges from 0 to 100, where 0 is defined as "no complications" and 100 is arbitrarily defined as "death of the patient". Table 4.1 shows the grades and the definitions of each grade in the Clavien-Dindo classification system.

Table 4.1. Clavien-Dindo Classification of Surgical Complications [204].

Grade	**Definition**
Grade I	Any deviation from the normal postoperative course without the need for pharmacological treatment or surgical, endoscopic, and radiological interventions.
Grade II	Requiring pharmacological treatment with drugs other than such allowed for grade I complications.
Grade III	Requiring surgical, endoscopic, or radiological intervention.
Grade IIIa	Intervention not under general anaesthesia.
Grade IIIb	Intervention under general anaesthesia.
Grade IV	Life-threatening complication (including CNS complications) requiring ICU management.
Grade IVa	Single organ dysfunction.
Grade IVb	Multiorgan dysfunction
Grade V	Death of a patient

Different studies have shown benefits of using the Clavien-Dindo classification system in different types of surgeries [205]–[213], demonstrating the system's applicability in varying cohorts of patients, and its usefulness as a measurement of standards in the quality management for surgical departments. However, there is an argument to be made about why cardiac surgery needs its own classification system for complications, as discussed in the next section.

4.2.2. Why Cardiac Surgery Needs its Own Classification System for Complications

Historically, the quality of cardiac surgery has always been defined by mortality rates. The widely used risk prediction scores used for audit purposes, such as EuroSCORE [164] and Parsonnet score [214] have been designed to predict mortality. The focus on mortality was even further ingrained into the system following a report investigating deaths of a large number of children following heart surgery at the Bristol Royal Infirmary [215]. The report gave 198 recommendations to improve the care standards in the National Health Service (NHS), including the recommendation that *"clinical* *audit should be compulsory for all healthcare professionals providing clinical care"* and *"clinical audit must be fully supported"*. The report also recommended that *"the indicators of performance should be comprehensible to the public as well as to healthcare professionals"* and that *"they should be fewer and of high quality, rather than numerous but of questionable or variable quality"*. According to the report the SCTS was already collecting data on mortality, hence, this was a convenient and straight-forward way to measure a cardiac centre's performance and quality. On the other hand, morbidity and postoperative complications vary from patient to patient and amongst hospitals, meaning that without any agreed definition and classification, morbidities are ill-defined and difficult to understand to both expert and patient. However, depending on severity, a complication can have a big impact on the patient, their family and on the healthcare system, just like mortality [4], [8], [9].

Hébert et al. [198] were the first to validate the Clavien-Dindo complications classification system for cardiac surgery patients. Even though Clavien-Dindo system is associated with numbers of comorbidities, length of surgery, length of hospital stay, and procedure complexity, the grading system is not specifically developed for cardiac surgery, meaning there can be a lot of room for subjectivity [198].

There are many reasons why outcome measures should be differentiated between cardiac and general surgery. Unlike cardiac surgery, general surgery covers a wide range of subspecialties, including breast, colorectal, endocrine, upper gastrointestinal and transplant surgeries. General surgeries also include a large amount of minimally invasive procedures, such as laparoscopic (or "keyhole") surgeries that result with less pain for patients, better outcomes and shorter postoperative recovery [216]. In addition, cardiac and general surgery differ also in terms of mortality rates, patient population, and hospital length of stay, all of which will be discussed below.

Different Mortality Rates

The risk of mortality for general surgery and cardiac surgery are very different: in Scotland, the mortality rates for general surgery patients have been under 0.8% since the year 2000 [217]. For cardiac surgery, it is around 2.5% [6], which is considered low, however significantly higher than it is for general surgery, and therefore proposes a higher risk for patients.

Furthermore, mortality rates differ among various cardiac surgeries. Between 2015-2016 the overall mortality rate for CABG patients was 1.0%. For combined CABG and valve surgery it was 4.0%, and for redo CABG 7.7%. For valve surgeries, in-hospital mortality was 4.6%, and for isolated aortic valve replacement (AVR) surgeries it was 2.7%. Combined AVR and CABG surgeries on average have resulted in an in-hospital mortality rate of 5.4% [6]. These statistics are a clear example of the higher surgical risk that cardiac surgery presents, compared to general surgery.

Different Patient Population

A study by Grant et al. [218] looked at trends and outcomes from 2002 to 2016 for cardiac surgery in the UK and found that there has been an increase in patient risk profile. According to that study, there has been a significant increase in patients' mean age from 64.2 years in 2002 to 66.4 years in 2010, with the age staying the same until 2016. Elderly patients are at higher risk of postoperative complications, especially for bleeding, infections, neurologic, pulmonary complications and renal problems [219] due to age-related changes in cardiovascular physiology, such as changes in the vessel wall and the myocardium [220]. In addition, the higher the age, the more likely the patient is to have weaker tissues and frailty, causing a lower tolerance to surgery [221].

The majority of cardiac patients in the UK have been men, however the proportion of female patients has also steadily increased [218]. Women often lack chest pain [222] which can delay diagnosis and therefore could lead to worsened myocardial infarction [223]. More patients have pulmonary disease and active endocarditis before surgery than before, increasing the risk of postoperative mortality and complications [224]–[227]. Grant et al. also note that the incidence of cardiac surgery for active endocarditis has more than doubled from 2002 to 2016 [218].

There were fewer isolated CABG procedures taking place in 2016 than there were in 2002, the number of which reduced by a third. There has been a consistent increase in patients undergoing some form of valve surgery, more specifically, isolated valve surgery, mitral valve surgery and aortic valve replacement [218]. Valve surgeries in general propose higher risk of postoperative mortality and complications to a patient than CABG procedures [228].

Different Hospital Length of Stay and Postoperative Complications

Different types of cardiac surgeries have a varying length of median hospital length of stay. According to the latest report by SCTS [6], in the UK in 2015-2016, the CABG patients stayed in hospital for the median time of 6 days (6 days for elective, 7 days for urgent, 8 days for emergency). For valve surgery patients, the median postoperative length of stay was 8 days. Isolated aortic valve replacement (AVR) surgery patients stayed in the hospital for the median time of 8 days. Combined AVR and CABG surgeries on average have resulted with 9 days median postoperative length of stay [6].

Hospital length of stay has been shown to be connected to postoperative complications [229]. Complications can vary, depending on type of surgery and the patient's medical history. Some common complications following general surgery include haemorrhage, basal atelectasis (minor lung collapse), blood loss, acute myocardial infarction, pulmonary embolism, septicaemia and low urine output [230]–[232].

While both general and cardiac surgery patients experience postoperative complications, cardiac patients have been shown to have a higher risk for reduced quality of life than patients undergoing general surgery, due to higher stress to the body the cardiac surgery adds. In addition, cardiac patients are more likely to experience poorer quality of life six months after cardiac surgery [233]. This is because cardiac patients can have major complications such as myocardial infarction, respiratory failure, renal failure, and stroke [7], [234].

It is important to note that measuring only mortality does not show long-term implications to the patient and to the healthcare system as a whole. Being able to improve collecting data on complications can help with managing patient's expectations. As postoperative complications, regardless of surgery, have an impact on both short and long term outcomes, it is important to improve the research on postoperative complications [229]. To be able to provide a more personalised and patient-centred care, and to move from general level to more granular level, defining and classifying postoperative complications for cardiac surgery patients would be highly beneficial to develop clinical support tools offering personalised risk prediction for these complications.

4.3. Methods

4.3.1. Ethical Statement

This study was approved by the University of Strathclyde Department of Computer and Information Sciences Ethics Committee (ID 837).

4.3.2. The Delphi Method

The questions stated in Introduction are aimed to be answered via the Delphi method. The Delphi method is a well-established experts consultation method based on the premise that group opinion is more valid and reliable than individual opinion [200]. It has been defined as a multi-staged survey system which has a goal to achieve consensus on an issue, where there was no consensus before [235].

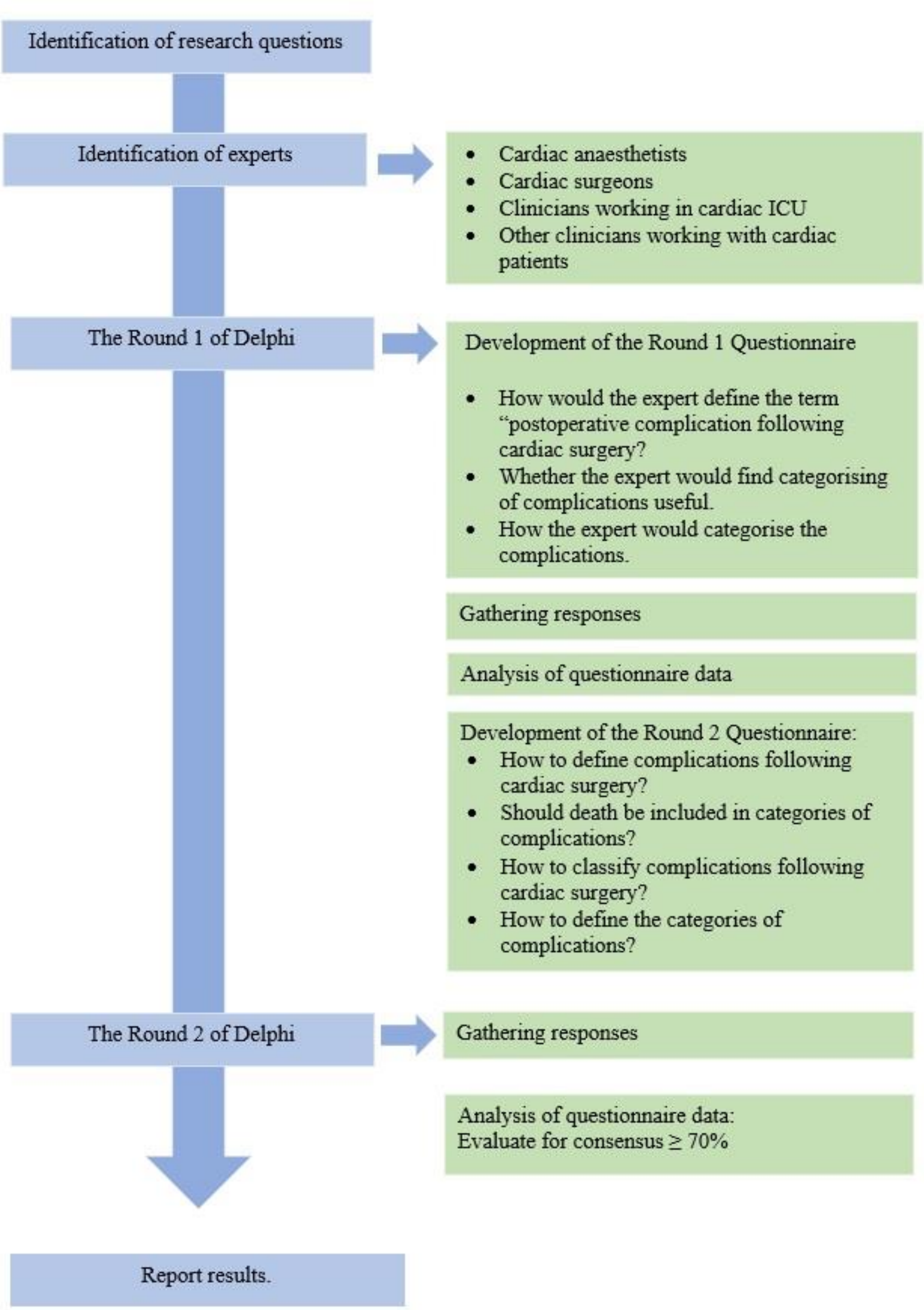
The original Delphi method, also known as the Classical Delphi, consists of two or more rounds of questionnaires administrated by post to an expert panel. The Round 1 focuses on the experts' opinion in an open-ended manner. After analysing the Round 1, the Round 2 asks the experts to rank the statements or questions according to the opinions stated in the previous round. Rounds continue until a consensus is reached on some or all the questions [200].

This study used the e-Delphi method, which is a similar process to the classical Delphi, but administered as an online web survey [200]. The overall study process is outlined in Figure 4.1.

To guarantee experts' anonymity in the study, the experts remained in both rounds anonymous, meaning the participants' responses in Round 1 and Round 2 were not linked. This decision was done due to choosing the "all-rounds" approach, where potential participants were invited to take part in subsequent rounds, regardless of whether they participated in the previous rounds. It has been shown that this approach can improve representation of opinions and can reduce the chances of false consensus [236].

The study rounds are further explained in Sections 4.4 and 4.5.

Figure 4.1. The flow of the Delphi study.


Identification of research questions
Identification of experts
Cardiac anaesthetists
Cardiac surgeons
Clinicians working in cardiac ICU
Other clinicians working with cardiac patients
The Round 1 of Delphi
Development of the Round 1 Questionnaire
How would the expert define the term "postoperative complication following cardiac surgery?
Whether the expert would find categorising of complications useful.
How the expert would categorise the complications.
Gathering responses
Analysis of questionnaire data
Development of the Round 2 Questionnaire:
How to define complications following cardiac surgery?
Should death be included in categories of complications?
How to classify complications following cardiac surgery?
How to define the categories of complications?
The Round 2 of Delphi
Gathering responses
Analysis of questionnaire data:
Evaluate for consensus ≥ 70%
Report results.

4.3.3. Identification of Experts

Cardiac surgery experts were identified as follows: cardiac anaesthetists, cardiac surgeons, and clinicians working with cardiac patients perioperatively or in intensive care. Since this was a study to develop a definition and classification for postoperative complications in cardiac surgery, mailing lists of the following professional associations were used to invite prospective participants to the Delphi study: *Association for Cardiothoracic Anaesthesia and Critical Care* (UK based, affiliated with the Royal College of Anaesthetists) [237], *European Association of Cardiothoracic Anaesthesiology and Intensive Care* (members from 35 European countries) [238], *The Society for Cardiothoracic Surgery* (UK and Ireland) [239], and *The UK Society for Computing and Technology in Anaesthesia* (UK based) [240]. Through these avenues, the invitation was sent to thousands of potentially eligible participants, depending on the number of members in each society.

In addition to the above, cardiac anaesthetists and cardiac surgeons in three Scottish cardiac centres were contacted directly via email: the Golden Jubilee National Hospital, the Royal Infirmary of Edinburgh, and Aberdeen Royal Infirmary (64 potential participants, 27 of them cardiac surgeons and 37 of them cardiac anaesthetists). Even though some international societies were contacted, it is important to note that the questionnaires were in English only.

4.3.4. Methods of Analysis

The Delphi process involves both qualitative and quantitative data analysis. The data in this study were collected through online questionnaires via Qualtrics [241]. Once the questionnaire was closed, the data were exported from Qualtrics and stored in Microsoft Excel spreadsheet for analysis. RStudio [242] and NVivo [50] were used for quantitative and qualitative analysis, respectively. Since the questionnaires were anonymous, each expert was coded as Rx.Py, where x is the number of the study round and y is the number of the participant in the round.

4.3.4.1. Consensus

The consensus level was determined to be 70%, similarly to other related studies in health research [243]–[245]. Descriptive statistics were used to analyse the opinions of experts, using frequencies of responses for questions that were not open-ended. If the frequency of a response was 70% or higher, the experts were deemed to have reached a consensus on this particular response.

All responses were considered in the analysis; however, the consensus was calculated based on how many experts answered each question. Partially filled responses were also included, as other published studies have done in the past [246], [247]. This is done because the experts eligible for this study are commonly under great pressure from work commitments, and hence including their responses is respectful towards the respondents' time and effort to participate.

4.3.4.2. Qualitative Analysis

The Round 1 of the study largely included open-ended questions to determine a variety of ways the experts would choose to define complications in cardiac surgery and to define categories of postoperative complications.

The thematic analysis framework [168] was used to analyse the responses to the open-ended questions in both study rounds. The answers to these open-ended questions were analysed and the results were included as options for responses in the subsequent rounds of the study [248].

Thematic analysis is a method used for identifying, analysing and reporting patterns within qualitative data, and has six phases [168]:

1. Familiarising with data
2. Generating initial codes
3. Searching for themes
4. Reviewing themes
5. Defining and naming themes
6. Producing the report.

Following the guidance of Hasson et al. [248], the statements that were identified as identical or similar were grouped as common concepts. Once specific themes were created, the statements within a thematic group were synthesised into a single summary statement. The wording was kept as close as possible to the statements that had been provided by the experts. Any unique statements provided by the experts with no related statement were kept as worded originally and included directly in Round 2. For reliability check, two participants' responses from Round 1 were randomly selected to be coded by also two other PhD students who have extensive experience in qualitative analysis methods (Dr Diane Morrow and Ramsay Meiklem). Any discrepancies were discussed with Dr Matt-Mouley Bouamrane at supervisory meetings and resolved, as appropriate.[5]

To capture the respondents' opinions as objectively as possible, a comment box was provided after each question to collect respondents' qualitative comments. This is to also increase respondents' ownership in Delphi studies [248]. These comments were also analysed, using the thematic analysis framework [168]. The collated themes of these comments were then discussed in the results of this chapter to further explain the position of experts regarding defining and categorising complications following cardiac surgery.

4.4. The Round 1 of the Study

4.4.1. Development of the Questionnaire

For the Round 1, the questionnaire was designed to explore the experts' general opinions regarding the definition of "postoperative complication following cardiac surgery" and categorising postoperative complications.

The questionnaire (see Appendix 4.1) started with a filter question to make sure that only the eligible experts would be included in the study: *"Are you in any way involved with cardiac surgery patients? (Can be preoperatively, intraoperatively and/or postoperatively.)"* If the answer to the question was "no", the expert was directed to the end of the survey.

The questionnaire consisted of three parts below:

1. The background of the expert.
2. How the expert would define the term "postoperative complication" following cardiac surgery.
3. Whether the expert would find categorising of complications useful; and if yes, how the expert would categorise the complications.

[5] How themes were allocated to participant responses can be found from https://doi.org/10.15129/fa6c4bc4-98dc-48cd-a568-d89c8a52dd10

The Round 1 questionnaire was sent out twice to professional societies and to other potential experts between 27th August 2019 and 24th September 2019. In total, the Round 1 of the questionnaire was open for 6 weeks and closed on 8th October 2019.

4.4.2. Expert Demographics

Overall, 71 experts took part in the Round 1 of the study based on being involved with cardiac surgery patient pathway. The majority, i.e., 67 respondents out of 71 (94%) of the respondents were based in the United Kingdom, two (3%) were from Saudi Arabia, one was from Australia (1%) and one from Bahrain (1%).

Most of the respondents (45 out of 71, 63%) specialised in both cardiac anaesthesia and cardiac critical care, 23 (32%) specialised in cardiac anaesthesia only and 3 (4%) specialised in cardiac critical care only. It is important to note that none of the participants stated to be cardiac surgeons. This is further discussed in Limitations (Section 4.6.1). In terms of experience, the mean number of years worked in the specialty was 16.63 (SD = 8.70) years and the median number of years was 16 (IQR = 12.5).

As shown in Table 4.2, most of the participating experts were involved with the surgery itself (67 out of 71, 94%), decision making (e.g., if patient is fit for surgery) (64 out of 71, 90%), preoperative assessment (63 out of 71, 89%) and cardiac intensive care unit (63 out of 71, 89%). Some respondents also were involved with long-term follow-up of the patient (8 out of 71, 11%) and in other ways (7 out of 71, 10%), such as acute and chronic pain management and perioperative echocardiography.

Table 4.2. Experts' involvement in cardiac patient pathway

	n/N (%)
The surgery itself	67/71 (94%)
Decision making (e.g., if patient is fit for surgery)	64/71 (90%)
Preoperative assessment	63/71 (89%)
CICU	63/71 (89%)
Long-term follow-up of the patient	8/71 (11%)
Other	7/71 (10%)

4.4.3. Defining the Term “Postoperative Complication”

Fifty experts commented on how they would define the term “Postoperative Complication” in cardiac surgery. The list of possible definitions that were collated from experts’ responses are as follows:

- An unplanned adverse event occurring after cardiac surgery that may be caused or compounded by the surgical process.
- An unplanned adverse event arising as a result of cardiac surgery, which was otherwise unlikely to have occurred in the same period.
- Any adverse event that impairs a patient’s physical, cognitive, psychological or emotional function and quality of life.
- Any deviation from the ideal recovery pattern after cardiac surgery.
- Unexpected, or expected but unwanted, outcome of cardiac surgery which significantly delays recovery from the procedure compared to the desired outcome or leads to the patient failing to derive the intended benefits of surgery.
- Any event resulting from surgery which lengthens the patient’s stay in hospital or reduces their quality of life beyond normal.
- Any unplanned clinical event that leads to a delay in hospital discharge or requires additional treatment or intervention to mitigate or reverse the event.
- Any deviation of any physiological system which adversely affects rapid recovery to good health.
- An event which may have an impact on patient’s survival or quality and longevity.

All these definitions focus on different impacts of complications on patient, institution, and surgery itself, e.g., delayed recovery, impact on patient’s quality of life and hospital length of stay. Hence, for simpler analysis, these statements were analysed thematically [168] and categorised under themes based on the definitions that the experts offered. For example, the definition *“An unplanned adverse event occurring after cardiac surgery that may be caused or compounded by the surgical process”* includes themes of “*The event can be unplanned*”, “*The event must be harmful or* *unfavourable*", "*The complication must be present following cardiac surgery, specifically*", and "*The event must occur after surgery and is unlikely to occur if the patient did not have the surgery*". For simplicity, these themes were then grouped under characteristics that complications could have (Table 3.4), with the example definition including the characteristics of "unplanned", "adverse event", "cardiac surgery" and "surgery".

Table 4.3 shows the eight themes that were collated in the analysis, together with their assigned characteristics. The responses for each definition were then mapped onto each characteristic to find out what the experts will deem to be important to define what constitutes a postoperative complication following cardiac surgery.

Table 4.3. Characteristics and their descriptions collated from how experts would define "postoperative complication following cardiac surgery".

Theme	Assigned Characteristic
The event must be harmful or unfavourable.	Adverse event
The event can have an impact on patient's survival or quality of life and longevity.	Affects quality of life
The event can have an impact on hospital length of stay.	Delay in hospital discharge
Due to the event the patient might have to stay in the hospital for longer and can adversely affect rapid recovery to good health.	Delay in recovery
The complication must be present following cardiac surgery, specifically.	Following cardiac surgery, specifically
The event must occur after surgery and is unlikely to occur if the patient did not have the surgery.	Due to surgical process
The event can be unexpected.	Unexpected
The event can be expected, but unplanned.	Unplanned

4.4.4. Usefulness of Classifying Postoperative Complications

Fifty-one experts answered the question as to whether they thought it was useful to define and classify postoperative complications for cardiac surgery. Out of 51 experts (Table 4.4), 23 (45%) thought it was "Extremely useful" and 20 (39%) thought it is "Very useful". Combining these percentages, based on the pre-determined consensus level of 70%, it can be concluded that the experts have reached the consensus that it is very useful to classify postoperative complications for cardiac surgery with a consensus level of 84%.

Table 4.4. Experts' opinion on usefulness to classify postoperative complications following cardiac surgery.

Usefulness	n/N (%)
Extremely useful	23/51 (45%)
Very useful	20/51 (39%)
Moderately useful	5/51 (10%)
Slightly useful	2/51 (4%)
Not at all useful	1/51 (2%)

The experts were also asked to justify their answer regarding the usefulness to classify complications following cardiac surgery. Following thematic analysis, the experts' responses could be grouped under four main themes: (1) audit and quality measurement, (2) planning and management, (3) risk management and communication, and (4) research. The collated themes and their subthemes are shown in Table 4.5.

Table 4.5. Themes and subthemes for the rationale for defining and categorising complications.

Theme	Subtheme
Audit and quality measurement	
	Quality improvement
	Benchmarking and comparison
	Gathering information
Planning and management	
	Care improvement
	Communication with different teams
	Management of care
	Resource allocation
Risk management and communication	
	Communication about risk
	Patients' and carers' consent
	Mitigation of risk
	Understanding about risk
Research	
	Patient outcomes
	Quality of research
	Comparison of studies
	New treatments

Audit and quality measurement

Based on experts' comments, defining and categorising complications could be useful for audit and quality measurement reasons. Currently cardiac centres are rigorously audited based on cases of mortality [6].

"Currently mortality is recorded universally, while morbidity isn't which has an impact on longevity and quality of life." (Expert R1.P76)

To discuss and learn from current mortalities and major morbidities following cardiac surgery, Mortality and Morbidity (M&M) meetings are held in cardiac centres. M&M meetings aim to support a systematic approach to the review of patient deaths or care complications, to improve patient care and provide professional learning [249]. However, currently M&M meetings have been shown to focus more on mortality only [250], [251], as also shown in Chapter 3. It was mentioned by experts that better markers of quality of care rather than mortality are needed.

One of the reasons why only mortality is vigorously audited and discussed at M&M meetings is poor data collection for postoperative complications. Experts expressed that defining and categorising complications would simplify gathering information on complications.

"If we do not know what is going wrong, we cannot work out how to stop it happening." (Expert R1.P23)

Accurate recording of frequency of complications would not only help with auditing in general, but also with the other themes found in this analysis, such as hospital planning and management, risk management and research.

Defining and categorising complications would also help with quality improvement by determining effects of quality improvement strategies, and benchmarking units' and hospital performances.

"[Defining and classifying complications] allows comparison of outcomes which may be more useful than pure mortality data." (Expert R1.P48)

Planning and management

It was expressed that defining and classifying complications following cardiac surgery is necessary for perioperative planning and management reasons. It is especially important to create methods for care improvement and to prevent complications.

"Categorising events is useful if a common cause can be found and addressed; more often it is important to treat and recognise the complications early." (Expert R1.P21)

Knowing how complications are categorised would also help with communication, and hence could potentially enhance discussion amongst multi-disciplinary teams by providing a common terminology.

Experts also mentioned how a standardised approach could help with management of care, including perioperative management and adoption of different working practices.

Finally, it was also said that categorisation of complications could also help with resource allocation and bed planning.

> *"Classification may help to understand causative factors and allocation of resources in prevention." (Expert R1.P56)*

Risk management and communication

It was thought that defining and categorising complications helps to manage and communicate about risk. Experts also highlighted the importance of communication about risk, not only between experts, but also to patients and their families to help with informed decision making.

> *"This [classification of complications] could then be used to good effect in discussions with patients and families as they would gain consistent information from various members of the MDT." (Expert R1.P13)*

Good communication about risk will also result in better understanding about potential adverse outcomes, and helps the patient to give an informed consent to have a procedure as the patient will have a fuller understanding of the potential adverse consequences.

> *"[Classification of complications] may allow better explanation of procedures and complications to patients to allow more thorough imparting of knowledge prior to consenting." (Expert R1.P48)*

Research improvement

Experts thought defining and categorising complications following cardiac surgery would help with improving research in postoperative complications. Through research, having a standardised approach could help to improve patient outcomes. The experts said that research could then help to understand causative factors to later prevent complications.

> *"[Categorising complications would be useful] to facilitate [...] research and to target therapies appropriately to prevent or decrease incidence." (Expert R1.P61)*

In addition to patient outcomes, experts also commented that having a standardised approach could also help to improve the quality of research itself.

> *"[Defining and categorising complications is extremely useful] to standardise outcome and adverse event reporting in studies to allow more meaningful research, especially systematic reviews to occur." (Expert R1.P10)*

The quality of research could also be improved by achieving better comparability of studies and would allow the impact of new therapies to be determined.

4.4.5. Classification of Postoperative Complications in Cardiac Surgery

Overall, N=48 experts stated how many categories postoperative complications should have. Most of the respondents wanted 3 to 5 grades to categorise complications (Table 4.6), where 16 out of 48 respondents (33%) vote for 3 grades, 12 respondents (25%) for 4, and 14 respondents (29%) voted for 5 grades.

Table 4.6. How experts voted for how many grades should there be to categorise complications.

Number of grades	n/N (%)
2 grades	3/48 (6%)
3 grades	16/48 (33%)
4 grades	12/48 (25%)
5 grades	14/48 (29%)
6 grades	3/48 (6%)

Some (26 respondents out of 48, 54%) also named the categories they offered, and it became clear that respondents offered the following variations as a common answer:

"Mild /Moderate/ Severe"

"None / Mild / Moderate / Severe"

"Mild / Moderate / Severe / Death"

"None / Mild / Moderate / Severe / Death"

This means that the consensus was reached that the categories for postoperative complications for cardiac surgery will be classified as “Mild”, “Moderate” and “Severe”. Since many respondents did offer to add “Death” also as a separate class, the experts were asked to decide whether to add that to the categories in Round 2 of the study. Since no complication would be categorised as “None”, this was not added into the categories.

4.4.6. Defining the Categories of Postoperative Complications

Experts also provided possible definitions for each grade that they proposed. To analyse the suggested definitions, the thematic analysis, explained in detail in Section 4.3.4.2, focused on characteristics that each complication category could have. Like in Section 4.4.3, the characteristics provided by experts for each category of complications were collated so that similar characteristics were merged into one, and unique characteristics were left in their initial form [235]. The final list of characteristics, proposed by experts, were as follows:

- Effect on overall length of stay in hospital
- Effect on final outcome
- Length of the complication
- Clinical relevance
- Impact on the patient
- Occurrence of the complication
- Clinical intervention is required
- Impact on the institution

These factors were then related to a level of complication. For example, the question “What is the effect on overall length of stay in hospital?” was turned into “No notable effect on overall length of stay” for Mild level of complication, “Some effect on overall length of stay” for Moderate and “Extended length of stay” for Severe complication. These statements were then used in Round 2 of the Delphi study so experts could vote on which characteristics are most important to define each complication category.

4.5. The Round 2 of the Study

4.5.1. Development of the Questionnaire

The Round 2 survey (see Appendix 4.2) of the Delphi study was sent out to the same societies and contact list from the Scottish cardiac centres as described in Section 4.3.3. To take part in Round 2, the experts were not required to have taken part in Round 1 of the study, as per the "all rounds" approach [236]. Just like in Round 1, the experts had to answer the filter question to make sure they were eligible to participate.

The aims of the Round 2 of the study were to reach a consensus regarding the following:

1. how the experts would define what constitutes to a "postoperative complication" following cardiac surgery based on the responses from the Round 1 of the study;
2. whether death should be included in the categories of complications; and
3. how the experts would define each category of complications based on the characteristics collated from the Round 1 of the study.

The choices for answers for the questions were collated based on the results of Round 1 of the study. Just like in Round 1, descriptive statistics were used to analyse the experts' opinions, using frequencies of responses for questions that were not open-ended. If the frequency of a response was 70% or higher, the experts were deemed to have reached a consensus on this particular response.

Round 2 of the questionnaires were sent out on 2nd June 2020 and a reminder was sent out on 16th June 2020. The survey was open for 4 weeks (closed on 30th June 2020)[6].

[6] The author is aware of the limitation the notable gap between the time of Round 1 of the study and the Round 2 of the study (8 months) can bring. The possibility of losing the momentum of the study might be the reason why less participants took part in the Round 2. The Round 2 of the study took place later than planned due to unexpected circumstances the author experienced. Furthermore, as the Round 2 took place during the height of the COVID-19 pandemic, it can be expected that the expert group targeted in this study played a crucial role in managing the pandemic, making them less likely to take part of the study.

Overall, 46 experts took part in the survey and 37 of them finished the survey. As done in previous round, this time also responses from participants that filled the survey partially, were included.

4.5.2. Experts' Definition of What Constitutes "Postoperative Complications Following Cardiac Surgery"

Thirty-eight experts voted for each characteristic (see Section 4.4.3) to define what constitutes a complication after cardiac surgery. A consensus was reached that all characteristics (Table 4.7), apart from "Unexpected" should be included in the final definition.

Combining these characteristics into a sentence, resulted in the following definition:

> *A complication following cardiac surgery is an* ***unplanned adverse event*** *that occurs* ***following cardiac surgery*** *that can cause* ***delay in recovery, delay in hospital discharge*** *and* ***affect patient's quality of life*** *and is likely to happen* ***due to surgical process****.*

Table 4.7. How experts voted for each characteristic that defines the term "postoperative complication after cardiac surgery".

Theme	Complication Characteristic	n/N (%)
The event can have an impact on patient's survival or quality of life and longevity.	Affects quality of life	35/38 (92%)
The complication must be present following cardiac surgery, specifically.	Following cardiac surgery, specifically	33/38 (87%)
The event must occur after surgery and is unlikely to occur if the patient did not have the surgery.	Due to surgical process	33/38 (87%)
The event must be harmful or unfavourable.	Adverse event	28/38 (74%)
The event can have an impact on hospital length of stay.	Delay in hospital discharge	28/38 (74%)
Due to the event the patient might have to stay in the hospital for longer and can adversely affect rapid recovery to good health.	Delay in recovery	28/38 (74%)
The event can be expected, but unplanned.	Unplanned	27/38 (71%)
The event can be unexpected.	Unexpected	23/38 (61%)

4.5.3. Including "Death" in the Classification of Postoperative Complications

Out of 37 experts, 31 (84%) thought that "Death" should be included in the classification of postoperative complications. As a result, a consensus has been reached that the complications should be categorised in four levels:

"Mild", "Moderate", "Severe" and *"Death"*.

The respondents offered various explanations regarding this opinion, overall stating that death is the ultimate complication.

> *"It [Death] is the ultimate undesired complication." (Expert R2.P1)*

> *"This is obvious, complications can be mild/moderate/severe, but death must be recognised separately. Death is the real "yes or no" decision maker for the patients as well." (Expert R2.P34)*

However, the experts agreed with the justification of this study that more emphasis needs to be put into researching and recording complications as well.

> *"It [death] is a finite endpoint. However, the focus needs to move to other complications rather than the current focus on death per se as the main outcome." (Expert R2.P2)*

4.5.4. Defining the "Mild", "Moderate", "Severe" Complication Categories

Based on the proposed characteristics that were collated from experts' responses (described in Section 4.4.6) a consensus was reached on definitions for "Mild" complications (Table 4.8). Hence, a complication following cardiac surgery is classified as "Mild" if the complication has the following characteristics:

- The complication has no consequential effect on the final patient outcome (28 out of 37 (76%) respondents).
- The complication has a minimal impact on patient (27 out of 37 (73%) respondents).

Table 4.8. The characteristics of "Mild" complications.

Characteristic	**n/N (%)**
Minimal impact on patient	28/37 (76%)
No consequential effect on final outcome	27/37 (73%)
No or only short-term clinical relevance	19/37 (51%)
No or small amount of intervention required	19/37 (51%)
No notable effect on overall length of stay	17/37 (46%)
Mildly debilitating	7/37 (19%)
Common	7/37 (19%)
Minimal impact on institution	6/37 (16%)
Lasting 1 week – 1 month	4/37 (11%)

Similarly, as shown in Table 4.9, a complication following cardiac surgery is classified as "Severe" if the complication is:

- Potentially life-threatening (34 out of 37 (92%) of respondents).
- There is a consequential or long-standing impact on the patient (31 out of 37 (84%) respondents).
- A notable amount of intervention is required due to this complication (26 out of 37 (70%) respondents).

Table 4.9. The characteristics of the "Severe" complications.

Characteristic	**n/N (%)**
Potentially life-threatening	34/37 (92%)
Consequential or long-standing impact on the patient	31/37 (84%)
Notable amount of intervention required	26/37 (70%)
Extended length of stay	25/37 (68%)
With sustained relevance and life-limiting	25/37 (68%)
Severely debilitating	21/37 (57%)
Lasting 3 months – 1 year	7/37 (19%)
Notable or long-standing impact on institution	5/37 (14%)
Uncommon	2/37 (5%)

The experts did not reach a consensus on the definition for "Moderate" complications due to none of the characteristics receiving 70% or more votes (Table 4.10). However, one could argue that the definition of moderate is known as it is neither mild nor severe. This is further discussed in Section 4.6.1.

Table 4.10. The characteristics of "Moderate" complications.

Characteristic	n/N (%)
Some effect on overall length of stay	23/37 (62%)
Acutely important, but less clinical consequence long-term	22/37 (59%)
Some intervention required	22/37 (59%)
Some effect on final outcome	20/37 (54%)
Moderately debilitating	19/37 (51%)
Limited impact on patient	18/37 (49%)
Lasting 1 – 3 months	4/37 (11%)
Less common	4/37 (11%)
Limited impact on institution	4/37 (11%)

Finally, experts were asked to provide examples for each proposed complication level (see Appendix 4.3), which included atrial fibrillation, constipation and pain that is resolved with analgesia as "Mild" complications; pneumonia, bleeding, and prolonged sedation as "Moderate" complications; and acute renal failure, cardiac arrest and stroke as "Severe" complications. These examples, however, should be interpreted with caution as postoperative outcomes and the actions taken to avoid these should be patient-focused, rather than institution-focused, as also stated by experts.

> *"[The way we look at complications] should be patient-centred, rather than dependent on institutional or team consequences." (Expert R2.P11)*

Furthermore, as seen in Appendix 4.3, the examples of the experts reflect the non-consensus in what constitutes to a moderate complication. For example, atrial fibrillation can be categorised as "Mild" in some cases, however, depending on its effect on patient's well-being, it can also be categorised under "Moderate" at more severe cases. The example complications, together with the definitions and characteristics of the complications, were used as a guide to group the postoperative complications reported in CaTHI database (see Section 5.5.1) to predict "Severe" postoperative complications (Chapter 6).

4.6. Discussion

This chapter showed the results of a Delphi study which aimed to define and categorise complications following cardiac surgery. The study reached a consensus on the following:

- It is useful to define and categorise complications following cardiac surgery

- how the complications following cardiac surgery are defined; and
- how the complications following cardiac surgery are classified.

The experts justified the usefulness of defining and categorising surgical complications following cardiac surgery by stating it could help with audit and quality control, planning and management, risk management and communication, and research.

Consensus was reached on the characteristics of postoperative complications, and hence the following definition was formed:

> *A complication following cardiac surgery is an unplanned adverse event that occurs following cardiac surgery that can cause delay in recovery, delay in hospital discharge and affect patient's quality of life and is likely to happen due to the surgical process.*

In the Clavien-Dindo classification system, complications were defined as *"any deviation from the normal postoperative course"*, and conditions which are inherent to the procedure and are expected, were termed to be "sequelae" [204]. However, the definition from this Delphi study provides a more precise explanation on what a complication is. Also, as the Clavien-Dindo definition was created for general surgery, the definition presented in this study makes an important point that the Clavien-Dindo definition does not: a complication following cardiac surgery is an event that is unlikely to happen without surgery, and in this case, cardiac surgery. When it comes to the definition of "sequelae", it can be argued that some adverse events following surgery can be expected, especially with existing and emerging preoperative prediction models. With improved data collection in electronic health records, more models predicting complications following surgery can be developed, meaning that many complications can be predicted and monitored on a real-time basis. Various studies have been published to predict fluid requirement [252], septic complications [253], hypotensive episodes [69] and clinical deterioration in general [254].

This study achieved a consensus on how to categorise complications following cardiac surgery, and how the categories are defined. It was agreed that the categories should be: "Mild", "Moderate", "Severe", and "Death". According to the experts, a "Mild" complication is a complication that has no consequential effect on the final patient outcome and has minimal impact on patient. The experts agreed that a "Severe" complication is a complication that is potentially life-threatening, requires notable amount of intervention, and has a consequential or long-standing impact on the patient.

4.6.1. Limitations

4.6.1.1. Study Sample

In the Round 1 and Round 2 of the study, 51 (out of 71) and 37 (out of 46) experts completed the study, respectively. Notably, a considerable number of participants dropped out from the questionnaires (29% and 20%, respectively). However, according to publications discussing the Delphi method, both rounds of the study had a sufficiently large sample size as it does not depend on statistical power, but rather on group dynamics for coming to a consensus among experts. Hence, the expert panel usually consists of 10 – 30 experts [255].

As seen from the results of the study, most experts consisted of cardiac anaesthetists and intensivists, however, no cardiac surgeons took part in the study. Historically, the decision as to whether a patient will be operated upon will be mainly made by the surgeon, as also shown in the findings of Chapter 3. Understanding surgeons' view on defining and classifying complications in cardiac surgery would be useful. However, 90% of the participants in this study were involved with decision making, which is common with the creation of preassessment clinics, where decisions about patient care are made by multi-disciplinary teams [256].

While this study achieved a consensus on the definition and classification of what constitutes to postoperative complications following cardiac surgery, this study may be considered as a "pilot" study from a medical contribution perspective. To overcome this, in future work a more international panel of experts is needed to increase the impact of the classification system. While experts within European Association of Cardiothoracic Anaesthesiology and Intensive Care were invited, most of the professional societies were UK-based societies, which explains the lack of responses from international experts. Since the standards in cardiac surgery are alike internationally [256], it is likely that results would be similar, however, the consensus would be more representative and more reliable to be put into practice. In addition, the societies were mostly related to cardiac anaesthesia, only one (The Society for

Cardiothoracic Surgery) being specific for cardiac surgeons. This explains why no cardiac surgeons took part in the study. Hence, in the future study, cardiac centres will be contacted directly to allow for more international panel, and more efforts will be directed towards recruiting more cardiac surgeons to participate.

4.6.1.2. Defining “Moderate” Complications

No consensus on the definition of “Moderate” complication was reached. Delphi studies do not always reach a consensus on all aspects of the study [257]. Categorisation decisions are often made based on the extreme categories rather than based on the middle category [258]. This has been addressed with, for example, American Society of Anesthesiologists (ASA) Classification [259], where there is no “Moderate” category. Historically, there have been concerns about the subjectivity of the ASA status [260] and the same problem can occur with the complication classification in this chapter. In order to categorise complications appropriately, actions and consequences of each category need to be considered. With “Mild” complication, some medicines might have to be administered, for example for urinary retention, but in general no notable action that requires time and resources is needed. With “Severe” complication, whether it is kidney failure or a stroke, dialysis or thrombectomy, respectively, might be needed. Both interventions are time-consuming and resource intensive. When it comes to a “Moderate” category, however, it is uncertain whether it is more on a “Mild” or a “Severe” side. On one hand it gives an unclear indication for general understanding regarding what action needs to be taken, however, on the other hand, it provides the users with a spectrum of categories and therefore a possibility for offering more nuance to the problem. As shown by Mayhew et al., for ASA physical status classification, objectivity has been improved and variability in classification has been reduced through bringing example cases for each classification level [260]. Hence, we also asked experts to provide examples for each category. However, further work is needed to provide examples, and hence it important to keep in mind that for personalised use, each complication, regardless of which category it falls into, needs individual approach for treatment, depending on the patient’s current state and medical history.

"Grading may be useful in quality management and comparing outcomes from different surgical units. Simple grading does however not give any indication of cause or prevention." (Expert R1.P56)

However, following the results of this study, it is understood that the classification of "Mild", "Moderate", "Severe" and "Death" is simple and clear to experts and to patients when communicating about risk, and can offer understanding for urgency for action when the developed model predicts these categories.

4.7. Conclusion

Using the Delphi method, this study shows cardiac anaesthetists' and cardiac intensivists' requirement for a standardised definition and classification for postoperative complications in cardiac surgery. The standardisation of complication identification, recording and reporting in cardiac surgery could help the development of future quality benchmarks, clinical audit, care quality assessment, resources planning, risk management, performance comparisons, communication, and research. The proposed definition and classification will be used in this thesis to develop preoperative prediction models predicting "Severe" postoperative complications in cardiac patients (Chapter 6).

Chapter 5. Predictive Modelling Methods

5.1. Introduction

Based on the results of studies undertaken in Chapters 2, 3 and 4, predictive modelling experiments are undertaken to predict severe postoperative complications, acute kidney injury and delirium. From the prior three chapters, it was evident that currently the focus in cardiac outcome reporting is on mortality (as shown in Chapters 2 and 3), which explains why the majority of prediction models preoperatively [14], [16] and in the ICU (shown in Chapter 2) are developed to predict postoperative mortality. However, as shown in Chapters 3 and 4, clinicians working with cardiac patients find investigating into cardiac surgery complications important, especially due to aging population and more patients with co-existing conditions.

It is important to note that all analysis (including predicted outcomes, handling of variables, predictive modelling methods and performance measures) was discussed with the clinical supervisor Prof Stefan Schraag, who is a consultant cardiac anaesthetist at the Golden Jubilee National Hospital. While this thesis makes a number of contributions to knowledge in medicine, the author’s expertise lies mainly in statistics and data science, and hence the input from a clinical expert was necessary to develop a fit-for-purpose prediction model for the predicted outcomes.

As explained in Chapter 2, the definition of mortality is straight-forward, however complications can have various ways of diagnosing them. Hence, experiments were undertaken in Study 1 to predict severe complications, using the definition from the Delphi study (Chapter 4), further explained in Section 5.6.1. In addition, experiments were undertaken in both Study 1 and Study 2 to predict complications that cardiac clinicians deemed to be serious (Chapter 3). Since septic complications, which are considerably well defined, were predicted often by other models, found in the literature review (Chapter 2), other complications that had internationally recognised diagnosis definitions were considered. Acute kidney injury (AKI) was predicted by only one paper [74] that had many limitations (explained in Chapters 2), and no papers that fitted the inclusion criteria in Chapter 2 were identified to predict postoperative delirium (further discussed in Chapter 8). Since both AKI and delirium have internationally recognised diagnosis criteria available, and are not commonly predicted, these two outcomes were chosen (further defined in Sections 5.6.2 and 5.6.3, respectively).

This chapter provides information on the methods used in this thesis. The methods are described based on the guidance of the TRIPOD statement [261]. The TRIPOD checklist for prediction model development and validation for the models developed in this thesis can be found from Appendix 5.1.

The methods are described for two studies:

> **Study 1:** Using preoperatively available data to develop preoperative prediction models (Chapter 6) to predict:
>
> > **1.1:** severe complications following cardiac surgery.
> >
> > **1.2:** acute kidney injury following cardiac surgery.
> >
> > **1.3:** delirium following cardiac surgery.
>
> **Study 2:** Using preoperative and intensive care unit data to develop hourly prediction models in intensive care unit to predict:
>
> > **2.1:** acute kidney injury following cardiac surgery (Chapter 7).
> >
> > **2.2:** delirium following cardiac surgery (Chapter 8).

5.2. Ethics

This study has an ethical approval from the Health Research Authority (REC18/YH/0366). In addition, Linda Lapp has signed an honorary contract with the Golden Jubilee National Hospital, giving her the right to read-only access to the databases under the supervision of the Database Managers Sadia Aftab (CaTHI) and Debbie McKechnie (Centricity CIS).

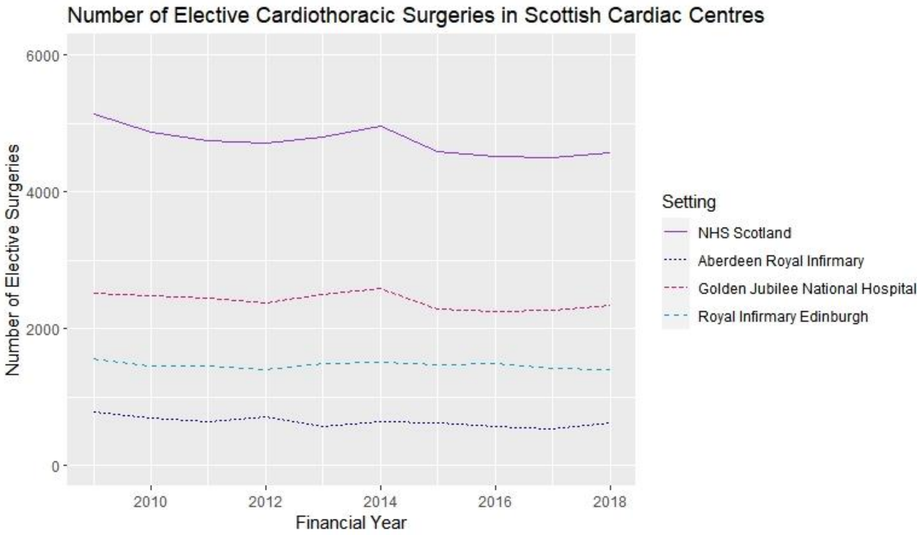
As this is a retrospective study of existing electronic health records, no patient contact was part of this PhD research.

5.3. Setting

In Scotland, there are three hospitals that are specialised in cardiac surgery: Golden Jubilee National Hospital (GJNH), Royal Infirmary Edinburgh and Aberdeen Royal Infirmary, GJNH being the largest.

GJNH is Scotland’s flagship hospital, offering world class centres for heart and lung services, orthopaedics, ophthalmology, and diagnostic imaging. It acts as a separate NHS Board and is also a National Waiting Times Centre to assist other NHS Boards with reducing patient waiting times. The Scottish Advanced Heart Failure Service, including the Heart Transplant Unit, the Scottish Pulmonary Vascular Unit and the Scottish Adult Congenital Cardiac Service are all located at the GJNH. NHS Golden Jubilee also hosts the Golden Jubilee Research Institute and the Golden Jubilee Innovation Centre.

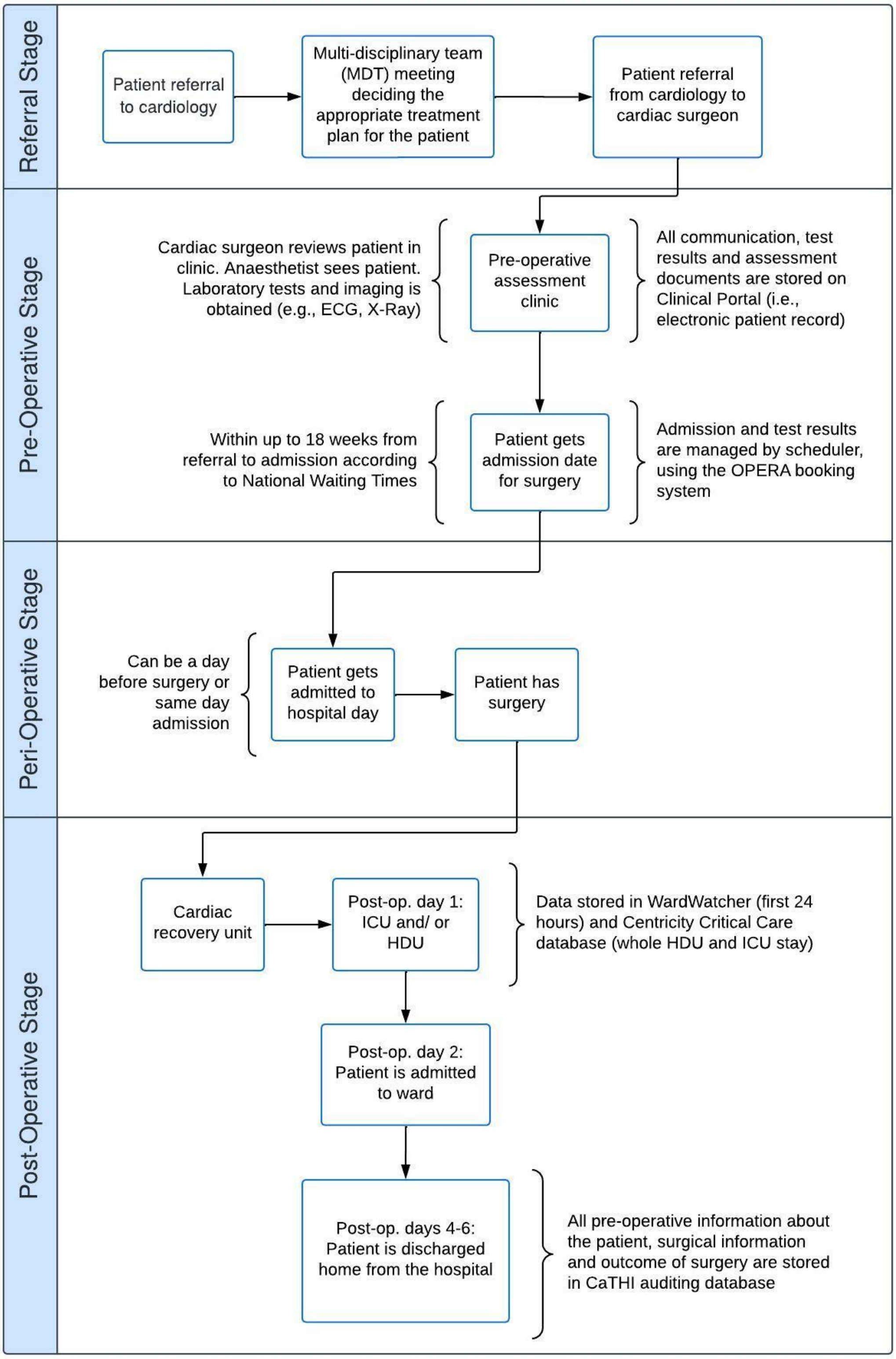
Figure 5.1. Number of elective cardiothoracic surgeries in Scottish cardiac centres and NHS Scotland in general per financial year (Data from Public Health Scotland [262]).



The Figure 5.1 shows the number of elective cardiothoracic surgeries at Scottish cardiac centres and in NHS Scotland from the financial years starting in 2009 to the end of 2018. In this timeframe, GJNH has carried out on average 2,407 (SD = 119.0) elective cardiothoracic surgeries per year, which makes just over 50% of all elective cardiothoracic surgeries in NHS Scotland. Royal Infirmary Edinburgh carries out around 30% and Aberdeen Royal Infirmary around 20% of the elective cardiothoracic surgeries in NHS Scotland [262].

The Figure 5.2 shows the patient's journey who has elective cardiac surgery at the GJNH. The figure also shows in which databases certain information about the patient are stored in.

Figure 5.2. Patient's journey for elective cardiac surgery at the Golden Jubilee National Hospital.



Referral Stage
Patient referral to cardiology
Multi-disciplinary team (MDT) meeting deciding the appropriate treatment plan for the patient
Patient referral from cardiology to cardiac surgeon
Pre-Operative Stage
Cardiac surgeon reviews patient in clinic. Anaesthetist sees patient. Laboratory tests and imaging is obtained (e.g., ECG, X-Ray)
Pre-operative assessment clinic
All communication, test results and assessment documents are stored on Clinical Portal (i.e., electronic patient record)
Within up to 18 weeks from referral to admission according to National Waiting Times
Patient gets admission date for surgery
Admission and test results are managed by scheduler, using the OPERA booking system
Peri-Operative Stage
Can be a day before surgery or same day admission
Patient gets admitted to hospital day
Patient has surgery
Post-Operative Stage
Cardiac recovery unit
Post-op. day 1: ICU and/ or HDU
Data stored in WardWatcher (first 24 hours) and Centricity Critical Care database (whole HDU and ICU stay)
Post-op. day 2: Patient is admitted to ward
Post-op. days 4-6: Patient is discharged home from the hospital
All pre-operative information about the patient, surgical information and outcome of surgery are stored in CaTHI auditing database

5.4. Participants

In total, preoperative data for 7354 adult patients was extracted from the CaTHI database. Patients undergoing coronary artery bypass graft (CABG), valve, and combined CABG and valve surgery at the GJNH between the 1st of April 2012 and 21st of December 2018 were included in this study.

Patients who had not been discharged from the hospital by the time of data extraction were also excluded from the analysis due to not having their final outcome recorded in the CaTHI database (i.e., dead or discharged, see Section 5.5.1 for information about the database).

Patients with "salvage" priority for surgery were excluded. Patients who received dialysis regardless of renal function prior to surgery were also excluded. Patients with unknown NYHA grade, unknown previous myocardial infarction (MI) status and unknown hypertension history were excluded. This is due to a very small group of patients having these characteristics.

For numerical variables, patients with obviously incorrect entries were excluded, such as negative ICU hours and body mass index (BMI) less than 10 or higher than 70. Finally, only the records that occurred in the dataset for the patient for the first time (unique entries) were included in the analysis.

Table 5.1. Number of patients in each study and experiment.

Study		**Procedure Dates**	**Final Number of Patients**
Study 1	1.1: Severe Complications	01/04/2012 – 31/12/2018	N = 6839
	1.2: Acute Kidney Injury	01/04/2012 – 31/12/2018	N = 6839
	1.3: Delirium	01/01/2016 – 31/12/2018	N = 3344
Study 2	2.1: Acute Kidney Injury	01/04/2012 – 31/12/2018	N = 6294
	2.2: Delirium	01/01/2016 – 31/12/2018	N = 3322

The final number of patients included into each study is shown in Table 5.1 The patient population for delirium prediction studies was notably smaller due to delirium diagnosis being recorded in this institution in the ICU since 2016 onwards. The delirium diagnosis was done, using CAM-ICU score. This will be further discussed in Section 5.6.3. The Study 2 population is smaller than it is for Study 1 in general, because patients who had the predicted outcome (i.e., AKI or delirium) within the first hour of ICU admission were excluded from the analysis due to making the event unpredictable in the ICU. The outcomes of each study will be discussed more in detail in Section 5.6.

5.5. Databases

In this thesis, two databases were used to analyse postoperative complications following cardiac surgery. These databases are: (1) Cardiac and Thoracic Health Information (CaTHI) database, which stores information gathered at preoperative clinics at the GJNH; and (2) Centricity^{TM} CIS critical care database, which stores laboratory and vital sign information for patients in cardiac ICU. For the Study 1, the outcomes of the AKI and delirium were derived from the Centricity^{TM} CIS database, and the prediction was undertaken using the variables from the CaTHI database only. For the Study 2, the prediction was undertaken, using the variables from both databases.

5.5.1. Cardiac, Cardiology and Thoracic Health Information (CaTHI) Database

The CaTHI database was developed at the GJNH and is currently used in all three cardiac centres in Scotland: GJNH, Edinburgh Royal Infirmary and Aberdeen Royal Infirmary. As the name suggests, the database consists of cardiac, cardiology and thoracic patients' diagnostic assessments, surgical procedures, and discharge information. Data collected from preoperative clinics are stored in this database to calculate the logistic EuroSCORE for each patient for audit purposes.

Logistic EuroSCORE is one of the most commonly used cardiac preoperative risk stratification system [164]. The score was initially developed to predict 30-day mortality and is indeed used for that purpose in GJNH. The model uses a limited number of commonly available variables meaning that it can be conveniently implemented in a wide variety of clinical contexts [16].

5.5.1.1. Missing Data

Since CaTHI is a clinical audit database, most variables were consistently recorded. In cases where categorical variables had missing data, the blank fields were coded as "Unknown". The variables with "Unknown" entries included renal impairment (43.38%), rhythm (7.97%), smoking status (36.24%), and left main stem disease (48.76%). If a numerical variable was not recorded for less than 80% of the patients, the variable was excluded from the analysis. The only variable excluded for that reason was preoperative haemoglobin level.

Therefore, the final dataset used for the analysis consisted of 24 preoperative variables (Appendix 6.2), including patient characteristics, preoperative variables about patients' cardiac status and comorbidities, as well as other surgical variables.

5.5.2. Centricity^{TM} CIS Critical Care Database

The Centricity^{TM} CIS critical care database was developed by General Electric Company who develop various medical devices, data analytics applications and services [263].

The Centricity^{TM} CIS consists of eight databases, which are narrowed down into database tables. For the analysis in this thesis the database "Patient" was used. This database records all patient-specific values of medication (e.g., dose and rate), vital signs and laboratory results, together with timestamps for each new recorded value. Also, care notes are saved into this database together with the timestamp of when the note was recorded. The database content changes rapidly due to new patient data being continuously added both manually and automatically.

5.5.2.1. Description of Centricity^{TM} CIS Variables Used in This Thesis

Arterial Base Excess

The arterial base excess shows the acid-base balance in the blood. It is derived from blood pH (potential of hydrogen) and P_aCO_2 (partial pressure of carbon dioxide). It is defined as the amount of acid required to restore a litre of blood to its normal pH at a P_aCO_2 of 40mmHg (millimetres of mercury). The arterial base excess increases in case of a metabolic alkalosis, which is a condition in which the body fluids have excess base. The value decreases in case of a metabolic acidosis, which is a condition where there is too much acid in the body fluids [264]. A normal range for arterial base excess is -2 to +2 mEq/l (milliequivalents per litre). A value outside of the normal range can be caused by respiratory and/or metabolic problems [264].

Arterial Haematocrit

The arterial haematocrit is a measurement of the volume of red blood cells as a percentage of whole blood (red blood cells and plasma). As it is a percentage, it is expressed as a number without units between 0.00 and 1.00. Normal haematocrit levels vary based on age and sex. Normal levels for men can be between 40% and 54%. For women, it can be between 36% and 48%. If the haematocrit level is lower than normal, the person has not enough red blood cells [265].

Bicarbonate (HCO_3)

Bicarbonate (HCO_3) is a by-product of a body's metabolism. Blood takes bicarbonate to lungs, which is then exhaled as CO_2. Bicarbonate is also regulated by kidneys, which therefore regulates body's acid balance (pH), and is connected to sodium, potassium and chloride. Bicarbonate is tested to see if a person has a kidney disease, liver failure or other problems related to metabolism. A normal range for bicarbonate levels is between 23 to 30 mEq/l in adults. Higher than normal levels can cause a pH increase in tissue, which can happen due to vomiting and dehydration. A low level of bicarbonate can be caused by diarrhoea, kidney disease, and liver failure [266].

C-Reactive Protein

C-reactive protein is measured to determine whether there is inflammation in the body due to infection or to evaluate a person's risk of a heart attack. It is a protein synthesised by the liver, rising in response to inflammation. Some causes for high levels of c-reactive protein are acute and chronic conditions and trauma. Currently, there is no standard for what a normal level for c-reactive protein is, however, levels of less than 0.3 mg/l is considered normal [267].

Creatinine

Creatinine is a product resulting from breaking down creatine phosphate from muscle and protein metabolism, which is released at a constant rate by the body. It is the main indicator of kidney health, as it is removed from the blood mainly by the kidney through glomerular filtration. Creatinine levels can vary based on age and sex as it is directly linked to muscle mass. However, the normal range for creatinine is between 53 to 115 µmol/l (micromoles per litre). High creatinine levels are caused by insufficient filtration in the kidneys [268].

Daily Fluid Balance

Fluid balance describes the input and output of fluids in the body to allow metabolic processes to function correctly [269]. Daily fluid balance can be affected by injury or illness. Due to dehydration vital organs can have lower volumes of circulation. Fluid overload can occur because of poor cardiac or renal function. As the name suggests, the fluid balance is calculated based on how much fluid the patient gets orally or intravenously, and what is the urine and bowel output for the patient. Hence, the value can be positive in case of urine retention and negative in case of dehydration due to, for example, diarrhoea [270].

Haemoglobin

Haemoglobin is the protein contained in red blood cells that delivers oxygen to the tissues of the body. The haemoglobin levels that are considered to be normal are 140 to 180 g/l (grams per litre) for men and 120 to 160 g/l for women. When the haemoglobin level is low, the patient has anaemia [265].

Hydrogen Ion

Hydrogen ions are produced continuously through body's metabolic processes and are excreted through the kidneys. Hydrogen ion is inversely related to blood pH, which measures blood's acidity or alkalinity. Higher hydrogen ion results in lower blood pH. Lower than normal hydrogen ion levels can be because of respiratory disturbance, such as mechanical ventilation, which is common in ICU. Higher than normal levels can indicate respiratory problems, such as pneumonia, metabolic disturbance, such as renal failure, or gastrointestinal issues, such as diarrhoea [271]. The normal range for hydrogen ion for humans is between 44 and 36 nmol/l (nanomoles per litre) [272].

Lactate

High lactate levels are most used as a marker for septic complications (e.g., sepsis and septic shock), but also for cardiac complications (e.g., cardiogenic, obstructive, haemorrhagic shock, and cardiac arrest), respiratory complications (severe lung disease, respiratory failure, pulmonary oedema) or any trauma. A normal blood lactate level is considered to be between 0.5 to 1 mmol/l. For patients with critical illness normal lactate range is considered to be less than 2 mmol/l [273].

Potassium

Normal potassium levels are considered to be between 3.5 to 5.0 mmol/l. To maintain normal levels of potassium, kidneys flush excess potassium out of the body. Hence, abnormal levels of potassium can indicate kidney disease, heart problems, but also gastrointestinal complications (e.g., vomiting and/or diarrhoea). High potassium levels can reduce heart muscle activity [274].

Sodium

Sodium plays an important part at maintaining healthy blood pressure and regulating body's fluid balance. Normal sodium levels are considered to be between 135 and 145 mEq/l. Lower than normal sodium levels can be caused by congestive heart failure, kidney complications or liver complications. High sodium levels are often caused by dehydration [275].

Urea

Urea is measured to estimate kidney function together with creatinine. Blood urea nitrogen to creatinine can indicate kidney problems early, for example, urea concentration being high compared to creatinine can indicate a prerenal problem. Normal urea levels can be between 2.2 to 7.2 mmol/l [276].

Urine Output

Urine output is used as one of the markers of acute kidney injury as decreased urine output can be associated to lower glomerular filtration rate and hence decrease in kidney function. Decreased urine output is also considered to mirror a decrease in creatinine clearance, which also indicates kidney function [277].

Medications

Dobutamine is a medication used to treat cardiogenic shock and severe heart failure.

Dopamine is a vasopressor agent used in hypotensive patients.

Noradrenaline, also known as norepinephrine, is a vasopressor agent used to manage septic shock.

Vasopressin is a vasopressor agent, also used to manage septic shock.

5.5.2.2. Missing Data

The problem of missing data was approached in several ways.

Firstly, all data were checked for obviously incorrect values based on the literature and with the guidance of the clinical supervisor. Some laboratory values in the ICU system were impossible for patients to have. This might have happened due to errors in the equipment recording the values or manual data entry errors. Hence, the following laboratory variables were adjusted as shown in Table 5.2.

Table 5.2. Laboratory values recorded in the ICU database, their normal values and ranges that were marked as NA.

Variable	Normal values	Range	Marked as
Arterial Haematocrit	40% to 54%	<=0 and >100	NA
Bicarbonate (HCO3)	23 to 30 mEq/l	=0	NA
Haemoglobin	140 to 180 g/l	=0 and >1000	NA
Hydrogen Ion	44 to 36 nmol/l	=0 and >100	NA
Lactate	0.5 to 1 mmol/l	=0	NA
Potassium	3.5 to 5.0 mmol/l	=0	NA
Sodium	135 to 145 mEq/l	=0 and >1000	NA
Urine Output	Patient-dependent	>10000	NA

The Table 5.3 shows the completeness of each variable. If a patient had a timestamp recorded for a missing value of a variable, then the previously recorded value was carried forward to the next timestamp. The maximum sequence length of where the values were carried forward was 23 for the hydrogen ion variable. The table also shows the number of complete data points and the number of measurements carried forward to the next timestamp to replace NAs.

The only variable recorded for all ICU patients was haemoglobin, with 100% completeness. Creatinine was recorded for almost all patients with 98.32% completeness. Noradrenaline, dobutamine, dopamine and vasopressin were recorded for less than 41% of the patients. This is understandable as these four variables are medicines, meaning that not all patients would require them. Hence, to use the informative missingness [127], medicine variables were marked as “Yes” if a patient had it recorded for them and “No” if not. This is because, if left as numerical variable, the variation would be too small for prediction models to pick up on the differences.

Table 5.3. Completeness of each laboratory variable, number of complete data points and how many data points were carried forward.

Laboratory variable (Unit)	Proportion of Patients Recorded For	Number of complete data points	Number of data points carried forward
Arterial Base Excess (mmol/L)	70.04%	170890	13
Arterial Haematocrit (%)	70.02%	160826	29
Bicarbonate(mEq/L)	68.86%	167193	16
C-Reactive Protein (µmol/L)	69.79%	20632	146
Creatinine (µmol/L)	98.32%	30211	97
Daily Fluid Balance	63.59%	33159	0
Haemoglobin (g/L)	100.00%	236528	30
Hydrogen Ion (mmol/L)	69.21%	168947	2251
Lactate (mmol/L)	69.21%	139059	347
Potassium (mmol/L)	70.04%	171166	31
Sodium (mmol/L)	70.04%	171164	29
Urea (mmol/L)	69.88%	21823	68
Urine (L per day)	99.70%	675746	0
Medicines			
Dobutamine (dose)	29.55%	85987	5558
Dopamine (dose)	5.75%	14939	1101
Noradrenaline (dose)	39.58%	114648	7454
Vasopressin (dose)	2.32%	9967	2254

For example, for hydrogen ion, NA occurred after mean of 1.60 hours (SD = 1.12). The maximum time the NA measurement was taken was after 6.42 hours since the previous measurement. The differences of time were even shorter between the missing values recorded for medicines. For example, for dobutamine, NA occurred after mean of 0.66 hours (SD = 0.98). The maximum time difference between the NA value and last recorded value was at 23.92 hours. While this happened for one patient only, this can be a limitation to the models developed in this thesis.

It is worth noting that while the treatment of missing values was discussed with clinical supervisor Prof Stefan Schraag, by carrying forward values an assumption is made that the patient’s status stays the same as the time changes. This can especially be a problem when carrying forward values that are usually recorded only once a day (e.g., urea, where NA appears in mean of 14.06 hours (SD = 10.64) and maximum NA occurred at 24.23 hours since the last value was recorded. While this can be a limitation to the data preparation in this thesis, the number of values carried forward for each laboratory value was very small (Table 5.3), and therefore it can be assumed that the effect of these carried forward values would have a minimal effect on the models' performance.

Also, while carrying forward is a common practice in dynamic prediction modelling, the reason as of why some measurements were not taken at certain times for certain patients is unknown. However, two speculations can be made: either there was a system error where the measurement was not taken correctly and hence the information was not entered into the database. Another reason could be that a patient was deemed to be generally well based on previous test results, and therefore the concept of "informative missingness" could be applied. As shown in the literature review (Chapter 2), this concept was applied by Huddar et al. [65], and should be further investigated in the future work.

In addition to the methods described here, experiments were undertaken in Study 2 to approach the problem of missing data, further described in Section 5.9.3.

5.5.2.3. Laboratory Value Measurement Frequency

The data shows (Table 5.4) that there were clear differences in the frequency of laboratory tests in ICU after 24 hours of admission. According to the clinical supervisor and database managers, the tests were done routinely for each patient in the first 24 hours and after 24 hours the tests were done more on ad-hoc basis. The median period in ICU was 22 (IQR = 24) hours, and mean was 48.46 (SD = 102.00) hours.

The Figures 5.3 to 5.5 reflect the Table 5.4, where the majority of the patients have the laboratory tests undertaken at similar frequency. Most patients have the tests for creatinine, urea and c-reactive protein between 20 to 30 hours since ICU admission (Figure 5.3).

The Figures 5.4 and 5.5 show that arterial base excess, arterial haematocrit, bicarbonate, haemoglobin, hydrogen ion, lactate, potassium and sodium are mostly measured in the early hours since ICU admission, more regularly than the laboratory values shown in Figure 5.3.

Table 5.4. Median hours of when each laboratory variable is recorded in general and based on whether the patient has been in the ICU for 24 hours or less vs more than 24 hours.

	All patients in general	**ICU hours ≤24h**	**ICU hours >24h**
Variable	**Median (IQR) hours**	**Median (IQR) hours**	**Median (IQR) hours**
Every 24 hours			
Creatinine	23.67 (17.22)	0.00 (11.03)*	24.08 (1.15)
Urea	23.68 (16.90)	0.00 (11.10)*	24.08 (1.15)
C-reactive protein	23.78 (16.92)	0.00 (10.34)*	24.13 (1.02)
Every 2-3 hours			
Arterial Base Excess	2.35 (2.67)	1.18 (1.67)	3.32 (2.23)
Arterial Haematocrit	2.38 (2.68)	1.23 (1.72)	3.35 (2.27)
Bicarbonate	2.35 (2.67)	1.17 (1.67)	3.33 (2.22)
Haemoglobin	2.35 (2.73)	1.18 (1.68)	3.35 (2.28)
Hydrogen Ion	2.32 (2.67)	1.17 (1.65)	3.28 (2.23)
Lactate	2.78 (2.30)	1.92 (1.67)	3.37 (2.22)
Potassium	2.37 (2.67)	1.18 (1.67)	3.32 (2.20)
Sodium	2.35 (2.67)	1.18 (1.67)	3.32 (2.22)
Depending on patient			
Urine	0.85 (1.10)	0.87 (1.07)	0.83 (1.12)
Daily Fluid Balance	5.32 (23.95)	0.00 (0.00)*	22.55 (24.13)

*Usually measured only once when admitted to ICU – hence median = 0.00 hours.

Figure 5.3. Time of measurement of creatinine, urea and c-reactive protein.

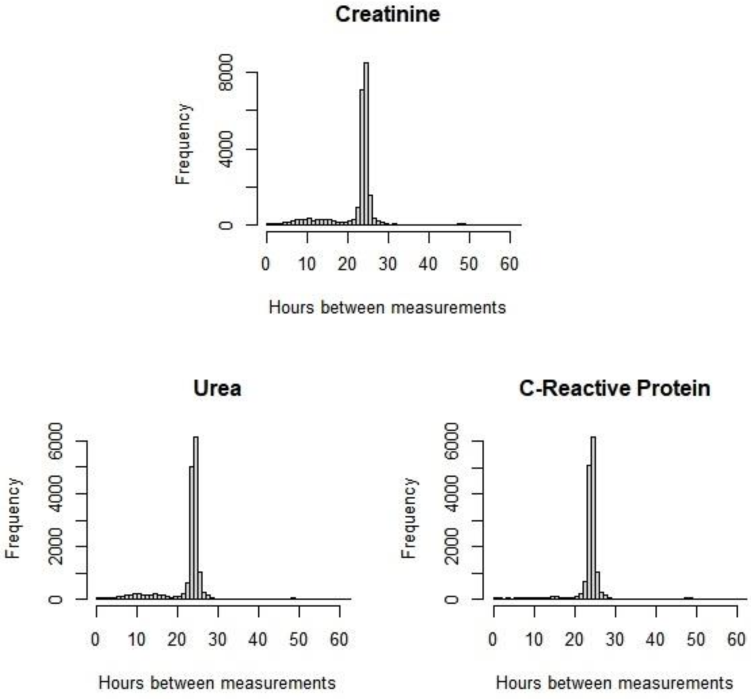


Figure 5.4. Time of measurement of arterial base excess, arterial haematocrit, bicarbonate and haemoglobin.

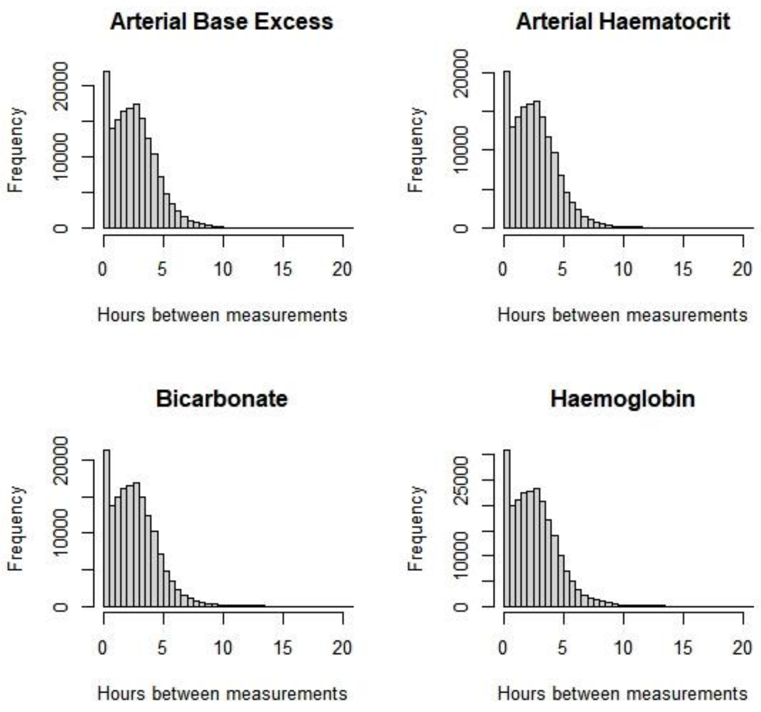
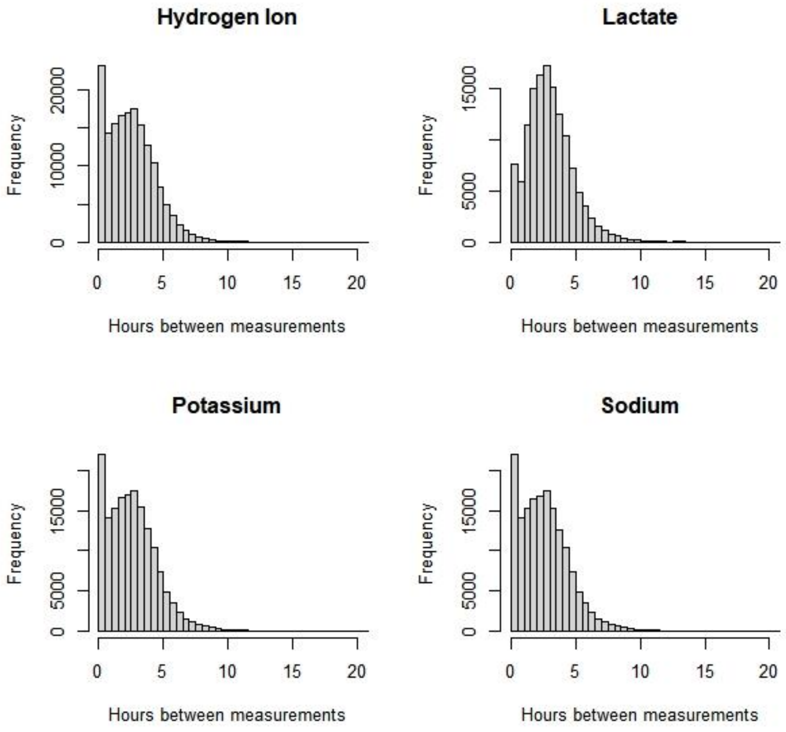


Figure 5.5. Time of measurement of hydrogen ion, lactate, potassium and sodium.



5.6. Predicted Outcomes of the Studies

For the preoperative models (Study 1), three outcomes are predicted:

- Whether a patient has severe postoperative complications
- Whether a patient has postoperative acute kidney injury
- Whether a patient has postoperative delirium.

For the hourly ICU prediction models (Study 2), two outcomes are predicted:

- Whether a patient has acute kidney injury within 25 hours since ICU admission
- Whether a patient has delirium within 21 hours since ICU admission

5.6.1. Severe Postoperative Complications

Since cardiac anaesthetists and surgeons expressed the need for a prediction model to predict combined complications (Chapter 3), a Delphi study was undertaken to define the complications and classify these. Since "severe" complications were agreed to have the biggest impact on patient's life, the amount of intervention required due to the complications, and a long-standing effect on the patient, this group of complications was decided to be predicted.

The exact definition for "severe" complications comes from Chapter 3, which presented the Delphi study, where postoperative complications following cardiac surgery were defined and classified. In the study, the experts, such as cardiac anaesthetists and clinicians working in cardiac ICU, reached a consensus that "severe" complications are defined as follows:

> *A severe complication following cardiac surgery is a complication that is potentially life-threatening, significant amount of intervention is required due to this complication, and there is a significant or long-standing impact on the patient.*

In addition to reporting overall surgical outcome (dead/alive) and total stay in hospital, the CaTHI database also records whether a patient had a complication following surgery. The reported complications were categorised into "mild", "moderate" and "severe" based on the results of the Delphi Study in Chapter 4. The complication categorisation within the CaTHI dataset resulted in 27 different "severe" complications. The list of complications and the prevalence of these can be found from

Appendix 6.1. In instances where patients had several complications from various levels of severity, those patients were assigned to the category that recorded the highest level.

The prediction of severe postoperative complications is a continuation of previous work in the author's MPhil, where postoperative complications following cardiac surgery were predicted [17]. The differences of the studies done in previous work can be found from Table 1.2 in Chapter 1. Part of the results of these prediction experiments have been published as conference proceedings [278].

It is important to note that the definition of "severe" complications is not based on laboratory variables, recorded in electronic health records, but is based on the list of complications reported in the CaTHI database. Hence, the time of occurrence of these "severe" complications is unknown, and therefore, the prediction of this outcome is undertaken only in Study 1, where only preoperative prediction models are developed.

5.6.2. Acute Kidney Injury

Acute kidney injury (AKI) is officially defined as *"an abrupt decrease in kidney function that includes, but is not limited to acute renal failure"* [106]. AKI is a broad term for various kidney problems, such as acute interstitial nephritis, acute glomerular and vasculitic renal disease, ischaemia, toxic injury, as well as prerenal azotaemia and acute postrenal obstructive nephropathy.

According to Kidney Disease Improving Global Outcomes (KDIGO) Clinical Practice Guideline for Acute Kidney Injury, AKI is defined as any of the following [106]:

- Increase in serum creatinine by ≥ 0.3 mg/dl (≥ 26.5 μmol/l) within 48 hours; or
- Increase in serum creatinine to ≥ 1.5 times baseline, which is known or presumed to have occurred within the prior 7 days; or
- Urine volume < 0.5 ml/kg/h for 6 hours.

In this thesis, AKI was defined using the baseline serum creatinine measurement, recorded in the CaTHI database as preoperative creatinine, and the subsequently recorded serum creatinine measurements in the Centricity^{TM} CIS database, in the ICU postoperatively, as done elsewhere [279]. A difference between each serum creatinine and preoperative creatinine measurement was calculated. If the difference was greater than or equal to 1.5 times the baseline, the patient was considered to have AKI. In addition, the timestamp when the creatinine difference occurred was recorded as AKI timestamp.

Using the KDIGO guideline to diagnose AKI based on the changes of creatinine levels compared to the creatinine level recorded in preoperative clinic is objective and reliable, and the method is internationally recognised [106]. In addition, as each value recorded in ICU has a timestamp, it is easy to assign the timestamp of when the creatinine change happened to assign to the AKI diagnosis, making it possible to develop a real-time prediction model for AKI in the ICU. Hence, this outcome will be predicted in both Study 1 and Study 2 in this thesis.

As shown in Chapter 6, according to CaTHI, 5.22% of the patients had a renal complication in this patient population. However, after applying the KDIGO guidelines, 18.93% of the patients had AKI. This very large difference in patients recorded to have renal complications shows that AKI is under-reported in the CaTHI database and using the KDIGO guidelines is the appropriate approach to diagnose AKI retrospectively.

Following cardiac surgery, up to 40% of patients can experience AKI, resulting in increased morbidity and mortality [280]. Patients who have postoperative AKI following cardiac surgery are at higher risk of postoperative infection, atrial fibrillation and prolonged stay in ICU and hospital [281].

There are various ways how clinicians have tried to prevent AKI in cardiac surgery. One attempt includes the development of off-pump CABG surgery, intended to be less damaging to kidneys, however, the effect of off-pump CABG for reducing AKI is inconclusive [281].

Known postoperative risk factors for AKI have shown to be haemodynamic instability[7], nephrotoxic[8], inotropic[9] and vasoconstrictor drugs[10] and systemic inflammation [282]. Other factors include postoperative anaemia, reduced cardiac output and sepsis.

5.6.3. Delirium

According to Oxford Dictionary, delirium is an *"acutely disturbed state of mind characterised by restlessness, illusions, and incoherence, occurring in intoxication, fever and other disorders"* [283]. In this thesis, delirium is defined, using the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU) diagnosis tool that has been reported in the Centricity^{TM} CIS database since 2016.

The CAM-ICU tool takes 2-3 minutes to use for a reliable result and requires little training. It is usable for clinicians without psychiatric training to monitor whether critically ill patients develop delirium [284]. The CAM-ICU tool is shown in Table 5.5.

The CAM-ICU assessment results are entered the Centricity^{TM} CIS database, meaning a timestamp for each assessment result is also recorded. Hence, as the time of the diagnosis is known, experiments predicting delirium are undertaken in both Study 1 and Study 2.

Delirium can affect up to 50% of hospital patients who are over the age of 65 years [285], which makes up the majority of cardiac patients (as shown in Chapter 6). Delirium is also relatively common in cardiac surgery patients, with the incidence between 26 to 52% [286]. Patients undergoing valve surgery are more likely to have postoperative delirium than patients undergoing CABG surgery. The risk is even higher for patients who require replacement of both the mitral and aortic valves [287].

[7] Haemodynamic instability is defined as one or more out-of-range vital sign measurements, such as low blood pressure [408].

[8] Nephrotoxic drugs are drugs that can have a significant damage on renal function [409].

[9] Inotropic agents, or inotropes, are medicines that change the force of the heart's contractions [410].

[10] Vasoconstrictor drugs contract the smooth muscle in blood vessels, causing the vessels to constrict. This helps to increase arterial blood pressure [410].

Table 5.5. The Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). Delirium is diagnosed when both Features 1 and 2 are positive, along with either Feature 3 or Feature 4. This assessment method was developed by Ely et al. [284].

Feature 1: Acute Onset of Mental Status Changes or Fluctuating Course	
	Is there evidence of an acute change in mental status from the baseline?
	Did the (abnormal) behaviour fluctuate during the past 24 hours, that is, tend to come and go or increase and decrease in severity?
	Sources of information: Serial Glasgow Coma Scale or sedation score ratings over 24 hours as well as readily available input from the patient's bedside critical care nurse or family.
Feature 2: Inattention	
	Did the patient have difficulty focusing attention?
	Is there a reduced ability to maintain and shift attention?
	Sources of information: Attention screening examinations by using either picture recognition or Vigilance A random letter test (further described in [284]). Neither of these tests require verbal response, and thus they are ideally suited for mechanically ventilated patients.
Feature 3: Disorganised Thinking	
	Was the patient's thinking disorganised or incoherent, such as rambling or irrelevant conversation, unclear or illogical flow of ideas, or unpredictable switching from subject to subject?
	Was the patient able to follow questions and commands throughout the assessment? "Are you having any unclear thinking?" "Hold up this many fingers." (Examiner holds two fingers in front of the patient) "Now, do the same thing with the other hand." (Not repeating the number of fingers)
Feature 4: Altered Level of Consciousness	
	Any level of consciousness other than "alert".
	Alert – normal, spontaneously fully aware of environment and interacts appropriately
	Vigilant – hyperalert
	Lethargic – drowsy but easily aroused, unaware of some elements in the environment, or not spontaneously interacting appropriately with the interviewer; becomes fully aware and appropriately interactive when prodded minimally
	Stupor – difficult to arouse, unaware of some or all elements in the environment or not spontaneously interacting with the interviewer; becomes incompletely aware and inappropriately interactive when prodded strongly
	Coma – unarousable, unaware of all elements in the environment, with no spontaneous interaction or awareness of the interviewer, so that the interview is difficult or impossible even with maximal prodding

Delirium is complex and has multiple factors causing it, and hence it has been shown that the most effective strategy to prevent delirium is risk factor analysis and predictive modelling, as explained by Inouye et al. [285]. On one hand, for an older people, especially for those with dementia and other underlying conditions, a single dose of sleeping medication may be enough to bring on delirium. On the other hand, for a young healthy patient, delirium usually develops only after a number of interventions, such as general anaesthesia, major surgery (e.g., open-heart surgery) and ICU stay [285].

A number of risk factors have been identified to be associated with delirium. Currently identified biggest risk factors are dementia or other cognitive impairment, functional impairment, vision impairment, history of alcohol abuse, and age > 70 years. In addition, a presence of a number of comorbidities and abnormal laboratory values (blood and urine tests) are also risk factors [285].

Delirium can have serious long-term effects on patients, such as permanent damage to cognitive ability. Due to potential functional decline, it can also lead to complications, such as infections or blood clots that weaken patients and increase the risk of mortality [288]. Delirium also has consequences, such as increased hospital length of stay [5], increased healthcare costs [5], and increased re-admission rates [289].

5.7. Descriptive Statistics Methods

Relevant descriptive statistics for each study were shown for preoperative and outcome data from CaTHI dataset (Appendix 6.2) and for laboratory data from the Centricity^{TM} CIS dataset (Appendices 7.1 and 8.1). For categorical variables, frequencies were calculated and shown in percentages. For numerical variables, mean and standard deviation were calculated for each variable. These procedures were undertaken for the total population as a whole, but also based on whether the patient had the predicted outcome, i.e., severe complication (explained in Section 5.6.1), acute kidney injury (based on KDIGO classification, explained in Section 5.6.2) or delirium (based on CAM-ICU score, explained in Section 5.6.3). The population that had the outcome was compared to the population that did not have the outcome based on descriptive statistics to understand whether there was a difference between the population with vs without the predicted outcome. The comparisons were made, using Pearson's chi-squared test of independence [290] for categorical variables and Welch's Two Sample t-test [291] for numerical variables.

For all tests in this thesis, a 95% significance level is used.

5.8. Study 1: Preoperative Prediction Model Methods

5.8.1. Classification Methods

As this is an imbalanced classification problem involving both categorical and numerical variables, various statistical and machine learning methods were used which have been shown to be effective for this kind of data analysis [292]: *logistic regression (LR), random forest (RF), naïve Bayes (NB), bagging classification and regression trees (BCART), support vector machine (SVM), AdaBoost (AB), gradient boosting model (GBM)* and two stacked models: *Stack RF* and *Stack generalised linear model (GLM).* All models were developed, using 10-fold cross-validation[11] in the training set, as is recommended in the literature [121].

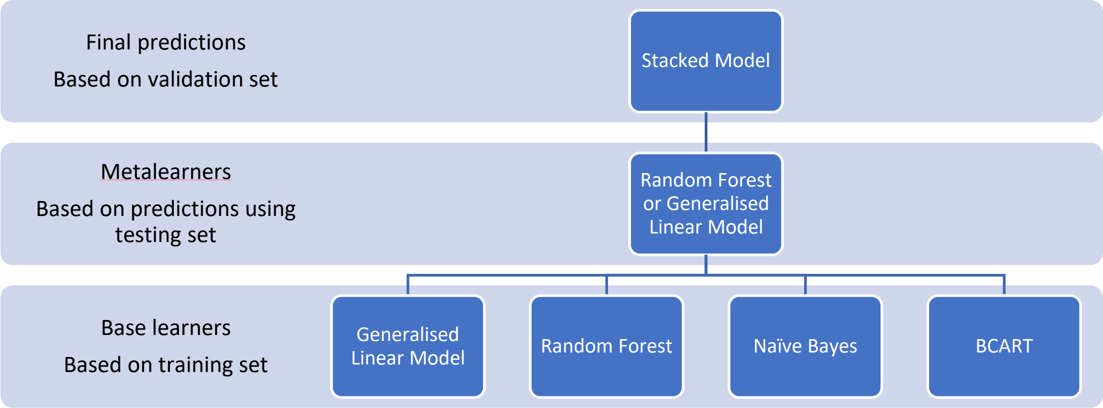
Logistic regression, random forest, naïve Bayes and BCART were developed, using the "*caret"* R package version 6.0.90 [293], with methods "glm", "rf", "naive_bayes" and "treebag", respectively. For support vector machine, the "*e1071"* R package version 1.7.9 [294] was used. For AdaBoost, the R package "*fastAdaboost*" version 1.0.0 [295] was used, and for gradient boosting, R package "*gbm*" version 2.1.8 [296] was used.

The hyperparameters were arrived at, using manual search [297], which means that various options were tested, and the parameters that produced the highest area under the receiver operating characteristic curve (AUC) were chosen.[12] The random forest model was developed, using 200 trees. For AdaBoost, 40 iterations were used, and for gradient boosting, the number of trees was set to 100, with the shrinkage of 0.01 and interaction depth of 4.

[11] The general code used for developing the prediction models for Study 1 can be found from DOI: 10.15129/9da23147-6be9-46f1-95be-6681ed2cc7e0.

[12] For random forest and gradient boosting, the optimal number of trees were found by testing 50 to 500 trees, increasing the number of trees by 50 tree increments. For AdaBoost, the minimum number of iterations tested was 10, and maximum was 100. The tests were made with 10 iteration increments.

Figure 5.6. Methods of how the stacked models were developed.



For the stacked models, the data were divided into training, testing, and validation sets (further explained in Section 5.8.1.1). As seen from Figure 5.6, the base learners (generalised linear model, random forest, naïve Bayes and BCART) were trained, using the training set. The predicted probabilities were calculated from the base learners, using the testing set. Two meta-learner algorithms were developed: one using a generalised linear model to make a prediction based on the predicted probabilities derived from the base learners, and the other using a random forest. These two stacked models were then evaluated, by finding the predicted probabilities based on validation set. All this analysis was undertaken, using the "*caret*" R package version 6.0.90 [293].

5.8.1.1. Training and Testing Datasets

The severe complications and acute kidney injury were analysed, using the data from 1st April 2012 to 31st December 2018 (6839 patient records, all unique patients). As delirium is recorded in the ICU at the GJNH since 2016, for this outcome, the data from 1st of January 2016 to 31st December 2018 is analysed (3344 patient records, all unique patients).

For models that are not stacked models, the respective datasets were divided into training (2/3 of data) and testing (1/3 of data) sets. The models were trained, using this training set. For stacked models, the respective datasets were divided into three:

training (1/2 of data), validating (1/4 of data) and testing (1/4 of data). The base-learners were trained, using the training data, and the stacked model was built, using the training, and validating set. The number of records in each training, testing and validating set for each outcome can be found from Table 5.6.

Table 5.6. Number of records in each dataset, based on predicted outcome and model type.

Outcome	Training (N)	Validating (N)	Testing (N)
Severe Complications			
Not stacked models	4583	-	2256
Stacked models	3420	1710	1709
Acute Kidney Injury			
Not stacked models	4583	-	2256
Stacked models	3420	1710	1709
Delirium			
Not stacked models	2241	-	1103
Stacked models	1672	836	836

5.8.1.2. Classification Experiments

Three experiments were undertaken in Study 1, using both original training data and upsampled training data.

Experiment 1: Predicting the outcome using variables that are significantly associated with the predicted outcome based on logistic regression. The number of these variables will depend on the outcome predicted, and hence will be described separately in Sections 6.2, 6.3 and 6.4 in Chapter 6.

Experiment 2: Predicting the outcome using variables that are included in the logistic EuroSCORE (15 variables). These variables are *age, sex, left ventricular function, extracardiac arteriopathy, previous myocardial infarction, angina status, active endocarditis, hypertension, pulmonary disease, neurological dysfunction, serum creatinine, previous cardiac surgery, type of surgery, critical preoperative state and surgical priority* [164].

Experiment 3: Predicting the outcome using all variables available preoperatively (24 variables). This is a classical machine learning approach, where all available variables are included in the model. These variables include patient characteristics, preoperative variables about patients' cardiac status and co-morbidities, as well as variables about surgery. More detail about the variables in the dataset can be found from Appendix

6.2. where also the descriptive statistics for the total patient population and for each outcome separately are presented.

5.8.1.3. Data Upsampling Experiments

As will become evident in Chapter 6, the instances of the predicted complications are far less frequent than the instances of no or non-severe complications, no AKI and no delirium. Class imbalance can generate classification difficulty due to the imbalanced class distributions, where some classes are highly underrepresented compared to other classes. This skewed distribution can make it difficult to develop ‘balanced’ predictive algorithms, which can both predict majority and minority class instances accurately [298].

One potential solution to address this problem is to use upsampling. In this chapter, the *upSample* function provided in R package *caret* [299] version 6.0.90 is used in order to obtain a balanced dataset in terms of the frequencies in each classification of the predicted outcomes. The *upSample* function randomly samples a data set so that all classes have the same frequency as the majority class. Simple random sampling is used, and all the original data are left intact, and additional samples are added to the minority class. Some samples are removed from the majority class with replacement [299]. The upsampling method is chosen, as opposed to widely popular SMOTE [130], due to the fact that SMOTE requires numerical variables only, and cannot handle categorical data, whereas upsampling can [299]. As shown in Appendix 6.2, all variables in this analysis are categorical variables.

Hence, in addition to developing prediction models for the outcomes using original training data, the models were also developed, using upsampled training data. The purpose of carrying out the experiments with balanced data also was to potentially improve the performance of the models in terms of AUC, sensitivity, and specificity.

Table 5.7 shows how the number of records in the training data changed through upsampling. As the prevalence of severe complications was noticeably lower than it was for the other outcomes, the number of added records through upsampling was higher.

The models' performance was evaluated, using testing data. It is important to note that for experiments using upsampling, the testing data were left as original to ensure applicability of the models in real-world scenarios, where the predicted outcomes occur considerably rarely. More information about models' performance evaluation is provided in Section 5.10.

Table 5.7. Details of the number of records changing through upsampling of the training data.

Outcome (Prevalence)	**Model Type**	**Original Training Data**	**Records Added Through Upsampling**	**Balanced Training Data**
Severe complications (5.91%)	Non-stacked	4583	4067	8650
	Stacked	3420	3044	6464
Acute kidney injury (18.93%)	Non-stacked	4583	2891	7474
	Stacked	3420	2174	5594
Delirium (12.47%)	Non-stacked	2241	949	3190
	Stacked	1672	1238	2910

5.9. Study 2: Hourly Prediction Model Methods

5.9.1. Predicted Outcomes of the Models

In the Study 2 acute kidney injury (AKI) and delirium are predicted on an hourly basis in the ICU. Both predicted outcomes were defined, as described in Sections 5.6.2 and 5.6.3.

In Study 2, the timestamps of when the predicted outcomes were recorded in the Centricity^{TM} CIS database were taken into account, enabling the development of hourly prediction models.

It is also important to note that in Study 2, the patients who had AKI or delirium in the first hour of ICU stay were removed from analysis. This was done, because it would have been impossible to predict an event that had already happened.

5.9.2. Data Preparation

5.9.2.1. Timestamps and Time Windows

Each laboratory value in the Centricity^{TM} CIS database was recorded with a timestamp attached to them. Each dataset corresponding to each laboratory value was stored separately as a .csv file. Each file was read into R. The timestamps for each recorded laboratory value were converted into time formats for R. The data were grouped by PatientID and then arranged in order of events based on the time of when the variable was recorded[13].

For the hourly prediction, time windows were used to firstly indicate the onset of the predicted outcome. Secondly, the time windows were used to develop prediction models for each time window before the event. In this thesis, the hourly prediction was undertaken for AKI within 25 hours of ICU stay and for delirium within 21 hours of ICU stay. These times were chosen based on the fact that it is common for patients to experience these two complications within these timeframes (see Chapter 7 and 8). For each these predicted outcomes, models were built for hourly lead times, based on the time windows. The lead times were chosen to be every hour from 1 to 24 hours ahead of AKI occurring within 25 hours since ICU admission and 1 to 13 hours ahead of delirium occurring within 21 hours since ICU admission (See chapters 7 and 8 for further explanation).

To assign time windows for the recorded laboratory values, a new column Time_Diff was created to calculate the difference of times when the laboratory value was measured based on each PatientID. This time difference was measured in hours.

A new column Time_Diff_cumulative was created to have the cumulative time difference for recorded values for each PatientID. This was done to generalise the hours of when measurements were taken to make the timestamps comparable.

This data file was then linked (left_join, *'dplyr'* package version 1.0.7 in R) with the PatientIDs present in the CaTHI dataset. All laboratory data files were linked with the

[13] General code of how the timestamps and time windows were achieved for Study 2 can be found from DOI: 10.15129/1ab360f7-0779-4cf3-8a9a-dae621892a51.

CaTHI dataset to ensure that the ICU patients that were analysed were undergoing the same surgeries. Also, this allowed for the inclusion of preoperative variables in the hourly models, as well as finding out about the outcome for each patient journey, including the hospital days, and whether they were alive or dead at the end.

To create time windows to start with the predictive modelling, a new column for each laboratory value dataset was created called Time_Window. This was created by rounding up the Time_Diff_cumulative column. This means that if the Time_Diff_cumulative = 15.67 hours, then the Time_Window = 16.

To predict the outcomes within a certain time window, only the entries within that certain Time_Window were included in that data. This means that if predicting AKI within 25 hours, only entries recorded up to the 25th Time Window were included.

In addition, for prediction of the outcomes in general, patients who had the predicted outcome recorded within the first hour since ICU admission were removed from analysis, as done in similar studies [67]. This is because the hourly prediction models are intended to be used in the ICU, and hence it is impossible to predict an outcome that has already happened. Hence, 545 patients who had AKI on admission to the ICU, and 22 patients who had delirium on admission to the ICU were excluded from the analysis for the Study 2.

5.9.2.2. Data Structure for Lead Times

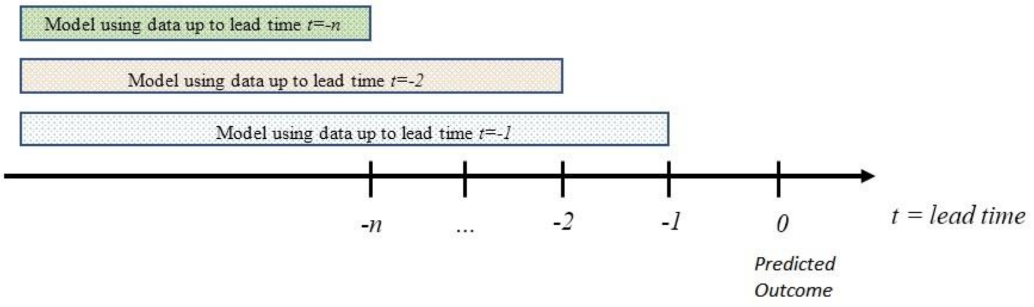
To simplify the data used in models, for each lead time the minimum, maximum, first and last measurement of a variable were used, helping to create a more consistent set of input data for the models, which might otherwise have had to deal with variations in the number of independent variables at each stage. This approach was also taken by Hug [300] and Johnson et al. [301] whose studies were found in the literature review (Chapter 2).This means that if the predicted outcome happened in time window = 6, for each variable first, last, min and max measurements that occurred in time windows 0-5 were calculated.

For example, if for a patient the predicted outcome happened in the 21st time window, if the prediction was made:

- 1 hour in advance, min, max, first and last measurement were calculated based on Time Window 0 to 20
- 2 hours in advance, min, max, first and last measurement were calculated based on data in Time_Window 0 to 19
- 10 hours in advance, min, max, first and last measurement were calculated based on data in Time_Window 0 to 11.
- And so on.

The Figure 5.7 shows how the models were developed for each lead time before the predicted outcome. The first model predicted the outcome 1 hour in advance and used all data that were collected until 1 hour before outcome. The second model predicted the outcome 2 hours in advance and used all data that were collected until 2 hours before the outcome. The nth model predicted the outcome n hours in advance and used all data that were collected until n hours before the outcome occurred.

Figure 5.7. Visualisation of how models were developed for each lead time.


Model using data up to lead time t=-n
Model using data up to lead time t=-2
Model using data up to lead time t=-1
-n
...
-2
-1
0
t = lead time
Predicted Outcome

The prediction models had a binary outcome (AKI = Yes/No or delirium = Yes/No), but only patients with AKI or delirium = Yes had a timestamp associated with the outcome was recorded for them. Hence, an arbitrary time as the end point was chosen for patients with AKI or delirium = No. Most patients had AKI between 20 and 25 hours and delirium between 10 and 14 hours since ICU admission. Hence, to cover all bases, an arbitrary end point of 25 hours for AKI prediction and 21 hours for delirium prediction was chosen.

5.9.3. Classification Experiments

The missing data where previously recorded values could not be carried forward to the next timestamp (Section 5.5.2.2) were approached in three different ways:

- **Experiment 1:** Using complete data only - In this approach all patients with missing data were removed from analysis. Using complete data allowed using different classification methods, such as logistic regression, random forest, AdaBoost, gradient boosting and support vector machine, BARTm and C5.0.
- **Experiment 2:** Using models that "handle" missing data – In this approach two machine learning methods – C5.0 and BARTmachine – were used to carry out the prediction of AKI and delirium. These methods were used as these can make a prediction, regardless of some patients having missing values in the data. Here, patients with more than 40% of missing variables were excluded from the analysis, as done elsewhere [300], [302].
- **Experiment 3:** Using imputation – In this approach missing values were replaced with other possible values. For usability, imputing 0 to replace missing values were tested, which is a similar approach to Pattalung et al.'s [79] approach, explained in Chapter 2. In addition, median and missForest imputation methods were experimented with. Here also, patients with more than 40% of missing variables were excluded from the analysis. The same classification methods were used as in Experiment 1.

Predicting AKI and delirium were imbalanced classification problems, meaning the number of patients with the outcomes was relatively small, compared to the number of patients without the outcome. Hence, as shown in Table 5.8, predictive modelling methods appropriate for this kind of data analysis were used for Experiment 1 and 3: *logistic regression*, *random forest*, *AdaBoost*, *gradient boosting model* and *support vector machine, BARTm* and *C5.0*. For Experiment 2 only *BARTm* and *C5.0* were used.

To develop the models, the datasets for each lead time were divided into training set (2/3 of data) and testing set (1/3 of data). For every experiment, the models were developed using the training data that did not have any missing values. The completeness of testing data depended on the experiment (Table 5.6). All models were developed on training data, using 10-fold cross validation[14], which is the recommended approach to developing a prediction model [261].

Table 5.8. Completeness of data and methods used for predicting AKI and delirium, based on experiment.

Experiment	Type of Data	Methods
Experiment 1	Complete data only	Logistic regression, random forest, AdaBoost, gradient boosting model, support vector machine, BARTm and C5.0
Experiment 2	Complete training data, missing values in testing data.	BARTm and C5.0
Experiment 3	Complete training data, imputation methods to replace missing values in testing data.	Logistic regression, random forest, AdaBoost, gradient boosting model, support vector machine, BARTm and C5.0

5.9.3.1. Classification Methods Used for Experiments 1 and 3

In this section, the classification methods used for Experiments 1 and 3 are described. These methods were selected due to being representative of a wide range of approaches that are appropriate for mixed data (numerical and categorical), and have also been shown to be appropriate for imbalanced classification tasks [303]. The hyperparameters were decided based on manual search, as done in Study 1 (Section 5.8.1).

Even though, as shown in Chapter 2, the use of neural networks is considerably common in developing dynamic prediction models in healthcare, this study did not use neural networks in its experiments. The main reason for this decision was the fact that patients stayed in the ICU for a relatively short time (just over two days, as shown in Chapter 7 and Chapter 8), and the laboratory variables included in the analysis are recorded less frequently than a deep neural network model would require for accurate prediction (resulting in an average of approximately 20 observations for 11 regularly measured variables per patient). The inclusion of more regularly recorded data and experiments with deep learning methods are further discussed in Future Work (Section 9.4).

[14] The general code for developing the classification models in Study 2 can be found from DOI: 10.15129/1ab360f7-0779-4cf3-8a9a-dae621892a51.

AdaBoost

AdaBoost is a boosting algorithm that repeatedly runs a given weak learning algorithm on various distributions over the training data, ultimately combining the weak classifiers into one classifier [304]. It has been shown to perform well by many studies.

The AdaBoost model was developed using the package *'fastAdaboost'* version 1.0.0 [295], which implements Freund and Schapire's AdaBoost.M1 algorithm [304], and for which n=40 iterations were conducted. AdaBoost.M1 algorithm is a binary classifier where the target variable is a factor with exactly two levels. The final model comprises of weak decision trees that are combined.

Gradient Boosting Model

Gradient boosting model has a similar approach to AdaBoost, where a weak learner is modified, i.e., boosted, to become a stronger learner. It works by adding weak learners one at a time to existing weak learners, resulting in a better-performing model. Because it is a tree-based model, it is also possible to understand the variable importance [305].

For the gradient boosting model, the package *'gbm'* version 2.1.5 [296] was used, which uses the Friedman's gradient boosting algorithm [306]. The number of trees was chosen to be n=100 and the shrinkage parameter as 0.01.

Logistic Regression

For logistic regression, a generalised linear model is used in this thesis. Logistic regression is a traditional statistical method used to develop predictive models for diagnosing or prognosing clinical outcomes [150]. Since the datasets in this thesis are not very large, and machine learning models are known to require more data than logistic regression [307], it is useful to see how the logistic regression models perform in comparison to different machine learning approaches. In addition, logistic regression is a highly explainable model though the ability to convert the model estimates to odds ratios. This helps clinicians to understand which variables are associated with the predicted outcome. For logistic regression, package *"caret"* R package version 6.0.90 with method "glm" was used [293].

Random Forest

Random forest is a popular machine learning algorithm, combining the output of multiple decision trees into single result. It is made of multiple decision trees, and therefore is an ensemble algorithm, which predicts more accurate results than individual decision trees. A clear benefit to using random forest is the ability to extract variable importance from the model. This means that it is easy to find out which variables are associated with the predicted outcome [308]. For random forest, the R package *"caret"* version 6.0.90 was used [293] with method "rf". The number of trees was set at n=200.

Support Vector Machine

Support vector machine is a machine learning method based that works by developing a hyperplane that separates observations into one class from another based on the variables in the data [309]. Even though support vector machine has been shown to be very successful at solving clinical big data classification problems, one drawback of the method is its mathematical complexity [310], meaning it is difficult to understand which variables are associated with the outcome. For support vector machine, the package *"e1071"* version 1.7.9 [294] was used.

5.9.3.2. Missingness of Data

There are some classification methods found in the literature that have been shown to handle missing values without imputation. Before carrying out any analysis dealing with missing data, it is important to understand why there are missing data. There are three types of missing data:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

For MCAR, the probability of missing values is the same for any variable. The missing values are independent from one another.

For MAR, there is a relationship between the missing values and some variable in the dataset. For example, emergency surgery patients might have less preoperative data available, and therefore have, for example, "Unknown" renal function.

For MNAR, happens when any variable causes omissions in data. For example, if height and weight are not recorded or are recorded incorrectly, the BMI value will be missing.

As we are looking at certain laboratory values that are recorded independently, we classify these as MCAR. When it comes to medicine data, these can be classified as MNAR as not all patients need the medication. However, the missing medicine data problem was solved by changing this into a categorical variable, i.e., does the patient have medication (Yes/No). For patients for whom the medication was not recorded, it was assumed that the patient did not get the medication.

5.9.3.3. Classification Methods Handling Missing Values

BARTm Classification

One of these methods is BARTmachine (Bayesian Additive Regression Trees), developed by Kapelner and Bleich [146]. This method has also been evaluated and compared with BART and Random Forest with imputation with missForest [311].

BARTm is a probabilistic approach to prediction which incorporates built-in estimates of uncertainty in the form of credible intervals as well as prior information on covariates. Kapelner and Bleich were the first to incorporate missing data into the BART algorithm. This was done by relying on the Metropolis-Hastings algorithm [312] which attempts to send missing data to whichever of the two daughter nodes increases the overall model likelihood. Hence, missingness becomes a valid splitting criterion. Their approach is applicable to continuous and nominal covariate data. As BART includes estimates of uncertainty, the amount of uncertainty increases with the amount of information lost due to missingness [146].

BART is a combination of many regression trees estimated via a Bayesian model. BART consists of a set of independent priors and likelihoods, with its posterior distribution estimated via Gibbs Sampling [313] and with a Metropolis-Hastings step [312]. There are three priors within BART which are designed to prevent overfitting:

1. The prior placed on the tree structure is designed to prevent trees from growing too deep, limiting the complexity that can be captured by a single tree.

2. The prior placed on the leaf value parameters, which are the predicted values found in the terminal nodes and is designed to shrink the leaf values towards the overall centre of the response's empirical distribution.
3. The prior is placed on the variance of the noise σ^2 and is designed to reduce overfitting by introducing noise into the model if it begins to fit too agreeably.

In machine learning, these priors can also be taught of as "tuning parameters" [146].

In addition to the BART's splitting attribute and split value, BARTm works by additionally offering a direction for records to be sent when the records have missing values. This means, if we have two options (e.g., left and right), then BARTm offers options for both of these options if a record has a missing value. This means that the prior on splitting rules is the same as the original BART, but with BARTm there is an additional consideration that the direction of missingness is equally likely to be left or right, conditional on the splitting attribute and value [146].

The authors say that the BARTm algorithm should yield better predictive performance than classical decision trees because of its ability to alter its trees by pruning and regrowing nodes or changing splitting rules [146]. It should be noted that because BARTm adjusts itself to predictor space for a location where the missing data would most increase the overall marginal likelihood, it therefore has great performance when dealing with missing data under MAR and NMAR conditions. However, when missingness does not depend on any other covariates, it can be more difficult to find appropriate ways to partition the missing data, and therefore BARTm can be less effective under MCAR condition [146].

In this thesis, to use this method, *"bartMachine"* package [314] version 1.2.6. in R was used.

C5.0 Classification

The C5.0 model is an extension of the C4.5 classification algorithm developed by Quinlan [315] in 1993. The decision trees like C4.5 and CART trees were among first algorithms which incorporate the handling of missing data into the algorithm itself. When a node is encountered, the decision tree tests a variable. If for that variable there is a missing value, then all outcomes are explored. Thus, for each possible sub-node a prediction is made. The distribution for each sub-node is kept and added. Finally, the class chosen for prediction is the class with the biggest density value [315].

As a summary, C5.0 algorithm is a tree-based algorithm that works by splitting the sample based on the field that provides maximum *information gain*. Each sub-sample generated based on the first split is then split again. This process continues to repeat until the sub-samples cannot be split any further. Then, the lower-level splits are re-examined, and those not significantly contributing to the outcome are removed.

In order to decide which feature to split upon, the algorithm uses entropy to calculate the change in homogeneity resulting from a split on each possible feature. This task is referred to as *information gain*. The information gain for a feature is calculated as the differences between the entropy in the segment before the split and the partitions resulting from the split:

As after a split, the data are divided into more than one partition, meaning the function to calculate entropy needs to consider the total entropy across all the partitions. This is accomplished by weighing each partition's entropy by the proportion of records falling into that partition. The higher the information gain, the better a feature is at creating homogenous groups after a split on that feature.

C5.0 algorithm is known to be quite robust when working with large number of data that also contains missing data. C5.0 algorithm, having similarly good performance to Neural Networks and Support Vector Machines, however, is easier to understand and interpret.

For the C5.0 model development in this thesis, R package *"C50"* [316] version 0.1.5. was used, together with the default of including missing values as the model can accommodate these.

5.9.3.4. Imputation Methods

In Experiment 3, to develop the prediction models, complete training data were used, but the models were evaluated, using testing data that had missing values replaced with three imputation methods: *median imputation*, *0 imputation*, and *missForest imputation*.

It was previously described (Section 5.5.2.2) that if a patient had a number of laboratory values recorded for them and had a missing value with an attached timestamp to it, the previous laboratory value was carried forward to the new timestamp with a missing value. The imputation methods, however, were used only for patients who did not have these laboratory values recorded for them at all, meaning that also no timestamps for these missing values were recorded. Hence, in this thesis, the missing laboratory variable summary statistics (minimum, maximum, first and last) were replaced with values, derived from the imputation methods. This was due to some laboratory variables not being measured for the patients, making it impossible to calculate the minimum, maximum, first and last value for them.

Similarly to Section 5.9.3.3, where methods handling missing data are described, the imputation methods can also be applied if the data are missing completely at random (MCAR).

Since the imputation methods are known to distort the variance of the data, patients with more than 40% of missing values were excluded from analysis, resulting in completeness ranging between 96.00% and 97.40% in testing datasets for each lead time for delirium prediction. This means that only up to 4% of the patients had missing data (Chapter 8, Section 8.5.1). The AKI prediction experiments, however, resulted in completeness between 60.66% and 63.00% in testing datasets, meaning that up to 40% of the patients had missing data (Chapter 7, Section 7.5.1). This means that the imputation methods can significantly change the variance in AKI data. Nevertheless, it is recommended that single imputation methods, such as median imputation, and in this case, 0 imputation, are only used for analysis, where only a small number of values are missing [317]. However, due to the simplicity of median and 0 imputation methods, and hence having a higher likelihood to be put into use in practice, these methods were decided to be experimented with.

Median Imputation

Median imputation is a single imputation method that replaces missing values, using the population median for the numerical predictor. This method has been shown to produce similar results to more sophisticated imputation methods, such as multiple imputation [116], [318].

The median imputation was chosen as most variables in this dataset were not normally distributed, making the median value more appropriate to replace the missing values with, than mean. There are three assumptions when median imputation is used:

1. The data are missing completely at random (MCAR).
2. The missing observations are likely to look like the majority of the observations in the variable (in this case median due to skewed distribution).
3. The missing values are most likely very close to the value of the median of the distribution.

As explained in Section 5.9.3.2, it can be assumed that the data for patients was missing completely at random, which satisfies the first assumption. In addition, most variables were not normally distributed, which means that most values lied close to the median value. This satisfies the second assumption. The third assumption is more difficult to be certain about, because the missing values might be key components that show why a patient was having a complication. Since patients with >40% of missing values were removed from analysis, the patients with the predicted outcomes had very low number of missing values, especially for delirium dataset (shown in Chapters 7 and 8). Hence, it can be assumed that for most patients who had missing values and had their data replaced with median imputation, the replaced values should be normal laboratory levels.

0 Imputation

The 0 imputation, being very similar to the median imputation, is a method in which the missing values are replaced with the value 0 for a numerical predictor. It was inspired by Pattalung et al.'s experiment [79], found in literature review (Chapter 2). This approach, not being a recognised imputation method, was experimented with due to its extreme simplicity, with a hypothesis that if this method is successful, it would be easy to apply into everyday practice. A development of a decision support tool where prediction of AKI or delirium can be made for individual patients, a replacement of missing values with the value 0 would be easy to understand for clinicians, and simple to develop. In addition, the calculation time would be fast as running sophisticated imputation methods would be unnecessary.

missForest imputation

The missForest imputation is a method, where missing values are directly predicted, using the random forest algorithm on the non-missing values in the dataset. It has also been shown to manage missingness rates up to 30%. Even though in this thesis only numerical variables were treated with imputation methods, unlike median imputation, missForest can also be applied for categorical missing values. In addition, the data handled by missForest can be mixed-type, non-parametric, and allows for non-linear effects [122].

The missForest has been shown to outperform other well-known advanced imputation algorithms, such as k-nearest neighbours, multivariate imputation by chained equations, and missingness pattern alternating lasso [122], [319].

To apply missForest onto the datasets used in this thesis, *'missForest'* R package version 1.4. [320] was used.

5.10. Models' Evaluation and Performance Measures

The models' performance measures were calculated using the testing set derived from each dataset. The models were evaluated based on the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). These measures were shown for each experiment in Study 1 and also for each model developed for each lead time in Study 2. The performance metrics were calculated, using the optimal cut-off points, where sensitivity and specificity were maximised. In addition, in Study 2, mean and standard deviation for each performance measure across all lead times were calculated, applicable for the outcome. This was done to understand and summarise the results in general.

The aim of this thesis is to develop a model that reaches the highest overall performance (AUC), sensitivity and specificity as possible. This is to ensure that the model recognises as many patients with the predicted outcome as possible (sensitivity) and as many patients without the predicted outcome as possible (specificity). It is also important to achieve a high NPV to ensure that the probability that the patient actually does not have the predicted outcome is high [196].

In addition to the models' discriminative ability, also calibration plots were assessed to understand the accuracy of risk estimates between the predicted and observed probabilities for the predicted outcome. Using calibration plots is recommended in this case as in this thesis the number of patients with the predicted outcomes is sufficient [321]. In addition to the plots, calibration is also assessed, using mean and standard deviation predicted probabilities, which are compared with the prevalence of the outcome. If the mean predicted probability was equal to the prevalence, the model estimated the risk of the outcome correctly. If the mean predicted probability was higher than the prevalence, the model overestimated the risk of the predicted outcome, and vice versa [321]. As the models developed in this thesis were only validated internally, in case of poor calibration, the models were not recalibrated as the average of predicted risks would match the event rate [321].

In addition to the above, to further understand the performance of the models, confusion matrices were derived for all Study 1 models and in Study 2 for models predicting AKI 12 hours in advance, delirium 13 hours in advance, and both AKI and delirium 1 hour in advance, using the optimal cut-off values.

Furthermore, because in clinical practice it is important to understand why the models give a certain probability to a patient to have the predicted outcome, variable importance for the models were presented, where possible. Variable importance was extracted from each model, based on the variable importance measures available for each algorithm, i.e., model coefficients for logistic regression, mean gini importance for random forest, proportion of times each variable is chosen for a splitting rule for BARTm, variable usage for C5.0, relative influence for gradient boosting. Due to the nature of the AdaBoost and support vector machine methods, variable importance for these algorithms was not available. Based on each algorithm's variable importance measures, the variables were ranked. For Study 2, because there were 81 variables in each model (counting the minimum, maximum, first and last values for each laboratory value), to summarise the variable importance, from each model developed for each time window, the top 20 variables were extracted. These were combined with all other models developed for each lead time, and then summarised based on the experiment.

Chapter 6. Study 1: Predicting Severe Postoperative Complications, Acute Kidney Injury and Delirium, using Preoperative Data

6.1. Introduction and Related Work

The 2011 National Confidential Enquiry into Patient Outcome and Death (NCEPOD) estimated that there are between 20,000-25,000 deaths among people undergoing a surgical procedure every year in the UK [322]. Approximately 80% of these deaths occur amongst a minority of 'high risk' patients, who make up approximately 10% of the overall surgical population. In addition to facing higher mortality rates, these patients also have increased risk of postoperative complications, and therefore require high levels of care and clinical resources before, during and after surgery [322].

Over the last two decades, an increasing number of hospitals have developed preoperative clinics and services [11] designed to triage patients well in advance of their surgery into 'low risk patients', suitable for day-care surgery, and 'high-risk patients', requiring additional management and admission as inpatients [323]. Data-driven risk scoring systems are now an integral component of these surgical pre-assessment clinics, and most of these generally focus specifically on predicting patients' risks of mortality [14].

According to the 2021 Blue Book published by the Society for Cardiothoracic Surgery, the in-hospital mortality rate after cardiac surgery has remained low: i.e., 2.8% [218], [324]. Although surgical mortality rates are decreasing [218], the increasingly comorbid patient population [218] has resulted with complications after surgery being common, as shown in previous work [17]. Postoperative complications can have an important impact on patients' quality of life [8], [9] and can also increase hospital length of stay [4], [139], [325] and healthcare costs [10], [326], [327]. Hence, a robust and reliable predictive model for postoperative complications would prove extremely useful for managing patient flows and clinical resources in surgical care. Therefore, to explore the feasibility of such prediction model, various methods were experimented with to predict severe postoperative complications, acute kidney injury (AKI) and delirium.

Based on the author's previous work as part of MPhil (see Table 1.2 in Chapter 1), there are currently no validated surgical risk prediction models available which can predict combined surgical complications and their severity [17]. As shown by two earlier systematic reviews [14], [16], the commonly used risk prediction models (such as Initial Parsonnet Score [214], Cleveland Clinic Score [328], Society of Thoracic Surgeons Score [329], EuroSCORE [171] and EuroSCORE II [193]) have been mainly developed to predict mortality. As found in the literature review, undertaken in author's MPhil [17], these common preoperative prediction models have also been evaluated at predicting a combination of complications.

However, all studies had a different set of complications, without justification for why exactly these particular complications were chosen. A recent systematic review [330] investigating preoperative risk prediction models for complications after cardiac surgery did not find any newer models that were not included in the review, conducted in author's MPhil, however, found also that the heterogeneity of predicted outcomes makes objective comparison of models impossible. Hence, in this chapter, the combined complications as a predicted outcome were defined based on the Delphi Study (Chapter 4), offering a more objectively defined outcome that is based on the consensus of cardiac anaesthetists and intensivists.

The commonly used prediction models have also been evaluated at predicting various individual complications [17], including prolonged ICU stay, kidney complications, prolonged ventilation and deep sternal wound infection. However, as stated in author's MPhil thesis [17], the models were found to have a slightly lower performance when predicting a combination of complications, as opposed to individual complications.

It was identified in Chapter 3 that in addition to predicting a set of complications, the prediction of acute kidney injury (AKI) and delirium would be beneficial. Although these two complications have internationally recognised diagnosis criteria (as shown in Section 5.6 in Chapter 5), recent developments in preoperative prediction models predicting these two outcomes are lacking. Only one model was found to predict AKI [279], and one to predict delirium [331], in cardiac surgical patients, using preoperatively available data only.

Birnie et al. [279] developed a preoperative risk prediction model for postoperative AKI in cardiac patients, using very large development and validation datasets (total n=30,854). Rudolph et al. [331] developed a preoperative prediction model for postoperative delirium, using the data of 231 patients.

While these models predicting AKI and delirium achieved moderately high performance (AUC = 0.74 for Birnie et al. and AUC = 0.75 for Rudolph et al.), like the aforementioned well-known preoperative risk prediction models, these were also developed, using logistic regression. While logistic regression is easy to understand and has been shown to be competitive with machine learning methods [150], [332], other approaches should be experimented with to find out whether other methods can improve the prediction of severe postoperative complications, postoperative AKI and delirium.

Hence, in this chapter, in addition to logistic regression, other methods are experimented with, including tree-based, boosting and ensemble methods with the aim to achieve a better performance than previously developed prediction models. Furthermore, as the percentage of patients with severe postoperative complications was relatively small compared to no or other complications, this chapter also faced an imbalanced classification problem, which is one of the biggest challenges in prediction modelling due to its presence in many real-world classification tasks [333]. Hence, upsampling was also explored to find out whether it aids with developing a better performing prediction model. This approach was not experimented with in the other studies.

This chapter aims to answer the following questions:

- Which classification method has the best performance when predicting outcomes following cardiac surgery, based on preoperative data?
- What is the optimal number of preoperative variables for predicting the outcome?
- How do balancing methods affect the performance of preoperative models' performance?

The full information about the methods used in this chapter, including description of databases, description of variables, definition of outcomes, methods for descriptive statistics, classification methods and performance measures, can be found from Chapter 5.

6.2. Study 1.1 Results: Predicting Severe Postoperative Complications Based on Preoperative Data

6.2.1. Variables Associated with Severe Postoperative Complications

Based on the CaTHI database, out of 6839 patients, the prevalence of severe postoperative complications overall in this patient population was 5.91% (95% CI 5.37-6.49%). The descriptive statistics of how the patient population was spread amongst variables in the dataset can be found from Appendix 6.2, where the patients with severe complications were also compared with patients without severe complications, using Chi-Squared Tests of Independence.

Based on this, patients with severe complications were significantly different ($p<0.05$) from patients without severe complications based on age category, sex, whether they had type II diabetes, what surgical procedure they had, their surgical priority, whether they were in a critical preoperative state, and whether they had had a cardiac surgery or a percutaneous coronary intervention (PCI) before. There were also significant differences within patients with or without severe complications in terms of whether they had extracardiac arteriopathy, left ventricular (LV) function, New York Heart Association (NYHA) grade, angina status, rhythm of the heart, renal function, preoperative creatinine levels, congestive cardiac failure, and active endocarditis.

The Table 6.1 shows which variables were associated with severe postoperative complications based on unadjusted and adjusted odds ratios. Following covariate adjustment, twelve variables were associated with a patient having severe complications. Namely, patients who were 75 or older were 1.54 (95% CI 1.10-2.15) times more likely to have severe complications than patients who were 60 years old or younger. Female patients were 1.37 (95% CI 1.08-1.75) times more likely to have severe complications than male patients. With type II diabetes, patients were 1.52 (95% CI 1.20-1.90) times more likely to have severe complications than patients without type II diabetes. Procedures-wise, patients were more likely to have severe complications if they had combined CABG and valve surgery (OR = 1.56, 95% CI 1.14-2.12). With preoperative creatinine levels of 100µmol/l or higher, patients were 1.63 (95% CI 1.26-2.10) times more likely to have severe complications than patients with creatinine levels of less than 100µmol/l. Patients were also more likely to have severe complications if they had an emergency surgery (OR = 5.19, 95% CI 2.79-9.40), if they had had a previous cardiac surgery (OR = 3.22, 95% CI 2.12 - 4.80), NYHA grade level IV (OR = 1.81, 95% CI 1.10 - 2.95), and active endocarditis (OR = 3.15, 95% CI 1.65 - 5.83). Also having abnormal rhythm (OR = 1.42, 95% CI 1.06 - 1.88) and congestive cardiac failure (OR = 1.45, 95% CI 1.07 - 1.95) increased the likelihood of having severe postoperative complications. Interestingly, patients with angina status level I were less likely to have severe postoperative complications than patients with level 0 (OR = 0.59, 95% CI 0.39 - 0.87). This could be due to the fact that angina status is a very subjective measure, where the patients are categorised based on their ability to carry out any physical movement based on a clinician's observation [334].

Table 6.1. Variables associated with severe postoperative complications, based on unadjusted and adjusted odds ratios. Only variables that are significant based on unadjusted odds ratios are included in this table.

Variable	Level	Unadjusted OR (95% CI)	P-value	Adjusted OR (95% CI)	P-value
Age Category	60 or under	1.00		1.00	
	61 to 67	1.17 (0.85 - 1.59)	0.3360	1.23 (0.88 - 1.72)	0.2143
	68 to 74	1.18 (0.87 - 1.59)	0.2800	1.04 (0.75 - 1.45)	0.8194
	75 to 99	1.93 (1.47 - 2.55)	<0.0001	1.54 (1.10 - 2.15)	0.0119
Sex	Male	1.00		1.00	
	Female	1.49 (1.20 - 1.84)	0.0002	1.37 (1.08 - 1.75)	0.0104
Type II Diabetes	No	1.00		1.00	
	Yes	1.49 (1.20 - 1.84)	0.0003	1.52 (1.20 - 1.90)	0.0004
Procedure	CABG	1.00		1.00	
	Valve	2.05 (1.63 - 2.57)	<0.0001	1.24 (0.90 - 1.70)	0.1920
	CABG and Valve	2.26 (1.72 - 2.96)	<0.0001	1.56 (1.14 - 2.12)	0.0054
Renal Function Before Surgery	Normal	1.00		1.00	
	Moderately Impaired	1.64 (1.27 - 2.10)	0.0001	1.12 (0.84 - 1.49)	0.4554
	Severely Impaired	3.20 (2.31 - 4.40)	<0.0001	1.26 (0.83 - 1.89)	0.2810
	Unknown	1.15 (0.86 - 1.53)	0.3280	0.90 (0.65 - 1.24)	0.5194
Preoperative Creatinine	<100µmol/l	1.00		1.00	
	100 or higher	2.33 (1.89 - 2.86)	<0.0001	1.63 (1.26 - 2.10)	0.0002
Priority	Elective	1.00		1.00	
	Emergency	8.32 (4.91 - 13.72)	<0.0001	5.19 (2.79 - 9.40)	<0.0001
	Priority	0.89 (0.65 - 1.20)	0.4498	0.83 (0.60 - 1.13)	0.2516
	Urgent	1.54 (1.20 - 1.96)	0.0005	0.93 (0.68 - 1.26)	0.6311
Critical Pre-op. State	No	1.00		1.00	
	Yes	3.86 (2.42 - 5.94)	<0.0001	1.12 (0.62 - 1.94)	0.7046
Previous Cardiac Surgery	No	1.00		1.00	
	Yes	4.49 (3.10 - 6.37)	<0.0001	3.22 (2.12 - 4.80)	<0.0001
Extracardiac Arteriopathy	No	1.00		1.00	
	Yes	1.48 (1.12 - 1.94)	0.0050	1.32 (0.98 - 1.76)	0.0610
Left Ventricular Function	Good	1.00		1.00	
	Moderate	1.40 (1.08 - 1.78)	0.0081	1.13 (0.86 - 1.47)	0.3686
	Poor	2.56 (1.62 - 3.88)	<0.0001	1.61 (0.97 - 2.59)	0.0557
NYHA Grade	I	1.00		1.00	
	II	1.06 (0.78 - 1.45)	0.7330	0.95 (0.69 - 1.32)	0.7534
	III	2.05 (1.51 - 2.82)	<0.0001	1.34 (0.96 - 1.89)	0.0947
	IV	5.30 (3.49 - 8.03)	<0.0001	1.81 (1.10 - 2.95)	0.0189

Variable	Level	Unadjusted OR (95% CI)	P-value	Adjusted OR (95% CI)	P-value
Angina Status	0	1.00		1.00	
	I	0.43 (0.29 - 0.62)	<0.0001	0.59 (0.39 - 0.87)	0.0103
	II	0.61 (0.48 - 0.78)	0.0001	1.00 (0.74 - 1.35)	0.9950
	III	0.70 (0.51 - 0.95)	0.0260	1.04 (0.71 - 1.51)	0.8281
	IV	1.44 (0.98 - 2.07)	0.0538	1.58 (0.98 - 2.52)	0.0575
Active Endocarditis	No	1.00		1.00	
	Yes	5.73 (3.33 - 9.45)	<0.0001	3.15 (1.65 - 5.83)	0.0004
Rhythm	Normal	1.00		1.00	
	Abnormal	2.23 (1.72 - 2.87)	<0.0001	1.42 (1.06 - 1.88)	0.0155
	Unknown	1.45 (0.96 - 2.12)	0.0665	1.13 (0.72 - 1.70)	0.5842
Congestive Cardiac Failure	No	1.00		1.00	
	Yes	3.13 (2.45 - 3.97)	<0.0001	1.45 (1.07 - 1.95)	0.0161

6.2.2. Models Predicting Severe Postoperative Complications

In this section the results of the three experiments are reported:

> **Experiment 1:** Predicting severe postoperative complications, using only variables that are significantly associated with the outcome, based on adjusted odds ratios. In this case these variables are *age, sex, type II diabetes, surgical procedure, preoperative creatinine, surgical priority, previous cardiac surgery, NYHA grade, angina status, active endocarditis, rhythm and congestive cardiac failure* (12 variables).
>
> **Experiment 2:** Predicting severe postoperative complications, using only variables that are part of the logistic EuroSCORE (15 variables). These variables are listed in Section 5.8.1.2.
>
> **Experiment 3:** Predicting severe postoperative complications, using all preoperative variables in the CaTHI database. These variables are listed in the Appendix 6.2.

The patient demographics in training and testing data are shown in Appendix 6.3.

6.2.2.1. Models' Discrimination

As seen from Table 6.2, random forest using all preoperative variables had the highest AUC of 0.713. Based on sensitivity, stacked model with generalised linear model using all preoperative variables and logistic EuroSCORE variables achieved both the highest performance (Sens = 0.772 for both). In terms of specificity, support vector machine with logistic EuroSCORE variables had the highest performance (Spec = 0.931). Because the prevalence of severe postoperative complications was very low (5.91%) all models had very low positive predictive values. Hence, as expected, negative predictive values were relatively high, stacked model with generalised linear model developed with all preoperative variables having the highest (NPV = 0.909).

Looking at the models' performance after upsampling the training data (Table 6.3), AdaBoost with 24 variables had the highest AUC of 0.706. Random forest with 24 variables had the highest sensitivity of 0.781. Stacked model with generalised linear model using 15 variables had the highest specificity of 0.889. Similarly to the models developed with original training data, here also the positive predictive values were very low due to low prevalence of severe postoperative complications. Negative predictive values, however, were quite high, BCART with all preoperative variables having the highest (NPV = 0.912).

Table 6.2. Models' performance predicting severe postoperative complications with the optimal cut-off points where the sensitivity (Sens), specificity (Spec), positive and negative predictive values (PPV and NPV) were calculated from. For each performance measure, 95% confidence intervals (CI) are shown. The highest result for each performance measure is marked in bold.

Model	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Experiment 1 (12 variables)						
AB	0.650 (0.573 - 0.727)	0.712 (0.639 - 0.785)	0.513 (0.432 - 0.594)	0.037 (0.006 - 0.068)	0.908 (0.861 - 0.955)	0.134
BCART	0.659 (0.582 - 0.736)	0.377 (0.298 - 0.456)	0.901 (0.853 - 0.949)	0.046 (0.012 - 0.080)	0.791 (0.725 - 0.857)	0.120
GBM	0.675 (0.599 - 0.751)	0.678 (0.602 - 0.754)	0.601 (0.522 - 0.680)	0.036 (0.006 - 0.066)	0.895 (0.845 - 0.945)	0.048
LR	0.668 (0.592 - 0.744)	0.603 (0.524 - 0.682)	0.678 (0.602 - 0.754)	0.039 (0.008 - 0.070)	0.885 (0.833 - 0.937)	0.051
NB	0.662 (0.585 - 0.739)	0.486 (0.405 - 0.567)	0.795 (0.730 - 0.860)	0.043 (0.010 - 0.076)	0.859 (0.803 - 0.915)	0.001
RF	0.654 (0.577 - 0.731)	0.445 (0.364 - 0.526)	0.815 (0.752 - 0.878)	0.045 (0.011 - 0.079)	0.857 (0.800 - 0.914)	0.005
Stack GLM	0.667 (0.591 - 0.743)	0.634 (0.556 - 0.712)	0.634 (0.556 - 0.712)	0.035 (0.005 - 0.065)	0.902 (0.854 - 0.950)	0.100
Stack RF	0.671 (0.595 - 0.747)	0.515 (0.434 - 0.596)	0.786 (0.719 - 0.853)	0.037 (0.006 - 0.068)	0.869 (0.814 - 0.924)	0.068
SVM	0.534 (0.453 - 0.615)	0.404 (0.324 - 0.484)	0.746 (0.675 - 0.817)	**0.052 (0.016 - 0.088)**	0.901 (0.853 - 0.949)	0.057
Experiment 2 (15 variables)						
AB	0.648 (0.571 - 0.725)	0.651 (0.574 - 0.728)	0.584 (0.504 - 0.664)	0.040 (0.008 - 0.072)	0.902 (0.854 - 0.950)	0.161
BCART	0.645 (0.567 - 0.723)	0.534 (0.453 - 0.615)	0.734 (0.662 - 0.806)	0.042 (0.009 - 0.075)	0.878 (0.825 - 0.931)	0.001
GBM	0.668 (0.592 - 0.744)	0.534 (0.453 - 0.615)	0.754 (0.684 - 0.824)	0.041 (0.009 - 0.073)	0.869 (0.814 - 0.924)	0.072
LR	0.662 (0.585 - 0.739)	0.596 (0.516 - 0.676)	0.689 (0.614 - 0.764)	0.039 (0.008 - 0.070)	0.883 (0.831 - 0.935)	0.055
NB	0.643 (0.565 - 0.721)	0.473 (0.392 - 0.554)	0.763 (0.694 - 0.832)	0.046 (0.012 - 0.080)	0.879 (0.826 - 0.932)	0.001
RF	0.659 (0.582 - 0.736)	0.507 (0.426 - 0.588)	0.768 (0.700 - 0.836)	0.043 (0.010 - 0.076)	0.869 (0.814 - 0.924)	0.010
Stack GLM	0.682 (0.606 - 0.758)	**0.772 (0.704 - 0.840)**	0.549 (0.468 - 0.630)	0.025 (0.000 - 0.050)	0.903 (0.855 - 0.951)	0.060
Stack RF	0.648 (0.571 - 0.725)	0.644 (0.566 - 0.722)	0.608 (0.529 - 0.687)	0.036 (0.006 - 0.066)	0.907 (0.860 - 0.954)	0.052
SVM	0.568 (0.488 - 0.648)	0.247 (0.177 - 0.317)	**0.931 (0.890 - 0.972)**	0.053 (0.017 - 0.089)	0.801 (0.736 - 0.866)	0.060
Experiment 3 (24 variables)						
AB	0.691 (0.616 - 0.766)	0.658 (0.581 - 0.735)	0.617 (0.538 - 0.696)	0.037 (0.006 - 0.068)	0.894 (0.844 - 0.944)	0.183
BCART	0.704 (0.630 - 0.778)	0.575 (0.495 - 0.655)	0.732 (0.660 - 0.804)	0.039 (0.008 - 0.070)	0.871 (0.817 - 0.925)	0.080

Model	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
GBM	0.680 (0.604 - 0.756)	0.500 (0.419 - 0.581)	0.775 (0.707 - 0.843)	0.043 (0.010 - 0.076)	0.867 (0.812 - 0.922)	0.073
LR	0.670 (0.594 - 0.746)	0.500 (0.419 - 0.581)	0.779 (0.712 - 0.846)	0.043 (0.010 - 0.076)	0.865 (0.810 - 0.920)	0.071
NB	0.663 (0.586 - 0.740)	0.425 (0.345 - 0.505)	0.825 (0.763 - 0.887)	0.046 (0.012 - 0.080)	0.856 (0.799 - 0.913)	0.001
RF	**0.713 (0.640 - 0.786)**	0.562 (0.482 - 0.642)	0.748 (0.678 - 0.818)	0.039 (0.008 - 0.070)	0.866 (0.811 - 0.921)	0.085
Stack GLM	0.685 (0.610 - 0.760)	**0.772 (0.704 - 0.840)**	0.516 (0.435 - 0.597)	0.027 (0.001 - 0.053)	**0.909 (0.862 - 0.956)**	0.066
Stack RF	0.681 (0.605 - 0.757)	0.356 (0.278 - 0.434)	0.912 (0.866 - 0.958)	0.042 (0.009 - 0.075)	0.798 (0.733 - 0.863)	0.060
SVM	0.607 (0.528 - 0.686)	0.507 (0.426 - 0.588)	0.687 (0.612 - 0.762)	0.047 (0.013 - 0.081)	0.899 (0.850 - 0.948)	0.059

Table 6.3. Models predicting severe postoperative complications, using upsampling in training data, with the optimal cut-off points where the sensitivity (Sens), specificity (Spec), positive and negative predictive values (PPV and NPV) were calculated from. For each performance measure, 95% confidence intervals (CI) are shown. The highest result for each performance measure is marked in bold.

Model	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Experiment 1 (12 variables)						
AB	0.661 (0.584 - 0.738)	0.678 (0.602 - 0.754)	0.568 (0.488 - 0.648)	0.038 (0.007 - 0.069)	0.902 (0.854 - 0.950)	0.130
BCART	0.640 (0.562 - 0.718)	0.534 (0.453 - 0.615)	0.729 (0.657 - 0.801)	0.042 (0.009 - 0.075)	0.880 (0.827 - 0.933)	0.080
GBM	0.662 (0.585 - 0.739)	0.493 (0.412 - 0.574)	0.769 (0.701 - 0.837)	0.044 (0.011 - 0.077)	0.871 (0.817 - 0.925)	0.537
LR	0.663 (0.586 - 0.740)	0.589 (0.509 - 0.669)	0.669 (0.593 - 0.745)	0.041 (0.009 - 0.073)	0.890 (0.839 - 0.941)	0.109
NB	0.663 (0.586 - 0.740)	0.500 (0.419 - 0.581)	0.768 (0.700 - 0.836)	0.043 (0.010 - 0.076)	0.870 (0.815 - 0.925)	0.109
RF	0.659 (0.582 - 0.736)	0.589 (0.509 - 0.669)	0.663 (0.586 - 0.740)	0.041 (0.009 - 0.073)	0.892 (0.842 - 0.942)	0.240
Stack GLM	0.673 (0.597 - 0.749)	0.579 (0.499 - 0.659)	0.728 (0.656 - 0.800)	0.040 (0.008 - 0.072)	0.868 (0.813 - 0.923)	0.070
Stack RF	0.644 (0.566 - 0.722)	0.404 (0.324 - 0.484)	0.807 (0.743 - 0.871)	**0.050 (0.015 - 0.085)**	0.870 (0.815 - 0.925)	0.096
SVM	0.654 (0.577 - 0.731)	0.705 (0.631 - 0.779)	0.559 (0.478 - 0.640)	0.035 (0.005 - 0.065)	0.900 (0.851 - 0.949)	0.385
Experiment 2 (15 variables)						
AB	0.648 (0.571 - 0.725)	0.767 (0.698 - 0.836)	0.475 (0.394 - 0.556)	0.033 (0.004 - 0.062)	0.908 (0.861 - 0.955)	0.066
BCART	0.632 (0.554 - 0.710)	0.493 (0.412 - 0.574)	0.742 (0.671 - 0.813)	0.045 (0.011 - 0.079)	0.883 (0.831 - 0.935)	0.120
GBM	0.647 (0.569 - 0.725)	0.493 (0.412 - 0.574)	0.739 (0.668 - 0.810)	0.045 (0.011 - 0.079)	0.884 (0.832 - 0.936)	0.537
LR	0.660 (0.583 - 0.737)	0.445 (0.364 - 0.526)	0.814 (0.751 - 0.877)	0.045 (0.011 - 0.079)	0.858 (0.801 - 0.915)	0.602
NB	0.658 (0.581 - 0.735)	0.527 (0.446 - 0.608)	0.740 (0.669 - 0.811)	0.042 (0.009 - 0.075)	0.877 (0.824 - 0.930)	0.071
RF	0.639 (0.561 - 0.717)	0.705 (0.631 - 0.779)	0.502 (0.421 - 0.583)	0.039 (0.008 - 0.070)	0.911 (0.865 - 0.957)	0.120
Stack GLM	0.661 (0.584 - 0.738)	0.351 (0.274 - 0.428)	**0.889 (0.838 - 0.940)**	0.050 (0.015 - 0.085)	0.816 (0.753 - 0.879)	0.055
Stack RF	0.652 (0.575 - 0.729)	0.553 (0.472 - 0.634)	0.692 (0.617 - 0.767)	0.044 (0.011 - 0.077)	0.886 (0.834 - 0.938)	0.118
SVM	0.658 (0.581 - 0.735)	0.651 (0.574 - 0.728)	0.596 (0.516 - 0.676)	0.039 (0.008 - 0.070)	0.900 (0.851 - 0.949)	0.406
Experiment 3 (24 variables)						
AB	**0.706 (0.632 - 0.780)**	0.534 (0.453 - 0.615)	0.756 (0.686 - 0.826)	0.041 (0.009 - 0.073)	0.868 (0.813 - 0.923)	0.166
BCART	0.651 (0.574 - 0.728)	0.747 (0.676 - 0.818)	0.467 (0.386 - 0.548)	0.036 (0.006 - 0.066)	**0.912 (0.866 - 0.958)**	0.040

Model	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
GBM	0.653 (0.576 - 0.730)	0.397 (0.318 - 0.476)	0.836 (0.776 - 0.896)	0.048 (0.013 - 0.083)	0.857 (0.800 - 0.914)	0.549
LR	0.662 (0.585 - 0.739)	0.616 (0.537 - 0.695)	0.633 (0.555 - 0.711)	0.040 (0.008 - 0.072)	0.896 (0.846 - 0.946)	0.491
NB	0.677 (0.601 - 0.753)	0.479 (0.398 - 0.560)	0.818 (0.755 - 0.881)	0.042 (0.009 - 0.075)	0.846 (0.787 - 0.905)	0.154
RF	0.689 (0.614 - 0.764)	**0.781 (0.714 - 0.848)**	0.482 (0.401 - 0.563)	0.030 (0.002 - 0.058)	0.905 (0.857 - 0.953)	0.065
Stack GLM	0.675 (0.599 - 0.751)	0.658 (0.581 - 0.735)	0.651 (0.574 - 0.728)	0.036 (0.006 - 0.066)	0.881 (0.828 - 0.934)	0.093
Stack RF	0.665 (0.588 - 0.742)	0.482 (0.401 - 0.563)	0.752 (0.682 - 0.822)	0.047 (0.013 - 0.081)	0.878 (0.825 - 0.931)	0.052
SVM	0.663 (0.586 - 0.740)	0.534 (0.453 - 0.615)	0.723 (0.650 - 0.796)	0.043 (0.010 - 0.076)	0.882 (0.830 - 0.934)	0.557

The Figure 6.1 shows that in general the performance, as the number of variables changed, stayed quite similar. When using the original training data, support vector machine's performance was visibly the lowest, regardless of the number of variables used. Apart from the support vector machine, the AUC varied very little, the maximum being 0.713 (random forest, 24 variables, original data) and minimum of 0.632 (BCART, 15 variables, upsampled data).

The Figure 6.2 shows the sensitivity and specificity for each model and experiment. Even though the models' performance did not seem to change drastically as the number of variables changed, there seemed to be quite high variability of sensitivity and specificity across models. The sensitivity ranged from 0.781 (random forest, 24 variables, upsampled data) to 0.247 (support vector machine 15 variables, original data). The specificity ranged from 0.931 (support vector machine, 15 variables, original data) to 0.467 (BCART, 24 variables, upsampled data). Visually it was quite clear that models with very high specificity tended to have lower sensitivity and vice versa.

Figure 6.3 confirms that the PPV for all models, regardless of experiment, were very low, which was expected due to low prevalence of severe complications. Conversely, NPV for all models was considerably high and similar across models and experiments, the highest being 0.912 (BCART, 24 variables, upsampled data) and lowest being 0.791 (BCART, 12 variables, original data).

Figure 6.2. Area Under the Receiver Operating Characteristic Curve for each model, based on the number of variables used and whether upsampling was used.

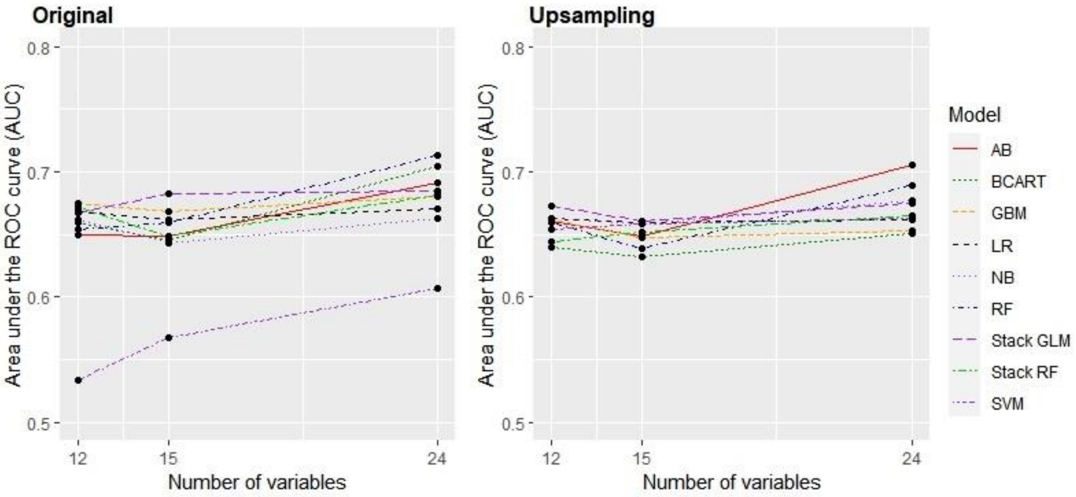


Figure 6.2. Sensitivity and specificity for each model based on the number of variables and whether upsampling was used.

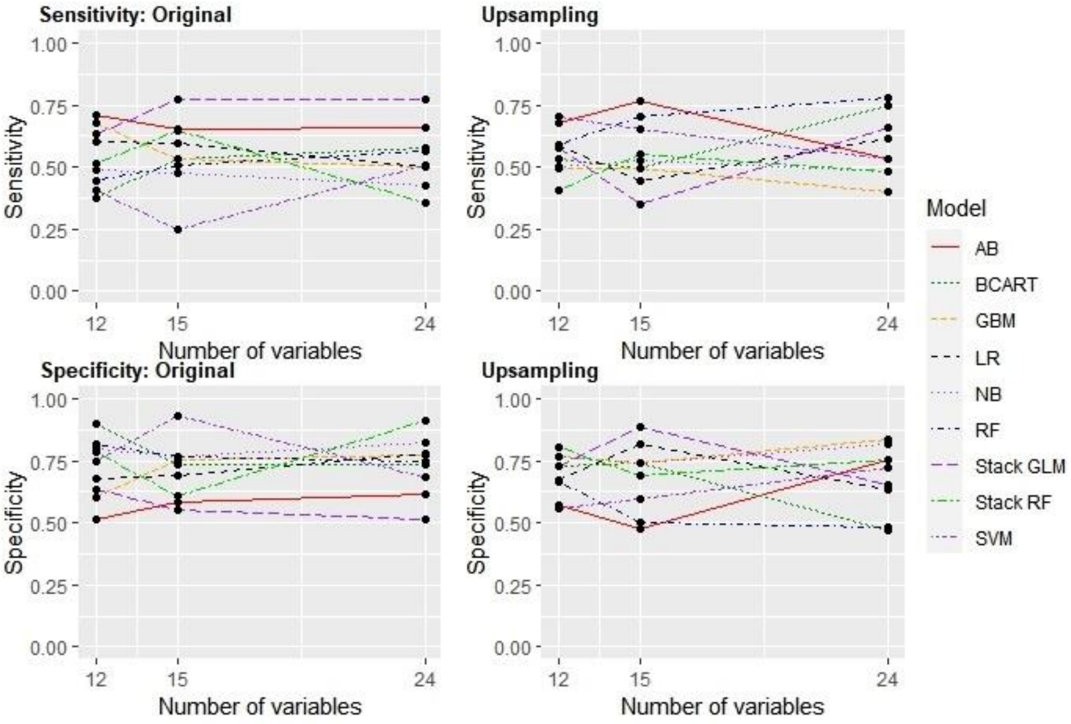
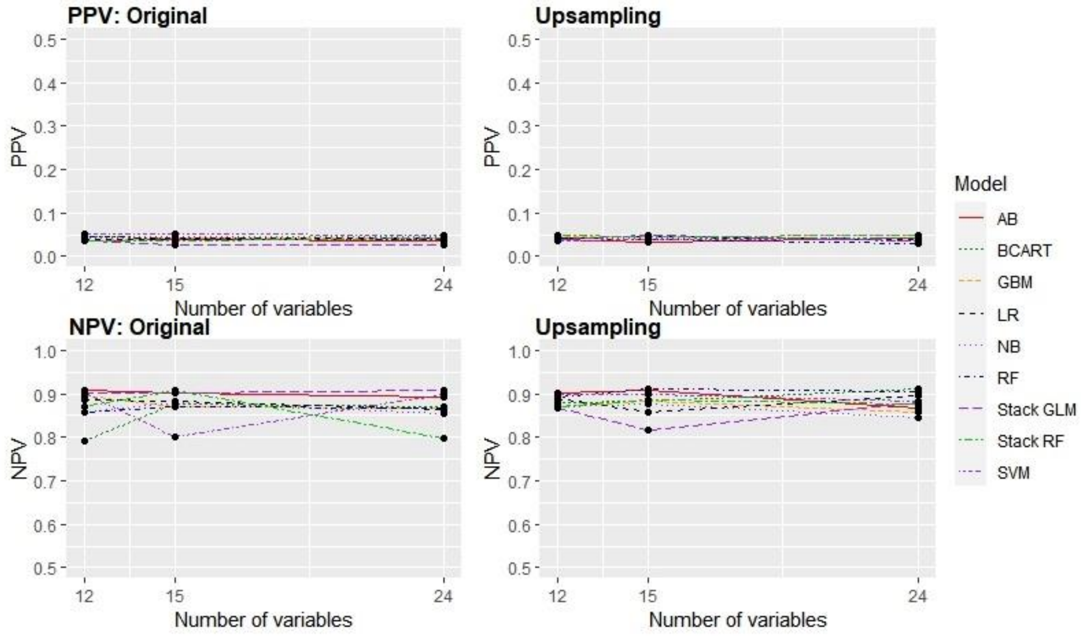


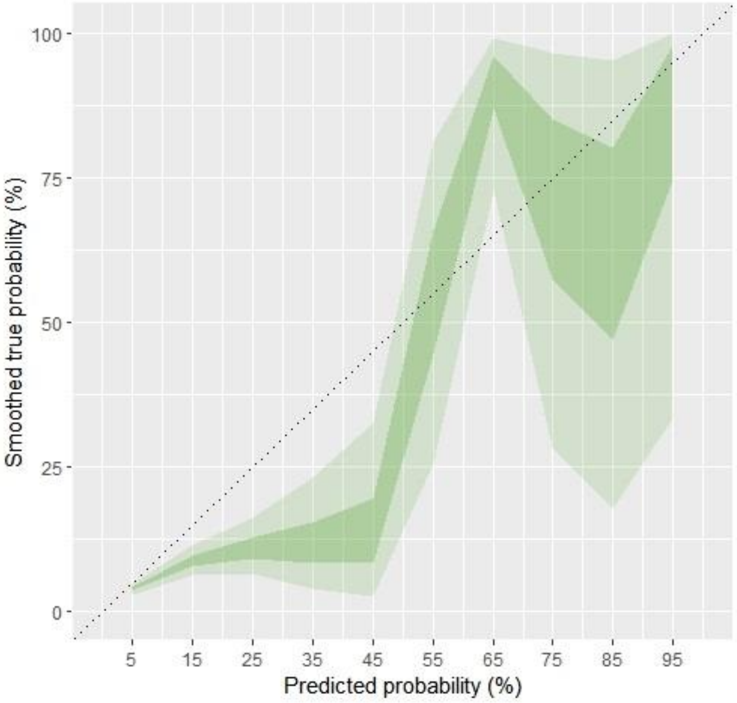
Figure 6.3. Positive and negative predictive values for each model based on the number of variables and whether upsampling was used.



6.2.2.2. Calibration and Variable Importance of the Best Performing Model

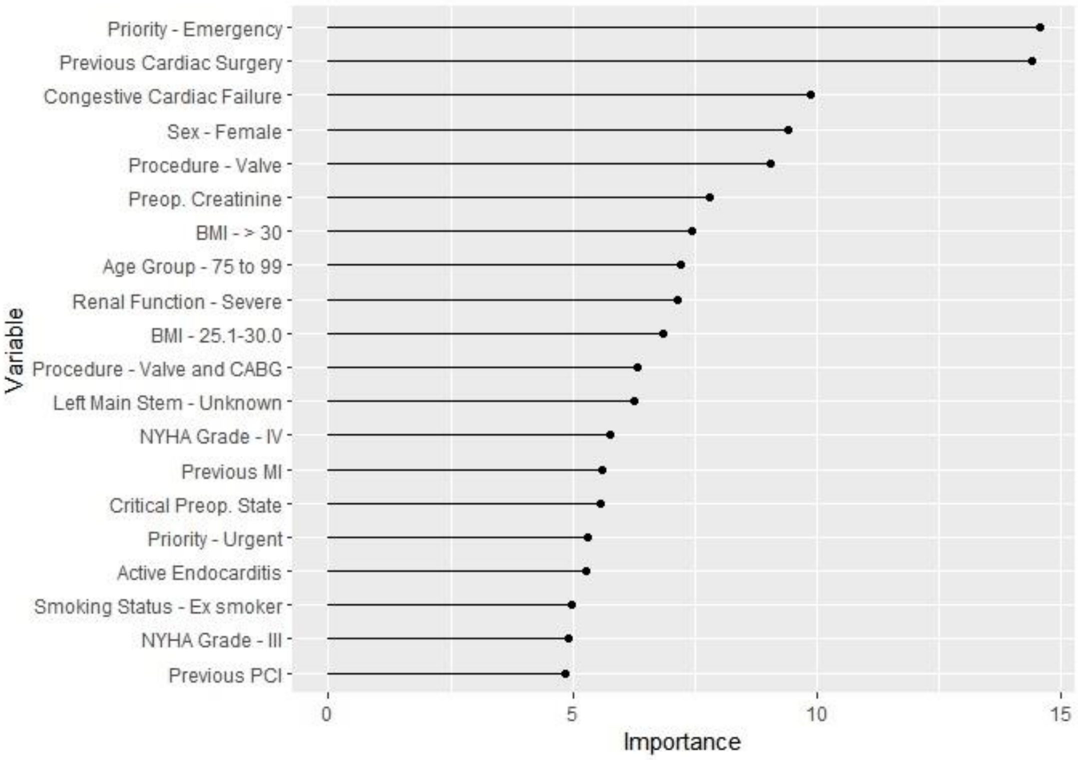
Since the random forest model using 24 preoperative variables and original data had the best discriminative performance, calibration for this model was assessed. The calibration plot (Figure 6.4) shows that for patients who had low true probability, the predicted probability for these people was higher. This means that for these patients the model overestimated their risk of having severe complications. For patients who had high true probability, however, the predicted probability was closer to the true probability, but in most cases stayed slightly on the lower side. This means that the model was better at estimating the risk for patients who had higher true probability for severe complications. The mean predicted probability for this model was 7.74% (SD = 11.26%), which is slightly higher than the prevalence of 5.91% severe complications. This means that the model in general overestimated patients’ risk for severe postoperative complications. In addition, the high standard deviation reflects the uncertainty about the estimated predicted probabilities, as also seen in Figure 6.4.

Figure 6.4. Calibration of the random forest model predicting “Severe” complications using 24 preoperatively available variables and original training data. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.



Since the random forest using 24 variables and original training data had the highest performance based on AUC, variable importance for this model is shown in Figure 6.5. The top 3 variables were surgical priority, previous cardiac surgery and congestive cardiac failure, all of which were also significantly associated with severe postoperative complications based on logistic regression. Interestingly, the random forest model also deemed BMI (top 7) and preoperative renal function (top 9) to be important, whereas based on adjusted odds ratios, these two variables were not significantly associated with severe complications. Due to the nature of the random forest model, the model coefficients for random forest were not available.

Figure 6.5. Variable importance for random forest using 24 variables and original training data.



Looking at the confusion matrices, (Figures 6.6 to 6.8), in general the models were not very good at classifying patients to have severe complications correctly. In terms of correctly classifying positive cases for severe complications, random forest with all preoperative variables and upsampled training data had the highest percentage classified correctly (78%).

Figure 6.6. Confusion matrices for models predicting severe postoperative complications, using only significant variables associated with the predicted outcome, using original training data (A) and upsampled training data (B).

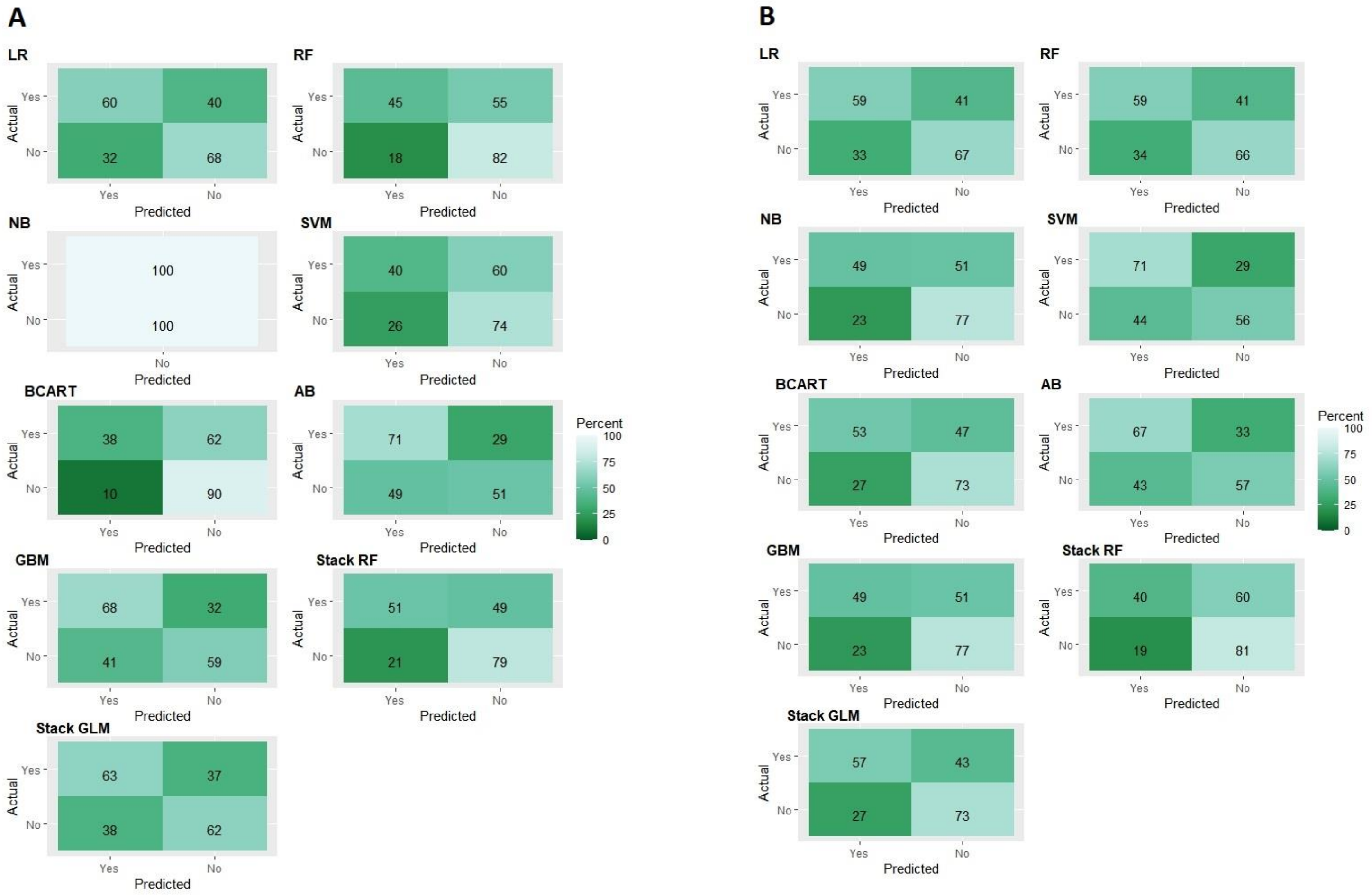


Figure 6.7. Confusion matrices for models predicting severe postoperative complications, using only logistic EuroSCORE variables, developed with original training data (A) and upsampled training data (B).

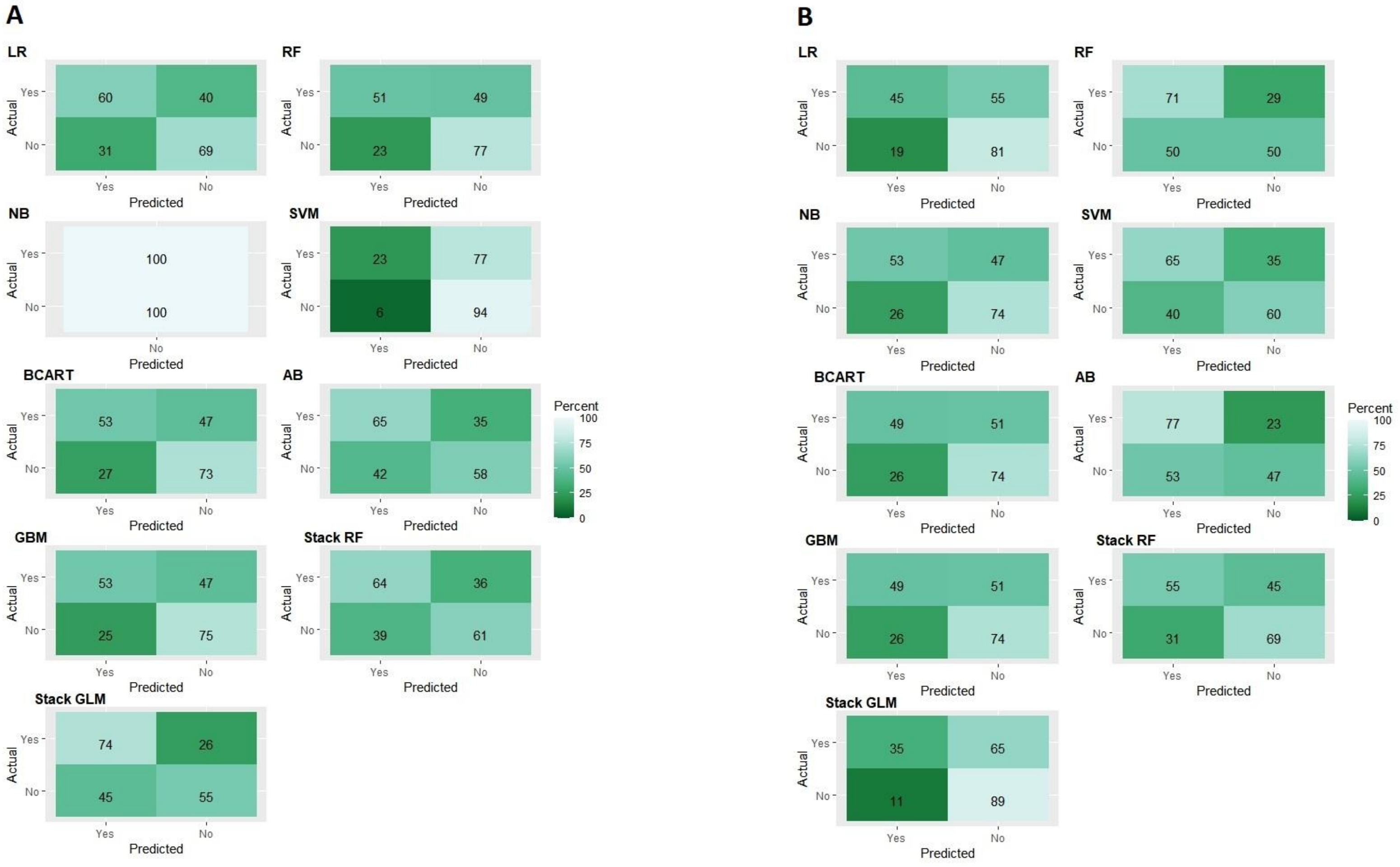
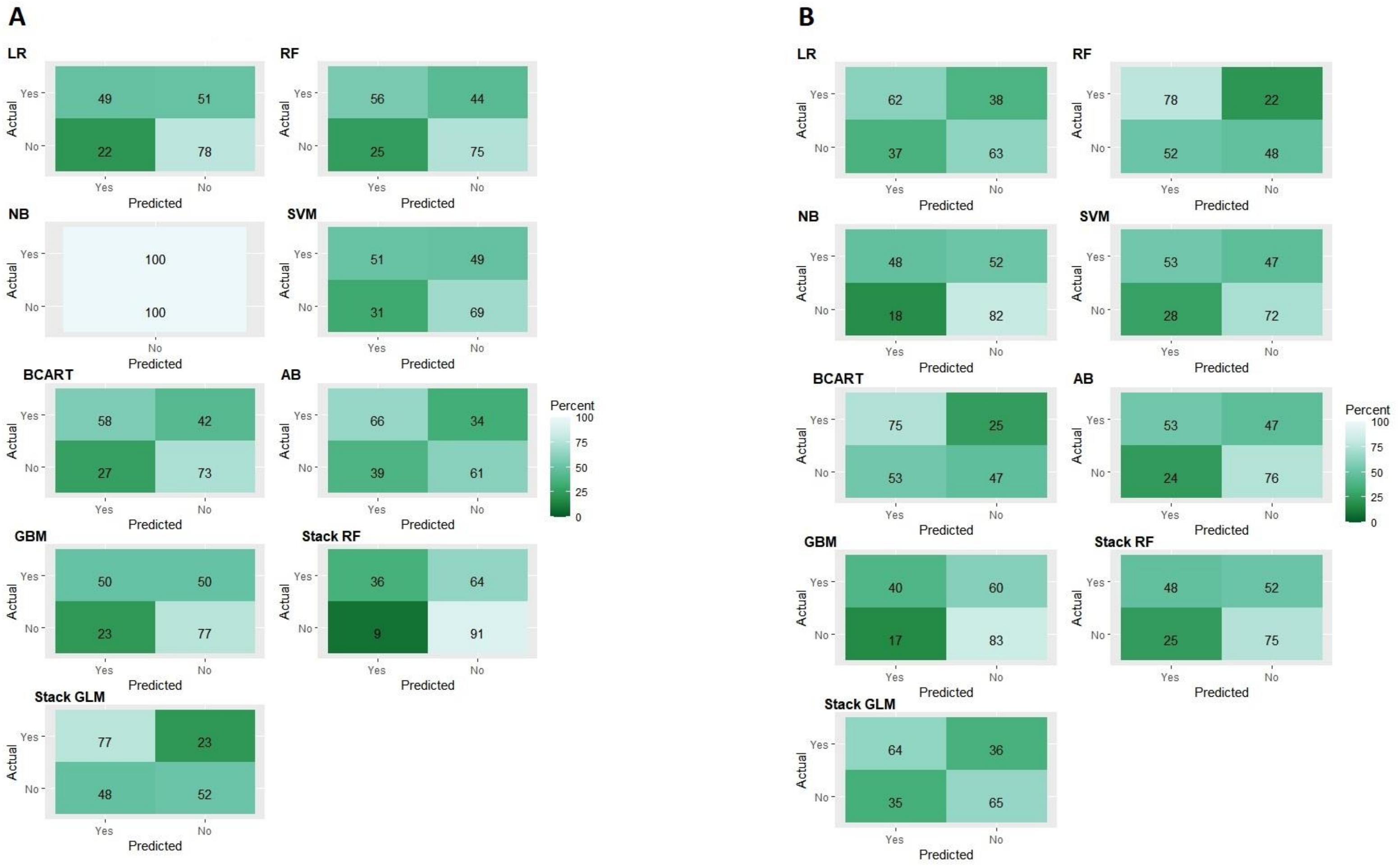


Figure 6.8. Confusion matrices for models predicting severe postoperative complications, using all preoperatively available variables, developed with original training data (A) and upsampled training data (B).



6.3. Study 1.2 Results: Models Predicting Acute Kidney Injury Preoperatively

Based on the complications recorded in the CaTHI database, 5.22% of the patients had renal complications ("acute renal failure", "acute renal dysfunction", "acute kidney injury", "other renal complication" and "postoperative elevated creatinine"). However, in this thesis, as explained in Chapter 5, Section 5.6, AKI is defined based on KDIGO guidelines, which follow the laboratory results of serum creatinine entered preoperatively in the CaTHI database and in the ICU in the Centricity^{TM} CIS database. Based on KDIGO guidelines, which are an internationally recognised way of diagnosing AKI [106], the prevalence of postoperative AKI in this patient population is 18.93% (95% CI 18.02 - 19.88). This shows that renal complications, including AKI, are underreported in the CaTHI database.

As done in previous section, here also the patient population with AKI was compared with the patient population without AKI, using the preoperatively available variables and Chi-Squared Test of Independence (Appendix 6.2). The patients with AKI were significantly different from patients with no postoperative AKI based on age group, sex, BMI, type II diabetes, surgical procedure, surgical priority, critical preoperative state, previous cardiac surgery, extracardiac arteriopathy, LV function, NYHA grade, angina status, rhythm, preoperative renal function, preoperative creatinine levels, pulmonary disease, hypertension history, congestive cardiac failure, and active endocarditis. The descriptive statistics of how patients with AKI are spread in the population can be seen from Appendix 6.2.

6.3.1. Preoperative Variables Associated with AKI

The Table 6.4 shows the variables that are associated with AKI based on unadjusted and adjusted odds ratios. Following covariate adjustment, 10 variables were associated with AKI. Patients who were between 75 and 99 years old are 1.53 (95% CI 1.17-1.99) times more likely to have postoperative AKI than patients who were 60 or younger. Patients who had a valve surgery were 1.53 (95% CI 1.23-1.99) times and patients undergoing combined CABG and valve surgery were 1.78 (95% CI 1.39-2.26) times more likely to have AKI than patients with CABG surgery.

As expected, patients who had severely impaired renal function preoperatively were 1.43 (95% CI 1.08-1.87) times more likely to have postoperative AKI than patients with normal renal function. This was also reflected in preoperative creatinine, which showed that patients with preoperative creatinine higher than 100 µmol/l were 2.10 (95% CI 1.73-2.55) times more likely to have AKI after surgery than patients with lower creatinine levels.

Patients who had emergency surgery (OR = 3.53, 95% CI 1.90-6.41) or urgent surgery (OR = 1.29, 95% CI 1.02-1.63) were more likely to have AKI than patients with elective surgery. Patients who had had a previous cardiac surgery were 1.70 (95% CI 1.12-2.55) times more likely to have AKI after surgery than patients who had a cardiac surgery for the first time.

When having extracardiac arteriopathy, patients were 1.39 (95% CI 1.09-1.75) times more likely to have AKI than with no extracardiac arteriopathy. Patients with hypertension history were 1.32 (95% CI 1.08-1.61) times more likely to have AKI than patients with no hypertension.

Interestingly, patients with level III angina status and with unknown left main stem status were less likely to have postoperative AKI than patients with no angina or with no left main stem problems. In terms of angina, the angina status is a highly subjective measure, where the patients are categorised based on their ability to carry out any physical movement based on a clinician's observation [334]. The fact that patients with unknown left main stem status were significantly associated with the outcome shows the importance of the improvement in data quality in clinical databases.

Table 6.4. How variables were associated with postoperative AKI based on unadjusted and adjusted odds ratios. Only variables significant based on unadjusted odds ratios are included in the table.

Variable	Level	Unadjusted OR (95% CI)	P-value	Adjusted OR (95% CI)	P-value
Age Category	60 or under	1.00		1.00	
	61 to 67	1.24 (0.97 - 1.58)	0.0862	1.14 (0.88 - 1.48)	0.3087
	68 to 74	1.44 (1.15 - 1.81)	0.0017	1.13 (0.87 - 1.45)	0.3615
	75 to 99	2.42 (1.96 - 3.01)	<0.0001	1.53 (1.17 - 1.99)	0.0017
Sex	Male	1.00		1.00	
	Female	1.39 (1.18 - 1.63)	<0.0001	1.20 (0.99 - 1.45)	0.0560
Type II Diabetes	No	1.00		1.00	
	Yes	1.21 (1.02 - 1.44)	0.0252	1.17 (0.98 - 1.41)	0.0881
Smoking Status	Never smoked	1.00		1.00	
	Ex-smoker	0.86 (0.71 - 1.05)	0.1356	0.92 (0.75 - 1.12)	0.4009
	Current smoker	0.66 (0.50 - 0.86)	0.0021	0.81 (0.61 - 1.08)	0.1543
	Unknown	0.85 (0.69 - 1.04)	0.1160	0.88 (0.70 - 1.10)	0.2612
Procedure	CABG	1.00		1.00	
	Valve	2.17 (1.82 - 2.57)	<0.0001	1.56 (1.23 - 1.99)	0.0003
	CABG and Valve	2.52 (2.04 - 3.11)	<0.0001	1.78 (1.39 - 2.26)	<0.0001
Renal Function Before Surgery	Normal	1.00		1.00	
	Moderately Impaired	0.61 (0.51 - 0.74)	<0.0001	0.98 (0.79 - 1.23)	0.8901
	Severely Impaired	2.49 (1.95 - 3.19)	<0.0001	1.43 (1.08 - 1.87)	0.0106
	Unknown	0.74 (0.59 - 0.91)	0.0055	0.94 (0.74 - 1.18)	0.5914
Preoperative Creatinine	<100µmol/l	1.00		1.00	
	100 or higher	2.73 (2.33 - 3.20)	<0.0001	2.10 (1.73 - 2.55)	<0.0001
Priority	Elective	1.00		1.00	
	Emergency	4.15 (2.42 - 6.94)	<0.0001	3.53 (1.90 - 6.41)	<0.0001
	Priority	0.90 (0.71 - 1.12)	0.3319	0.87 (0.68 - 1.10)	0.2475
	Urgent	1.37 (1.13 - 1.65)	0.0012	1.29 (1.02 - 1.63)	0.0323
Critical Pre-op. State	No	1.00		1.00	
	Yes	2.33 (1.47 - 3.60)	0.0002	1.14 (0.65 - 1.93)	0.6435
Previous Cardiac Surgery	No	1.00		1.00	
	Yes	2.35 (1.60 - 3.38)	<0.0001	1.70 (1.12 - 2.55)	0.0114
Extracardiac Arteriopathy	No	1.00		1.00	
	Yes	1.38 (1.11 - 1.71)	0.0037	1.39 (1.09 - 1.75)	0.0062
Left Ventricular Function	Good	1.00		1.00	
	Moderate	1.24 (1.02 - 1.50)	0.0269	1.11 (0.90 - 1.37)	0.3091
	Poor	1.79 (1.19 - 2.60)	0.0034	1.22 (0.78 - 1.85)	0.3766
NYHA Grade	I	1.00		1.00	
	II	0.92 (0.74 - 1.14)	0.4305	0.85 (0.68 - 1.07)	0.1652
	III	1.45 (1.16 - 1.82)	0.0012	1.03 (0.81 - 1.32)	0.8067
	IV	2.97 (2.08 - 4.22)	<0.0001	1.46 (0.95 - 2.21)	0.0801
Angina Status	0	1.00		1.00	
	I	0.70 (0.55 - 0.89)	0.0035	0.87 (0.67 - 1.12)	0.2801

		Unadjusted		Adjusted	
Variable	**Level**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
	II	0.59 (0.49 - 0.71)	<0.0001	0.85 (0.68 - 1.06)	0.1552
	III	0.48 (0.37 - 0.62)	<0.0001	0.66 (0.48 - 0.89)	0.0073
	IV	0.69 (0.47 - 0.98)	0.0457	0.66 (0.42 - 1.01)	0.0597
Active Endocarditis	No			1.00	
	Yes	2.86 (1.64 - 4.79)	0.0001	1.45 (0.77 - 2.64)	0.2337
Rhythm	Normal	1.00		1.00	
	Abnormal	1.86 (1.51 - 2.28)	<0.0001	1.12 (0.89 - 1.40)	0.3464
	Unknown	0.88 (0.61 - 1.24)	0.4750	0.88 (0.59 - 1.27)	0.4979
Left Main Stem Disease	No	1.00		1.00	
	Yes	0.77 (0.60 - 0.97)	0.0329	1.01 (0.77 - 1.32)	0.9305
	Unknown	0.76 (0.64 - 0.90)	0.0012	0.79 (0.66 - 0.95)	0.0122
Hypertension History	No	1.00		1.00	
	Yes	1.31 (1.09 - 1.56)	0.0036	1.32 (1.08 - 1.61)	0.0059
Congestive Cardiac Failure	No	1.00		1.00	
	Yes	2.15 (1.74 - 2.65)	<0.0001	1.09 (0.84 - 1.41)	0.5072

6.3.2. Models Predicting AKI Using Preoperative Data

In this section, classification methods were compared at predicting AKI, when using different number of variables. Since this was an imbalanced classification problem, upsampling was also experimented with.

The experiments in terms of the variables include the following:

> **Experiment 1:** Predicting AKI, using only the variables that were significantly associated with AKI based adjusted odds ratios. These variables are *age category, surgical procedure, preoperative renal function, preoperative creatinine, surgical priority, previous cardiac surgery, extracardiac arteriopathy, angina status, left main stem status* and *hypertension history* (10 variables).
>
> **Experiment 2:** Predicting AKI, using only the variables that are used to calculate logistic EuroSCORE (15 variables). These variables are listed in Section 5.8.1.2.
>
> **Experiment 3:** Predicting AKI, using all variables available in the dataset (24 variables). These variables are listed in Appendix 6.2.

The patient demographics in training and testing data are shown in Appendix 6.3.

6.3.2.1. Models' Discrimination

As seen from Table 6.5, gradient boosting developed with all preoperative variables had the highest performance in terms of AUC (0.666), closely followed by random forest (AUC = 0.665). In terms of sensitivity, stacked model with random forest using all variables achieved the highest performance (Sens = 0.671). Overall, models had higher specificity than sensitivity, support vector machine developed with 10 variables having had the highest (Spec = 0.876). For all models the positive predictive value (PPV) was very low, even though the prevalence of AKI was not very low (ca. 19%). This shows that the models were not very good at recognising patients with AKI, as also shown by sensitivity. The negative predictive values were considerably higher than positive predictive values, the stacked model with random forest using the logistic EuroSCORE variables (15 variables) having had the highest (NPV = 0.753).

Looking at the Table 6.6, where the performance of models using upsampled training data is shown, stacked model with generalised linear model had the highest performance of AUC = 0.667. This model used all preoperative variables. Based on sensitivity, random forest using 10 variables had the highest performance (Sens = 0.746). BCART with all preoperative variables had the highest specificity of 0.858. Even though the training set used upsampled data, using these models on the testing data that had the original imbalanced outcome resulted in very low positive predictive values for all models. The highest NPV, however, belonged to random forest using 10 variables (NPV = 0.774).

Table 6.5. Models predicting AKI based on different number of variables, using the original data with the optimal cut-off points where the sensitivity (Sens), specificity (Spec), positive and negative predictive values (PPV and NPV) were calculated from. For each performance measure, 95% confidence intervals (CI) are shown. The highest result for each performance measure is marked in bold.

Model	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Experiment 1 (10 variables)						
AB	0.562 (0.516 - 0.608)	0.370 (0.325 - 0.415)	0.739 (0.698 - 0.780)	**0.175 (0.140 - 0.210)**	0.739 (0.698 - 0.780)	0.391
BCART	0.597 (0.552 - 0.642)	0.561 (0.515 - 0.607)	0.589 (0.543 - 0.635)	0.156 (0.122 - 0.190)	0.747 (0.707 - 0.787)	0.001
GBM	0.634 (0.589 - 0.679)	0.468 (0.422 - 0.514)	0.769 (0.730 - 0.808)	0.147 (0.114 - 0.180)	0.666 (0.622 - 0.710)	0.212
LR	0.644 (0.600 - 0.688)	0.508 (0.462 - 0.554)	0.733 (0.692 - 0.774)	0.143 (0.111 - 0.175)	0.679 (0.636 - 0.722)	0.203
NB	0.632 (0.587 - 0.677)	0.541 (0.495 - 0.587)	0.701 (0.659 - 0.743)	0.140 (0.108 - 0.172)	0.690 (0.647 - 0.733)	0.001
RF	0.635 (0.590 - 0.680)	0.526 (0.480 - 0.572)	0.713 (0.671 - 0.755)	0.142 (0.110 - 0.174)	0.687 (0.644 - 0.730)	0.025
Stack GLM	0.625 (0.580 - 0.670)	0.401 (0.356 - 0.446)	0.794 (0.757 - 0.831)	0.155 (0.122 - 0.188)	0.679 (0.636 - 0.722)	0.211
Stack RF	0.589 (0.543 - 0.635)	0.506 (0.460 - 0.552)	0.649 (0.605 - 0.693)	0.156 (0.122 - 0.190)	0.741 (0.700 - 0.782)	0.066
SVM	0.564 (0.518 - 0.610)	0.283 (0.241 - 0.325)	**0.876 (0.846 - 0.906)**	0.169 (0.134 - 0.204)	0.638 (0.594 - 0.682)	0.181
Experiment 2 (15 variables)						
AB	0.563 (0.517 - 0.609)	0.254 (0.214 - 0.294)	0.868 (0.837 - 0.899)	0.176 (0.141 - 0.211)	0.676 (0.633 - 0.719)	0.469
BCART	0.602 (0.557 - 0.647)	0.624 (0.579 - 0.669)	0.551 (0.505 - 0.597)	0.145 (0.112 - 0.178)	0.744 (0.704 - 0.784)	0.040
GBM	0.636 (0.591 - 0.681)	0.586 (0.540 - 0.632)	0.638 (0.594 - 0.682)	0.139 (0.107 - 0.171)	0.713 (0.671 - 0.755)	0.176
LR	0.637 (0.593 - 0.681)	0.630 (0.585 - 0.675)	0.597 (0.552 - 0.642)	0.133 (0.102 - 0.164)	0.720 (0.678 - 0.762)	0.173
NB	0.628 (0.583 - 0.673)	0.559 (0.513 - 0.605)	0.654 (0.610 - 0.698)	0.143 (0.111 - 0.175)	0.713 (0.671 - 0.755)	0.001
RF	0.626 (0.581 - 0.671)	0.612 (0.567 - 0.657)	0.597 (0.552 - 0.642)	0.139 (0.107 - 0.171)	0.726 (0.685 - 0.767)	0.005
Stack GLM	0.618 (0.573 - 0.663)	0.419 (0.373 - 0.465)	0.782 (0.744 - 0.820)	0.153 (0.120 - 0.186)	0.682 (0.639 - 0.725)	0.156
Stack RF	0.587 (0.541 - 0.633)	0.539 (0.493 - 0.585)	0.601 (0.556 - 0.646)	0.157 (0.123 - 0.191)	**0.753 (0.713 - 0.793)**	0.294
SVM	0.551 (0.505 - 0.597)	0.296 (0.254 - 0.338)	0.832 (0.797 - 0.867)	0.174 (0.139 - 0.209)	0.696 (0.653 - 0.739)	0.181
Experiment 3 (24 variables)						
AB	0.625 (0.580 - 0.670)	0.481 (0.435 - 0.527)	0.706 (0.664 - 0.748)	0.155 (0.122 - 0.188)	0.711 (0.669 - 0.753)	0.370
BCART	0.630 (0.585 - 0.675)	0.499 (0.453 - 0.545)	0.687 (0.644 - 0.730)	0.153 (0.120 - 0.186)	0.716 (0.674 - 0.758)	0.240

Model	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
GBM	**0.666 (0.622 - 0.710)**	0.606 (0.561 - 0.651)	0.656 (0.612 - 0.700)	0.130 (0.099 - 0.161)	0.695 (0.652 - 0.738)	0.185
LR	0.658 (0.614 - 0.702)	0.590 (0.545 - 0.635)	0.648 (0.604 - 0.692)	0.136 (0.104 - 0.168)	0.706 (0.664 - 0.748)	0.177
NB	0.650 (0.606 - 0.694)	0.646 (0.602 - 0.690)	0.573 (0.527 - 0.619)	0.133 (0.102 - 0.164)	0.727 (0.686 - 0.768)	0.001
RF	0.665 (0.621 - 0.709)	0.626 (0.581 - 0.671)	0.630 (0.585 - 0.675)	0.129 (0.098 - 0.160)	0.704 (0.662 - 0.746)	0.050
Stack GLM	0.639 (0.595 - 0.683)	0.539 (0.493 - 0.585)	0.685 (0.642 - 0.728)	0.141 (0.109 - 0.173)	0.706 (0.664 - 0.748)	0.195
Stack RF	0.617 (0.572 - 0.662)	**0.671 (0.628 - 0.714)**	0.505 (0.459 - 0.551)	0.137 (0.105 - 0.169)	0.752 (0.712 - 0.792)	0.156
SVM	0.594 (0.549 - 0.639)	0.325 (0.282 - 0.368)	0.844 (0.810 - 0.878)	0.166 (0.132 - 0.200)	0.659 (0.615 - 0.703)	0.195

Table 6.6. Models predicting AKI based on different number of variables, using upsampled training data with the optimal cut-off points where the sensitivity (Sens), specificity (Spec), positive and negative predictive values (PPV and NPV) were calculated from. For each performance measure, 95% confidence intervals (CI) are shown. The highest result for each performance measure is marked in bold.

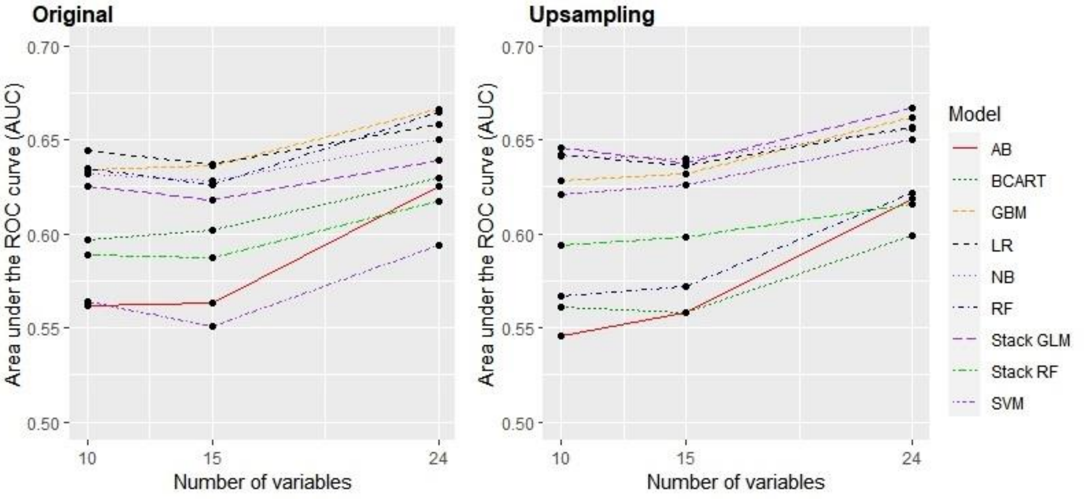
Model	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Experiment 1 (10 variables)						
AB	0.546 (0.500 - 0.592)	0.274 (0.233 - 0.315)	0.822 (0.787 - 0.857)	**0.180 (0.144 - 0.216)**	0.724 (0.683 - 0.765)	0.554
BCART	0.561 (0.515 - 0.607)	0.561 (0.515 - 0.607)	0.542 (0.496 - 0.588)	0.167 (0.133 - 0.201)	0.766 (0.727 - 0.805)	0.160
GBM	0.628 (0.583 - 0.673)	0.445 (0.399 - 0.491)	0.768 (0.729 - 0.807)	0.152 (0.119 - 0.185)	0.677 (0.634 - 0.720)	0.534
LR	0.642 (0.598 - 0.686)	0.595 (0.550 - 0.640)	0.639 (0.595 - 0.683)	0.136 (0.104 - 0.168)	0.709 (0.667 - 0.751)	0.493
NB	0.641 (0.597 - 0.685)	0.468 (0.422 - 0.514)	0.781 (0.743 - 0.819)	0.145 (0.112 - 0.178)	0.653 (0.609 - 0.697)	0.304
RF	0.567 (0.521 - 0.613)	**0.746 (0.706 - 0.786)**	0.366 (0.321 - 0.411)	0.147 (0.114 - 0.180)	**0.774 (0.735 - 0.813)**	0.050
Stack GLM	0.646 (0.602 - 0.690)	0.538 (0.492 - 0.584)	0.714 (0.672 - 0.756)	0.136 (0.104 - 0.168)	0.688 (0.645 - 0.731)	0.212
Stack RF	0.594 (0.549 - 0.639)	0.423 (0.377 - 0.469)	0.747 (0.707 - 0.787)	0.157 (0.123 - 0.191)	0.712 (0.670 - 0.754)	0.050
SVM	0.621 (0.576 - 0.666)	0.494 (0.448 - 0.540)	0.730 (0.689 - 0.771)	0.147 (0.114 - 0.180)	0.687 (0.644 - 0.730)	0.530
Experiment 2 (15 variables)						
AB -	0.558 (0.512 - 0.604)	0.252 (0.212 - 0.292)	0.842 (0.808 - 0.876)	0.181 (0.145 - 0.217)	0.716 (0.674 - 0.758)	0.526
BCART	0.558 (0.512 - 0.604)	0.637 (0.593 - 0.681)	0.465 (0.419 - 0.511)	0.163 (0.129 - 0.197)	0.772 (0.733 - 0.811)	0.120
GBM	0.632 (0.587 - 0.677)	0.550 (0.504 - 0.596)	0.674 (0.631 - 0.717)	0.142 (0.110 - 0.174)	0.705 (0.663 - 0.747)	0.493
LR	0.636 (0.591 - 0.681)	0.581 (0.535 - 0.627)	0.638 (0.594 - 0.682)	0.140 (0.108 - 0.172)	0.715 (0.673 - 0.757)	0.502
NB	0.640 (0.596 - 0.684)	0.528 (0.482 - 0.574)	0.705 (0.663 - 0.747)	0.143 (0.111 - 0.175)	0.692 (0.649 - 0.735)	0.146
RF	0.572 (0.526 - 0.618)	0.699 (0.657 - 0.741)	0.425 (0.379 - 0.471)	0.150 (0.117 - 0.183)	0.768 (0.729 - 0.807)	0.180
Stack GLM	0.638 (0.594 - 0.682)	0.601 (0.556 - 0.646)	0.638 (0.594 - 0.682)	0.132 (0.101 - 0.163)	0.713 (0.671 - 0.755)	0.183
Stack RF	0.598 (0.553 - 0.643)	0.703 (0.661 - 0.745)	0.464 (0.418 - 0.510)	0.134 (0.102 - 0.166)	0.759 (0.719 - 0.799)	0.202
SVM	0.626 (0.581 - 0.671)	0.528 (0.482 - 0.574)	0.693 (0.650 - 0.736)	0.145 (0.112 - 0.178)	0.700 (0.658 - 0.742)	0.497
AB	0.619 (0.574 - 0.664)	0.419 (0.373 - 0.465)	0.763 (0.724 - 0.802)	0.159 (0.125 - 0.193)	0.695 (0.652 - 0.738)	0.388
Experiment 3 (24 variables)						
BCART	0.599 (0.554 - 0.644)	0.272 (0.231 - 0.313)	**0.858 (0.826 - 0.890)**	0.174 (0.139 - 0.209)	0.678 (0.635 - 0.721)	0.440

Model	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
GBM	0.662 (0.618 - 0.706)	0.731 (0.690 - 0.772)	0.509 (0.463 - 0.555)	0.116 (0.086 - 0.146)	0.730 (0.689 - 0.771)	0.524
LR	0.657 (0.613 - 0.701)	0.599 (0.554 - 0.644)	0.643 (0.599 - 0.687)	0.134 (0.102 - 0.166)	0.706 (0.664 - 0.748)	0.489
NB	0.656 (0.612 - 0.700)	0.688 (0.645 - 0.731)	0.555 (0.509 - 0.601)	0.123 (0.093 - 0.153)	0.723 (0.682 - 0.764)	0.047
RF	0.622 (0.577 - 0.667)	0.425 (0.379 - 0.471)	0.768 (0.729 - 0.807)	0.157 (0.123 - 0.191)	0.687 (0.644 - 0.730)	0.335
Stack GLM	**0.667 (0.623 - 0.711)**	0.592 (0.547 - 0.637)	0.657 (0.613 - 0.701)	0.131 (0.100 - 0.162)	0.706 (0.664 - 0.748)	0.195
Stack RF	0.616 (0.571 - 0.661)	0.414 (0.368 - 0.460)	0.761 (0.722 - 0.800)	0.157 (0.123 - 0.191)	0.704 (0.662 - 0.746)	0.252
SVM	0.650 (0.606 - 0.694)	0.624 (0.579 - 0.669)	0.629 (0.584 - 0.674)	0.130 (0.099 - 0.161)	0.706 (0.664 - 0.748)	0.468

Overall, regardless of using upsampling or original training data, models had a moderate performance when predicting AKI, with AUC ranging from 0.546 to 0.667. The models were slightly better at recognising patients without AKI than patients with AKI, as shown by higher specificity and negative predictive values.

Looking at the Figure 6.9, for both original and upsampling models, when using a larger number of variables, the models tended to have slightly higher overall performance. However, the changes in performance were very small, where the highest performance goes from 0.646 to 0.667 when the number of variables changes from 10 to 24.

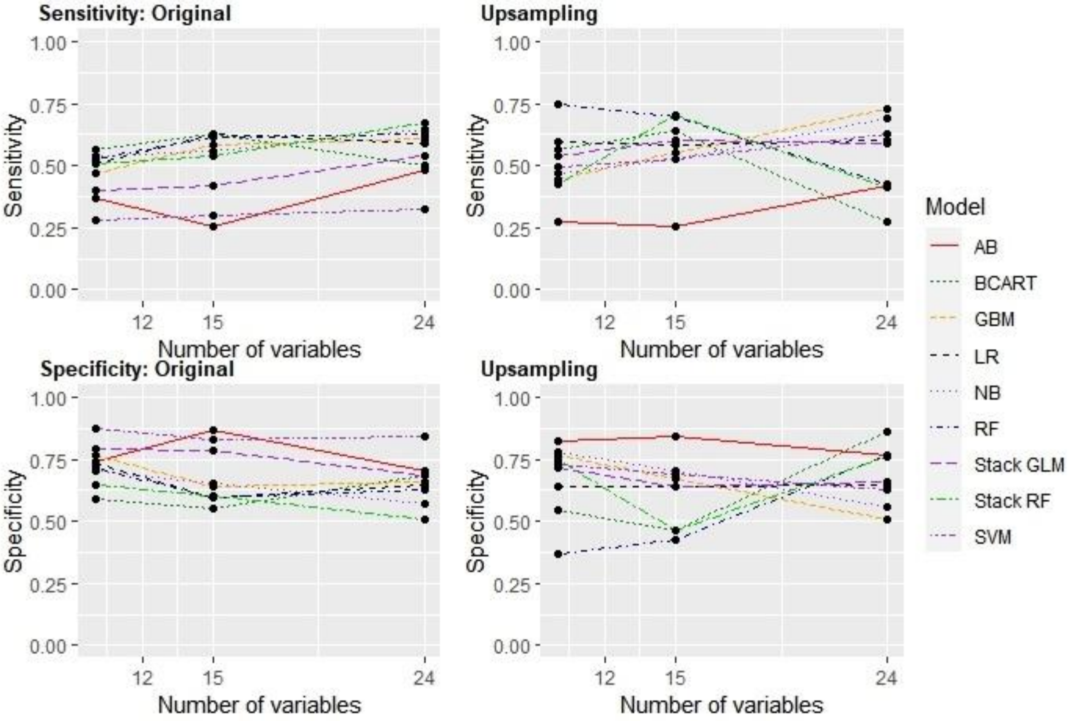
Figure 6.9. AUC for models predicting postoperative acute kidney injury (AKI), using different number of variables and original training data (left) or upsampled training data (right).



The Figure 6.10 shows that for some models the sensitivity improved slightly as the number of variables increased, however for random forest using upsampled training data the performance decreased recognisably (sensitivity from 0.746 at 10 variables to 0.425 at 24 variables). This could be due to random forest in this case not being able to handle the noise that higher number of variables can add, especially with upsampled training data.

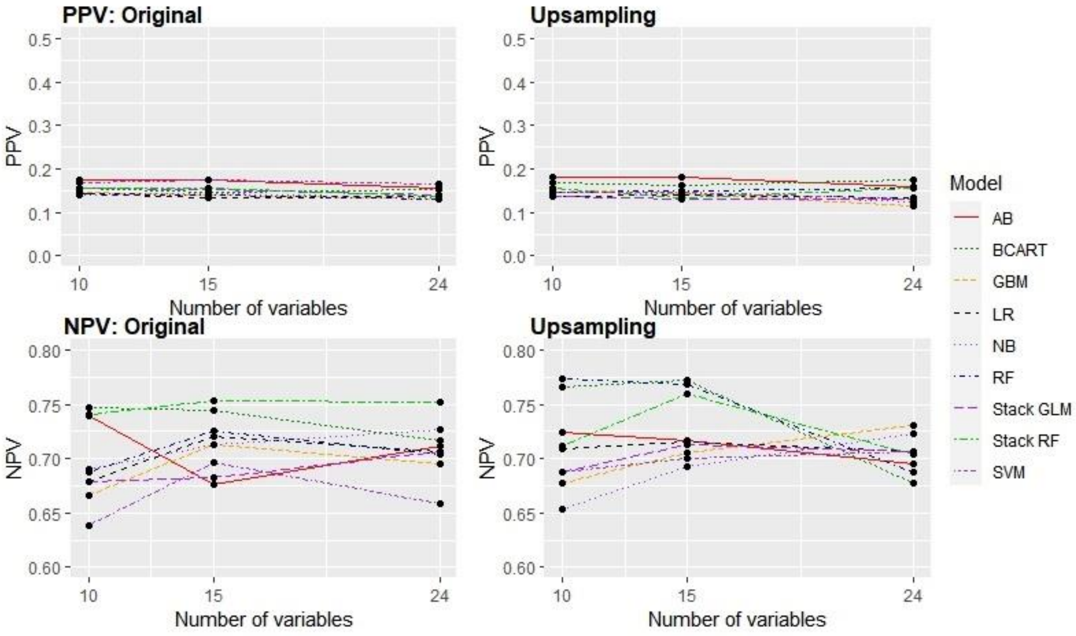
With original data, the specificity seemed to stay more stagnant for each model as the number of variables changed, however, the changes with upsampled training data were more visible. As models tended to have a trade-off between sensitivity and specificity, the random forest model's specificity increased noticeably as the number of variables increased (specificity from 0.366 with 10 variables and 0.768 with 24 variables).

Figure 6.10. Sensitivity and specificity for models predicting postoperative acute kidney injury based on different number of variables, using original training data (left) or upsampled training data (right).



Finally, looking at the Figure 6.11, the positive predictive values tended to stay constantly low for all models, regardless of using original or upsampled training data. Based on negative predictive values, there were more changes as the number of variables changed, especially for when using upsampled training data.

Figure 6.11. Positive and negative predictive values (PPV and NPV) for models predicting postoperative acute kidney injury based on different number of variables, using original training data (left) or upsampled training data (right).



6.3.2.2. Calibration and Variable Importance of the Best Performing Models

As the stacked generalised linear model with upsampled training data and 24 preoperatively available variables had the highest overall performance when predicting AKI (AUC = 0.667), calibration of this model was assessed. As seen from Figure 6.12, the predicted probabilities for patients with lower true probabilities were quite exact. This means that for these patients, the model estimated risk very well. As the true probabilities got higher, the model became less certain about the predicted probabilities. This could be explained by the use of upsampled training data, where the prevalence of AKI was 50%. In testing data, however, the prevalence of AKI was 19.50%, which is considerably lower than the prevalence of training data. The model's mean predicted probability of AKI was 19.68% (SD = 8.79%), which is very close to the prevalence of AKI (19.50%), meaning that the model was largely estimating the risk of AKI correctly.

Figure 6.12. Calibration plot for the stacked generalised linear model using 24 preoperatively available variables, developed with upsampled training data. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.

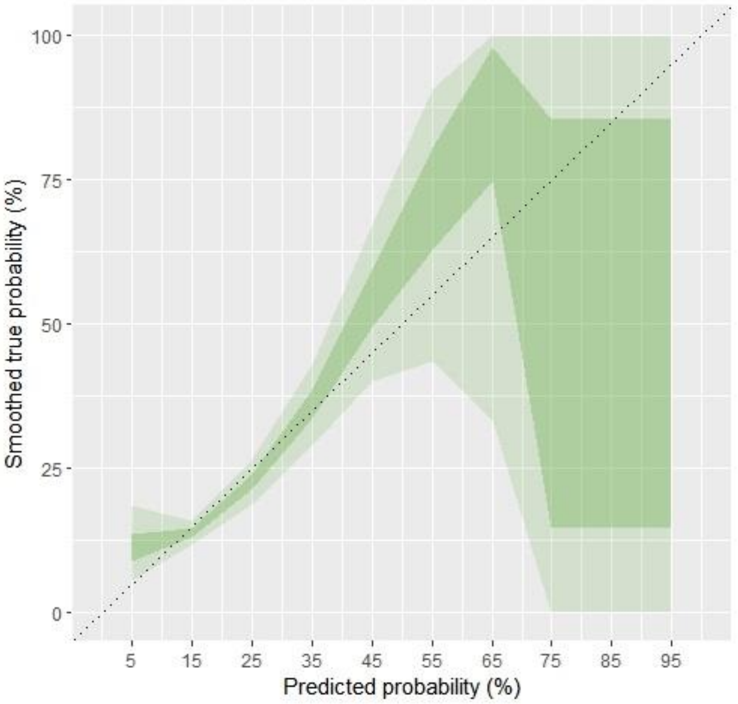
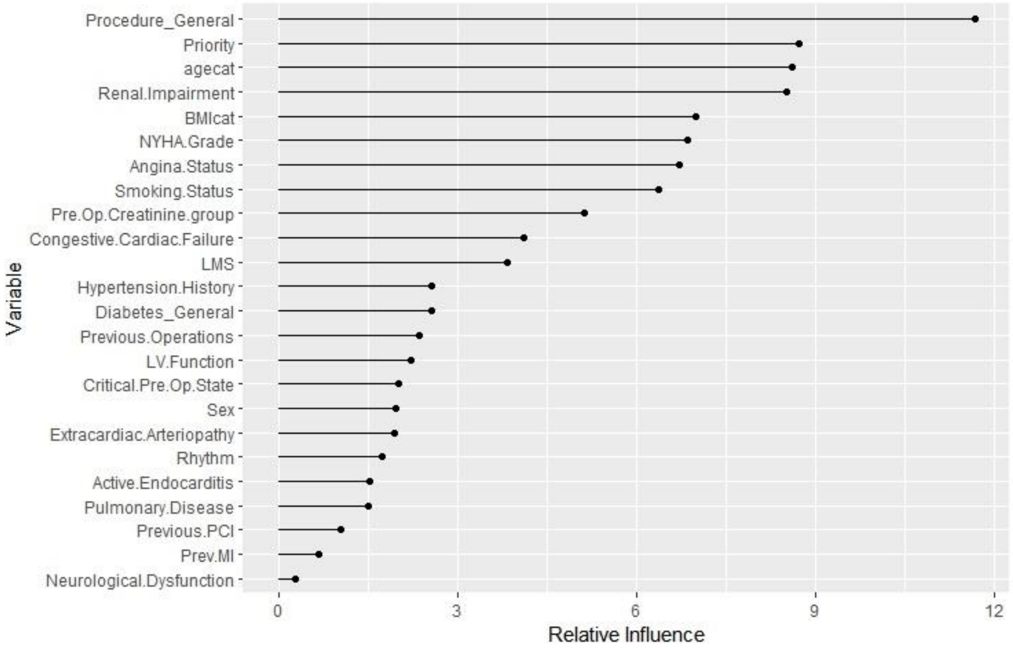


Figure 6.13. Variable importance for the gradient boosting model, using all preoperatively available variables and original training data.



Even though the stacked generalised linear model using all variables and upsampled training data had the highest performance (AUC = 0.667), the stacked models did not reveal the model coefficients and variable importance due to the complex nature of the method. Hence, the second-best performing model's variable importance is shown. Figure 6.13 shows which variables were important for the gradient boosting model using all variables and original training data (AUC = 0.666). The most important variables according to gradient boosting model were procedure, priority, age group and preoperative renal function. These four variables were also significantly associated with postoperative AKI according to adjusted odds ratios (see Section 6.3.1). It is important to note that gradient boosting model by its nature does not produce model coefficients, but only variables' relative influence.

Looking at the confusion matrices below (Figures 6.14 to 6.16), regardless of the number of variables or the training dataset used, models in general were not very good at classifying patients correctly to have postoperative AKI. Stacked random forest model using upsampled training data with logistic EuroSCORE variables had the highest percentage of AKI patients classified correctly (70% of the cases). As shown by moderately high specificity and negative predictive value, models were in general better at classifying patients not to have AKI correctly more often. Usually around 70% of the non-AKI patients were classified correctly. Support vector machine using the significant variables only and original training data classified 91% of non-AKI patients correctly. When using 15 or 24 variables, naïve Bayes tended to classify all patients as non-AKI patients when using the original training data. This highlights that if using accuracy as a performance measure, as opposed to AUC, sensitivity, specificity, PPV and NPV, then the understanding of the naïve Bayes' model performance would be easily misrepresented.

Figure 6.14. Confusion matrices for models predicting postoperative acute kidney injury, based on only variables that are significantly associated with the predicted outcome, using original training data (A) or upsampled training data (B).

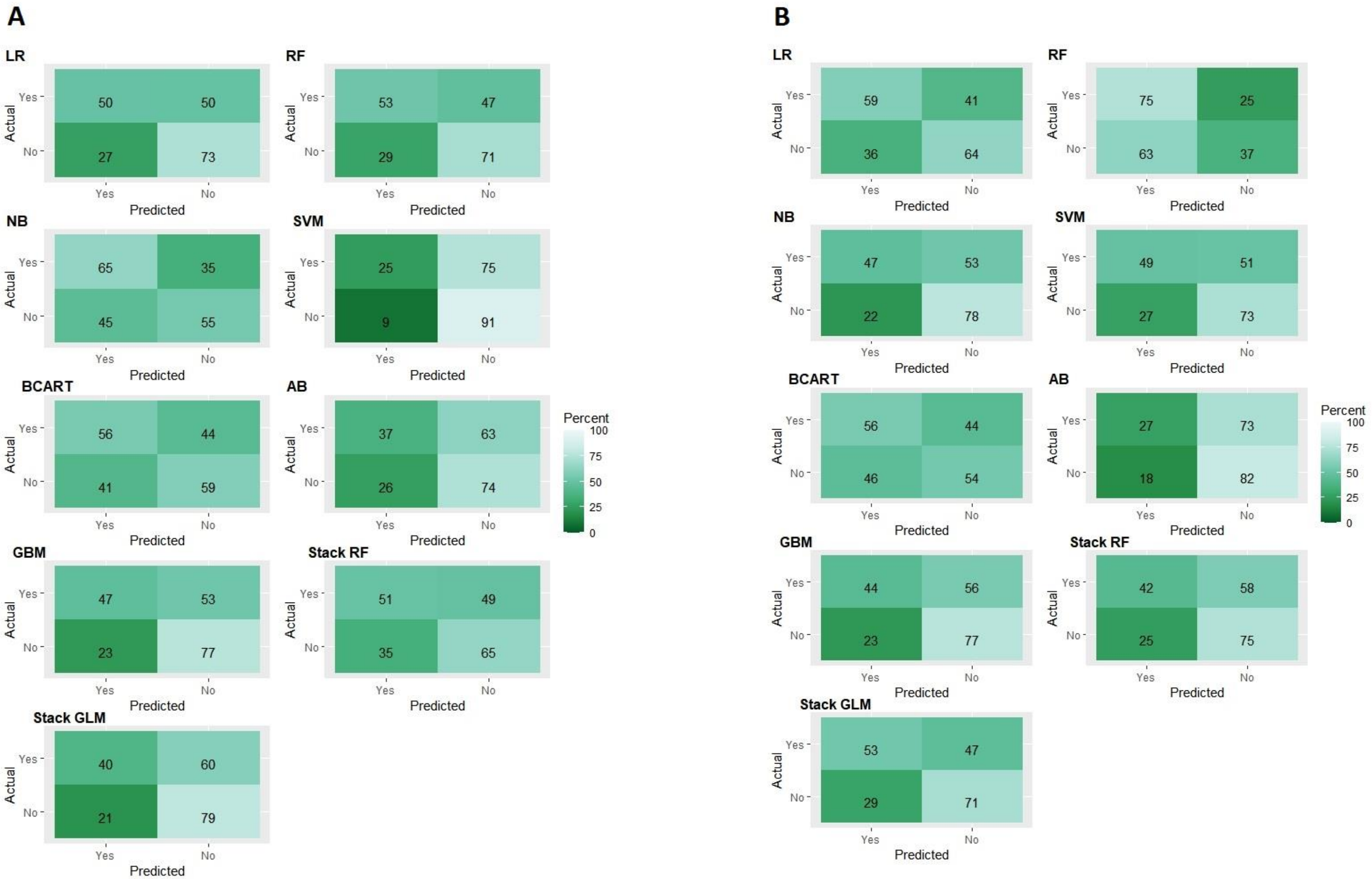


Figure 6.15. Confusion matrices for models predicting postoperative acute kidney injury, based on logistic EuroSCORE variables, using original training data (A) or upsampled training data (B).

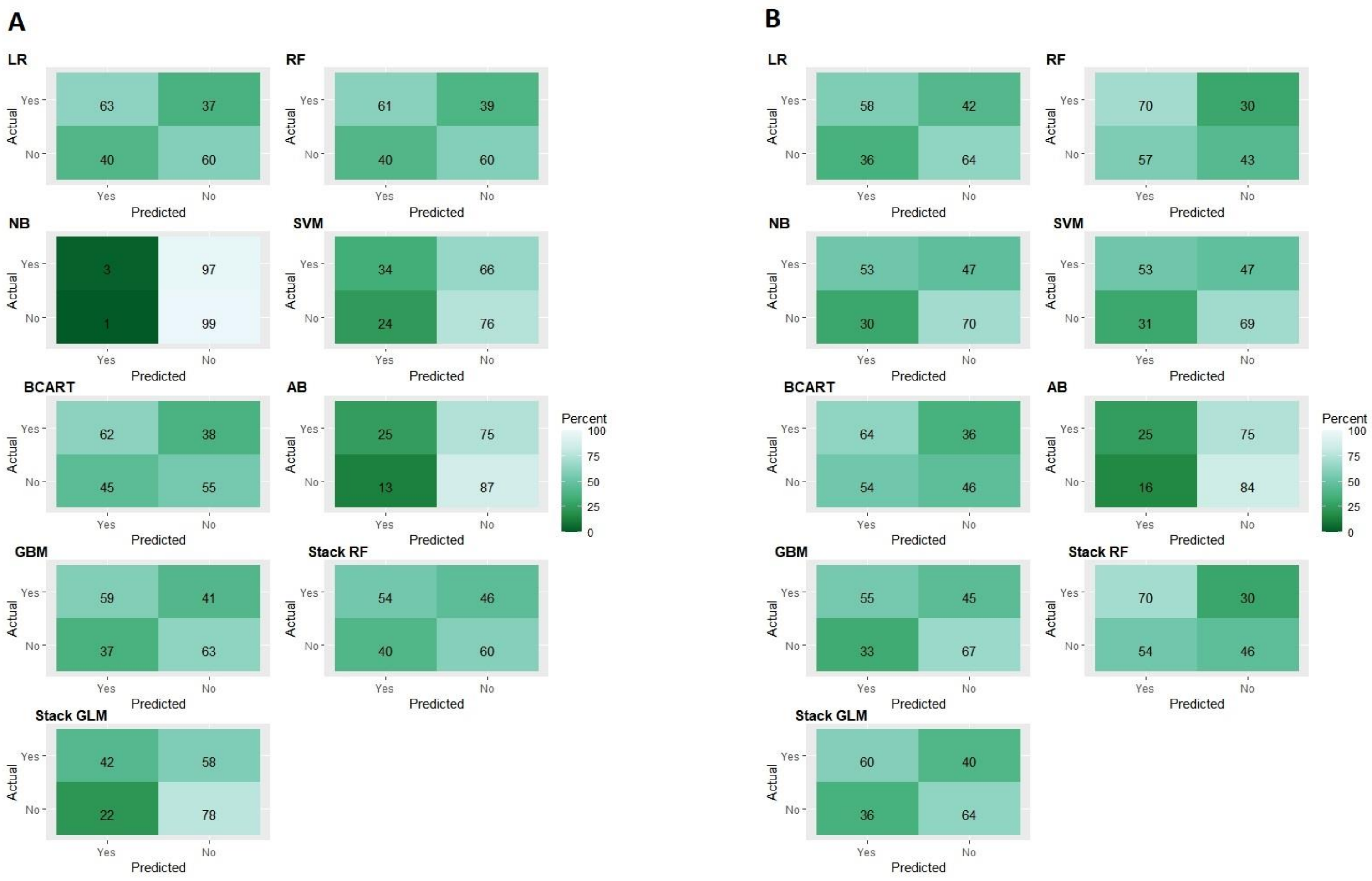
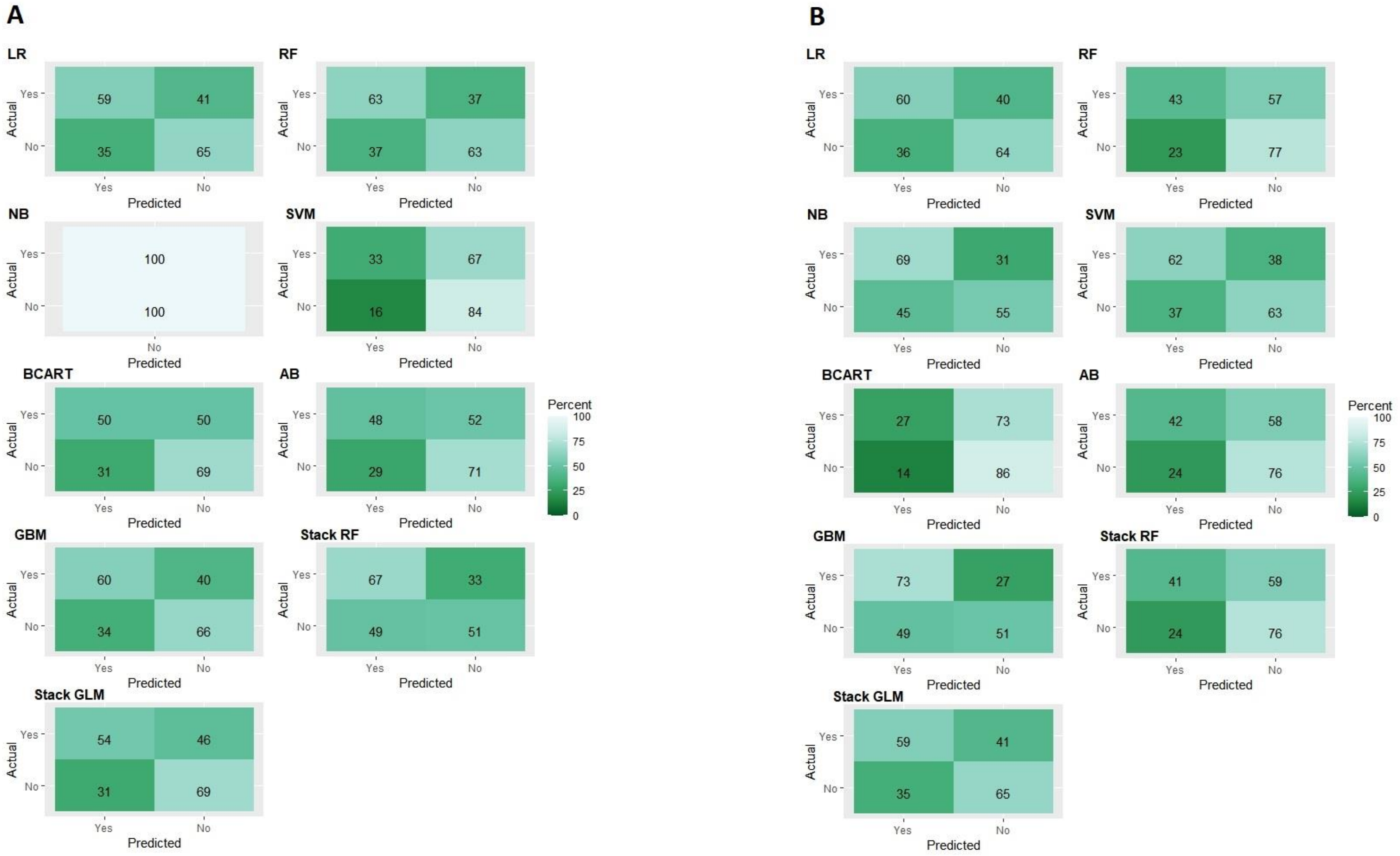


Figure 6.16. Confusion matrices for models predicting postoperative acute kidney injury, based on all preoperatively available variables, using original training data (A) or upsampled training data (B).



6.4. Study 1.3 Results: Preoperative Models Predicting Postoperative Delirium

As explained in Chapter 5, delirium is identified based on CAM-ICU assessment, which was recorded in the Centricity^{TM} CIS database together with the time of assessment. In this section, 3344 patient records were included in the analysis, where the prevalence of delirium according to CAM-ICU assessment was 12.47% (95% CI 11.39% – 13.63%). As seen with renal complications, delirium in the CaTHI database was also heavily underreported, only 1 person (0.01%, Appendix 6.1) being recorded to have delirium after surgery. Hence, the analysis in this thesis concerned delirium defined by CAM-ICU assessment as the predicted outcome.

The patients with delirium were compared to non-delirium patients, using the preoperative variables reported in the CaTHI database and Chi-Squared Test of Independence. The descriptive statistics of how the patient population overall, and patients with delirium were spread in the dataset can be found from Appendix 6.2.

The patients with postoperative delirium were significantly different from non-delirium patients based on age group, sex, surgical procedure, surgical priority, critical preoperative state, NYHA grade, rhythm, preoperative renal function, preoperative creatinine levels, left main stem disease and congestive cardiac failure.

6.4.1. Preoperative Variables Associated with Delirium

Table 6.7 shows the unadjusted and adjusted odds ratios for each variable that were significantly associated with postoperative delirium. Based on adjusted odds ratios, patients who were between 68 to 74 years old or older than 75 years were more likely to have postoperative delirium than patients who were 60 or under (OR = 1.44, 95% CI 1.03-20.3 and OR = 1.94, 95% CI 1.37-2.76, respectively). Patients who underwent combined CABG and valve surgery were 2.63 (95% CI 1.95-3.57) times more likely to have delirium in the ICU than patients who had only CABG surgery. With severely impaired preoperative renal function, patients were 1.85 (95% CI 1.22-2.80) times more likely to have delirium than patients with normal renal function. Emergency patients were 3.06 (95% CI 1.50-6.06) times more likely to have postoperative delirium than patients with elective surgery. Furthermore, patients with NYHA grade III and IV were more likely to have delirium than patients with NYHA grade I (OR = 1.55, 95% CI 1.10 - 2.20, OR = 1.89, 95% CI 1.12 - 3.18, respectively).

Table 6.7. How variables are associated with delirium, based on unadjusted and adjusted odds ratios (OR). The table includes only variables that are significant based on unadjusted odds ratios.

Variable	Level	Unadjusted OR (95% CI)	P-value	Adjusted OR (95% CI)	P-value
Age Category	60 or under	1.00		1.00	
	61 to 67	1.20 (0.85 - 1.70)	0.2872	1.11 (0.77 - 1.58)	0.5734
	68 to 74	1.72 (1.26 - 2.34)	0.0006	1.44 (1.03 - 2.03)	0.0343
	75 to 99	3.07 (2.30 - 4.12)	<0.0001	1.94 (1.37 - 2.76)	0.0002
Sex	Male	1.00		1.00	
	Female	1.46 (1.18 - 1.81)	0.0005	1.19 (0.93 - 1.52)	0.1679
Procedure	CABG	1.00		1.00	
	Valve	1.83 (1.43 - 2.33)	<0.0001	1.36 (0.98 - 1.88)	0.0658
	CABG and Valve	3.78 (2.90 - 4.94)	<0.0001	2.63 (1.95 - 3.57)	<0.0001
Renal Function Before Surgery	Normal	1.00		1.00	
	Moderately Impaired	1.71 (1.36 - 2.15)	<0.0001	1.14 (0.87 - 1.51)	0.3400
	Severely Impaired	3.76 (2.78 - 5.06)	<0.0001	1.85 (1.22 - 2.80)	0.0038
Preoperative Creatinine	<100	1.00		1.00	
	100 or higher	1.81 (1.45 - 2.26)	<0.0001	1.18 (0.88 - 1.57)	0.2584
Priority	Elective	1.00		1.00	
	Emergency	3.60 (1.93 - 6.46)	<0.0001	3.06 (1.50 - 6.06)	0.0016
	Priority	1.00 (0.77 - 1.28)	0.9720	1.00 (0.76 - 1.29)	0.9730
	Urgent	1.11 (0.86 - 1.44)	0.4190	0.99 (0.72 - 1.34)	0.9264
Critical Pre-op. State	No	1.00		1.00	
	Yes	2.85 (1.75 - 4.53)	<0.0001	1.68 (0.93 - 2.98)	0.0805
NYHA Grade	I	1.00		1.00	
	II	1.37 (1.02 - 1.88)	0.0435	1.22 (0.89 - 1.70)	0.2142
	III	2.10 (1.53 - 2.91)	<0.0001	1.55 (1.10 - 2.20)	0.0127
	IV	3.59 (2.28 - 5.60)	<0.0001	1.89 (1.12 - 3.18)	0.0166
Angina Status	0	1.00		1.00	
	I	0.71 (0.50 - 0.99)	0.0496	0.88 (0.60 - 1.25)	0.4747
	II	0.76 (0.59 - 0.98)	0.0388	1.04 (0.77 - 1.40)	0.7939
	III	0.84 (0.61 - 1.15)	0.2973	1.14 (0.77 - 1.67)	0.4951
	IV	1.00 (0.63 - 1.52)	0.9938	1.05 (0.61 - 1.77)	0.8443
Rhythm	Normal	1.00		1.00	
	Abnormal	1.58 (1.21 - 2.04)	0.0007	0.95 (0.70 - 1.26)	0.7168
	Unknown	1.13 (0.67 - 1.81)	0.6261	0.99 (0.56 - 1.66)	0.9710
Left Main Stem Disease	No	1.00		1.00	
	Yes	0.70 (0.50 - 0.96)	0.0336	0.84 (0.58 - 1.20)	0.3434
	Unknown	0.79 (0.63 - 1.00)	0.0529	0.82 (0.64 - 1.05)	0.1276
Congestive Cardiac Failure	No	1.00		1.00	
	Yes	1.84 (1.41 - 2.38)	<0.0001	1.19 (0.87 - 1.62)	0.2778

6.4.2. Models Predicting Delirium

In this section the results from three experiments are reported:

> **Experiment 1:** Predicting delirium, using only the preoperative variables that were significantly associated with delirium based adjusted odds ratios (5 variables). These variables are *age, surgical procedure, preoperative renal function, surgical priority,* and *NYHA grade*.
>
> **Experiment 2:** Predicting delirium, using only the variables that are used to calculate logistic EuroSCORE (15 variables). These variables are listed in Section 5.8.1.2.
>
> **Experiment 3:** Predicting delirium, using all variables available in the dataset (24 variables). These variables are listed in Appendix 6.2.

The patient demographics in training and testing data are shown in Appendix 6.3.

6.4.2.1. Models' Discrimination

As shown in Table 6.8, logistic regression using all preoperative variables had the highest performance (AUC = 0.675), followed closely by naïve Bayes with also all preoperative variables (AUC = 0.674). Based on sensitivity, AdaBoost with 15 variables had the highest sensitivity of 0.832. Support vector machine developed with 15 variables, however, had the highest specificity of 0.900. Even though the prevalence of delirium was not very low (12.47%), the positive predictive values were very low. On the other hand, negative predictive values were relatively high, AdaBoost developed with 15 variables having had the highest (NPV = 0.863). This was, because the models were in general better at predicting patients with no delirium, rather than patients with delirium, as shown also by confusion matrices (Figure 6.21).

The Table 6.9 shows that when using balanced training data, logistic regression using 5 variables had the highest AUC of 0.671. BCART using 15 variables had the highest sensitivity of 0.748 and stacked random forest model with all preoperative variables had the highest specificity of 0.771. Because only the training data were upsampled, and the testing data were left imbalanced to reflect the real-world prevalence of delirium, all models had very low positive predictive values. Random forest developed with 15 variables, however, had the highest NPV of 0.869.

Table 6.8. Models predicting delirium based on different number of variables, using original training data with the optimal cut-off points where the sensitivity (Sens), specificity (Spec), positive and negative predictive values (PPV and NPV) were calculated from. For each performance measure, 95% confidence intervals (CI) are shown. The highest result for each performance measure is marked in bold.

Model	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Experiment 1 (5 variables)						
AB	0.582 (0.498 - 0.666)	0.473 (0.388 - 0.558)	0.707 (0.629 - 0.785)	0.091 (0.042 - 0.140)	0.821 (0.755 - 0.887)	0.321
BCART	0.583 (0.499 - 0.667)	0.420 (0.335 - 0.505)	0.739 (0.664 - 0.814)	0.096 (0.046 - 0.146)	0.822 (0.756 - 0.888)	0.001
GBM	0.655 (0.574 - 0.736)	0.710 (0.632 - 0.788)	0.598 (0.514 - 0.682)	0.061 (0.020 - 0.102)	0.808 (0.741 - 0.875)	0.109
LR	0.672 (0.592 - 0.752)	0.687 (0.608 - 0.766)	0.629 (0.546 - 0.712)	0.063 (0.021 - 0.105)	0.800 (0.732 - 0.868)	0.108
NB	0.666 (0.585 - 0.747)	0.725 (0.649 - 0.801)	0.571 (0.486 - 0.656)	0.061 (0.020 - 0.102)	0.814 (0.747 - 0.881)	0.001
RF	0.608 (0.524 - 0.692)	0.450 (0.365 - 0.535)	0.751 (0.677 - 0.825)	0.090 (0.041 - 0.139)	0.804 (0.736 - 0.872)	0.001
Stack GLM	0.668 (0.587 - 0.749)	0.680 (0.600 - 0.760)	0.628 (0.545 - 0.711)	0.065 (0.023 - 0.107)	0.801 (0.733 - 0.869)	0.098
Stack RF	0.619 (0.536 - 0.702)	0.530 (0.445 - 0.615)	0.662 (0.581 - 0.743)	0.088 (0.039 - 0.137)	0.825 (0.760 - 0.890)	0.022
SVM	0.548 (0.463 - 0.633)	0.252 (0.178 - 0.326)	0.886 (0.832 - 0.940)	0.102 (0.050 - 0.154)	0.771 (0.699 - 0.843)	0.125
Experiment 2 (15 variables)						
AB	0.554 (0.469 - 0.639)	**0.832 (0.768 - 0.896)**	0.295 (0.217 - 0.373)	0.071 (0.027 - 0.115)	**0.863 (0.804 - 0.922)**	0.152
BCART	0.539 (0.454 - 0.624)	0.389 (0.306 - 0.472)	0.700 (0.622 - 0.778)	0.105 (0.053 - 0.157)	0.851 (0.790 - 0.912)	0.120
GBM	0.628 (0.545 - 0.711)	0.489 (0.403 - 0.575)	0.718 (0.641 - 0.795)	0.088 (0.039 - 0.137)	0.811 (0.744 - 0.878)	0.143
LR	0.664 (0.583 - 0.745)	0.725 (0.649 - 0.801)	0.549 (0.464 - 0.634)	0.063 (0.021 - 0.105)	0.822 (0.756 - 0.888)	0.100
NB	0.666 (0.585 - 0.747)	0.626 (0.543 - 0.709)	0.638 (0.556 - 0.720)	0.073 (0.028 - 0.118)	0.811 (0.744 - 0.878)	0.001
RF	0.593 (0.509 - 0.677)	0.573 (0.488 - 0.658)	0.581 (0.497 - 0.665)	0.090 (0.041 - 0.139)	0.844 (0.782 - 0.906)	0.005
Stack GLM	0.663 (0.582 - 0.744)	0.550 (0.465 - 0.635)	0.726 (0.650 - 0.802)	0.078 (0.032 - 0.124)	0.786 (0.716 - 0.856)	0.088
Stack RF	0.633 (0.550 - 0.716)	0.780 (0.709 - 0.851)	0.469 (0.384 - 0.554)	0.060 (0.019 - 0.101)	0.834 (0.770 - 0.898)	0.048
SVM	0.539 (0.454 - 0.624)	0.206 (0.137 - 0.275)	**0.900 (0.849 - 0.951)**	**0.106 (0.053 - 0.159)**	0.782 (0.711 - 0.853)	0.135
Experiment 3 (24 variables)						
AB	0.598 (0.514 - 0.682)	0.542 (0.457 - 0.627)	0.614 (0.531 - 0.697)	0.091 (0.042 - 0.140)	0.841 (0.778 - 0.904)	0.282
BCART	0.628 (0.545 - 0.711)	0.534 (0.449 - 0.619)	0.679 (0.599 - 0.759)	0.085 (0.037 - 0.133)	0.817 (0.751 - 0.883)	0.160

Model	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
GBM	0.658 (0.577 - 0.739)	0.595 (0.511 - 0.679)	0.690 (0.611 - 0.769)	0.073 (0.028 - 0.118)	0.794 (0.725 - 0.863)	0.138
LR	**0.675 (0.595 - 0.755)**	0.672 (0.592 - 0.752)	0.611 (0.528 - 0.694)	0.068 (0.025 - 0.111)	0.811 (0.744 - 0.878)	0.107
NB	0.674 (0.594 - 0.754)	0.641 (0.559 - 0.723)	0.654 (0.573 - 0.735)	0.069 (0.026 - 0.112)	0.800 (0.732 - 0.868)	0.001
RF	0.645 (0.563 - 0.727)	0.466 (0.381 - 0.551)	0.748 (0.674 - 0.822)	0.088 (0.039 - 0.137)	0.801 (0.733 - 0.869)	0.055
Stack GLM	0.634 (0.552 - 0.716)	0.520 (0.434 - 0.606)	0.774 (0.702 - 0.846)	0.078 (0.032 - 0.124)	0.761 (0.688 - 0.834)	0.091
Stack RF	0.578 (0.493 - 0.663)	0.330 (0.249 - 0.411)	0.815 (0.749 - 0.881)	0.100 (0.049 - 0.151)	0.805 (0.737 - 0.873)	0.172
SVM	0.562 (0.477 - 0.647)	0.443 (0.358 - 0.528)	0.673 (0.593 - 0.753)	0.100 (0.049 - 0.151)	0.846 (0.784 - 0.908)	0.130

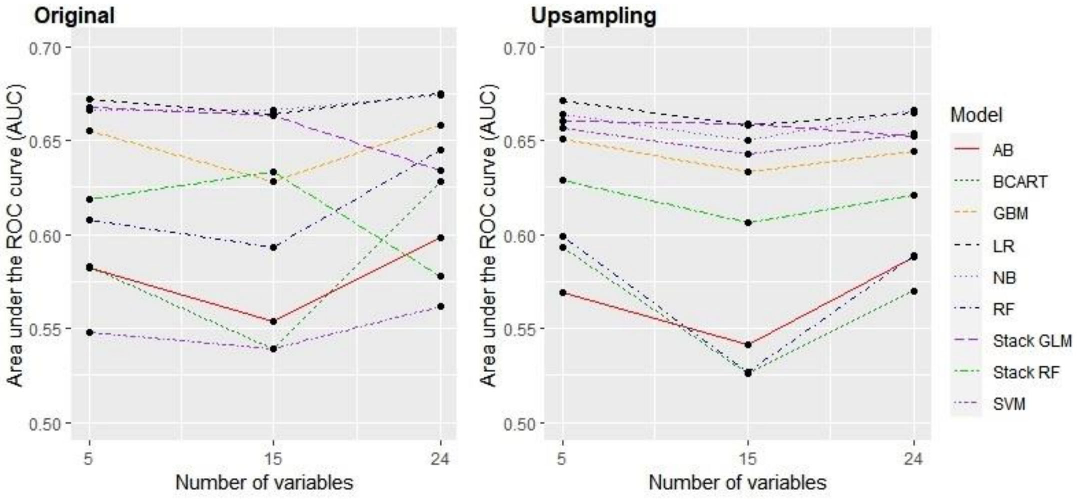
Table 6.9. Models predicting delirium based on different number of variables, using upsampled training data with the optimal cut-off points where the sensitivity (Sens), specificity (Spec), positive and negative predictive values (PPV and NPV) were calculated from. For each performance measure, 95% confidence intervals (CI) are shown. The highest result for each performance measure is marked in bold.

Model	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Experiment 1 (5 variables)						
AB	0.569 (0.484 - 0.654)	0.511 (0.425 - 0.597)	0.630 (0.547 - 0.713)	0.095 (0.045 - 0.145)	0.843 (0.781 - 0.905)	0.442
BCART	0.593 (0.509 - 0.677)	0.527 (0.442 - 0.612)	0.644 (0.562 - 0.726)	0.090 (0.041 - 0.139)	0.834 (0.770 - 0.898)	0.160
GBM	0.651 (0.569 - 0.733)	0.687 (0.608 - 0.766)	0.613 (0.530 - 0.696)	0.064 (0.022 - 0.106)	0.807 (0.739 - 0.875)	0.447
LR	**0.671 (0.591 - 0.751)**	0.672 (0.592 - 0.752)	0.631 (0.548 - 0.714)	0.066 (0.023 - 0.109)	0.803 (0.735 - 0.871)	0.457
NB	0.664 (0.583 - 0.745)	0.679 (0.599 - 0.759)	0.623 (0.540 - 0.706)	0.065 (0.023 - 0.107)	0.804 (0.736 - 0.872)	0.147
RF	0.599 (0.515 - 0.683)	0.710 (0.632 - 0.788)	0.467 (0.382 - 0.552)	0.077 (0.031 - 0.123)	0.848 (0.787 - 0.909)	0.025
Stack GLM	0.660 (0.579 - 0.741)	0.543 (0.458 - 0.628)	0.726 (0.650 - 0.802)	0.074 (0.029 - 0.119)	0.799 (0.730 - 0.868)	0.145
Stack RF	0.629 (0.546 - 0.712)	0.564 (0.479 - 0.649)	0.690 (0.611 - 0.769)	0.074 (0.029 - 0.119)	0.813 (0.746 - 0.880)	0.042
SVM	0.657 (0.576 - 0.738)	0.679 (0.599 - 0.759)	0.619 (0.536 - 0.702)	0.065 (0.023 - 0.107)	0.806 (0.738 - 0.874)	0.434
Experiment 2 (15 variables)						
AB	0.541 (0.456 - 0.626)	0.595 (0.511 - 0.679)	0.526 (0.440 - 0.612)	0.094 (0.044 - 0.144)	0.855 (0.795 - 0.915)	0.233
BCART	0.526 (0.440 - 0.612)	**0.748 (0.674 - 0.822)**	0.345 (0.264 - 0.426)	0.090 (0.041 - 0.139)	0.867 (0.809 - 0.925)	0.001
GBM	0.633 (0.550 - 0.716)	0.565 (0.480 - 0.650)	0.660 (0.579 - 0.741)	0.082 (0.035 - 0.129)	0.817 (0.751 - 0.883)	0.493
LR	0.658 (0.577 - 0.739)	0.618 (0.535 - 0.701)	0.649 (0.567 - 0.731)	0.073 (0.028 - 0.118)	0.808 (0.741 - 0.875)	0.478
NB	0.650 (0.568 - 0.732)	0.718 (0.641 - 0.795)	0.534 (0.449 - 0.619)	0.067 (0.024 - 0.110)	0.828 (0.763 - 0.893)	0.159
RF	0.527 (0.442 - 0.612)	0.679 (0.599 - 0.759)	0.393 (0.309 - 0.477)	**0.099 (0.048 - 0.150)**	**0.869 (0.811 - 0.927)**	0.085
Stack GLM	0.659 (0.578 - 0.740)	0.638 (0.556 - 0.720)	0.625 (0.542 - 0.708)	0.068 (0.025 - 0.111)	0.822 (0.756 - 0.888)	0.121
Stack RF	0.606 (0.522 - 0.690)	0.457 (0.372 - 0.542)	0.730 (0.654 - 0.806)	0.086 (0.038 - 0.134)	0.823 (0.758 - 0.888)	0.132
SVM	0.643 (0.561 - 0.725)	0.573 (0.488 - 0.658)	0.658 (0.577 - 0.739)	0.080 (0.034 - 0.126)	0.816 (0.750 - 0.882)	0.458
Experiment 3 (24 variables)						
AB	0.588 (0.504 - 0.672)	0.534 (0.449 - 0.619)	0.664 (0.583 - 0.745)	0.086 (0.038 - 0.134)	0.824 (0.759 - 0.889)	0.283
BCART	0.570 (0.485 - 0.655)	0.435 (0.350 - 0.520)	0.707 (0.629 - 0.785)	0.097 (0.046 - 0.148)	0.833 (0.769 - 0.897)	0.240

Model	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
GBM	0.644 (0.562 - 0.726)	0.679 (0.599 - 0.759)	0.597 (0.513 - 0.681)	0.068 (0.025 - 0.111)	0.815 (0.749 - 0.881)	0.490
LR	0.665 (0.584 - 0.746)	0.702 (0.624 - 0.780)	0.588 (0.504 - 0.672)	0.064 (0.022 - 0.106)	0.813 (0.746 - 0.880)	0.437
NB	0.666 (0.585 - 0.747)	0.733 (0.657 - 0.809)	0.543 (0.458 - 0.628)	0.062 (0.021 - 0.103)	0.822 (0.756 - 0.888)	0.075
RF	0.589 (0.505 - 0.673)	0.626 (0.543 - 0.709)	0.526 (0.440 - 0.612)	0.088 (0.039 - 0.137)	0.849 (0.788 - 0.910)	0.155
Stack GLM	0.652 (0.570 - 0.734)	0.596 (0.512 - 0.680)	0.701 (0.623 - 0.779)	0.068 (0.025 - 0.111)	0.799 (0.730 - 0.868)	0.136
Stack RF	0.621 (0.538 - 0.704)	0.447 (0.362 - 0.532)	**0.771 (0.699 - 0.843)**	0.083 (0.036 - 0.130)	0.802 (0.734 - 0.870)	0.186
SVM	0.654 (0.573 - 0.735)	0.641 (0.559 - 0.723)	0.627 (0.544 - 0.710)	0.072 (0.028 - 0.116)	0.812 (0.745 - 0.879)	0.431

The Figure 6.17 shows that regardless of the balance of the outcome in the training data, the AUC for models varied quite visibly. The highest AUC overall was 0.675 for logistic regression, and lowest for BCART (AUC = 0.526). The changes of AUC as the number of variables increased were not that remarkable for original data. Interestingly, with upsampled training data, the AUC reduced notably for BCART and random forest when changing from 5 variables to 15 variables. This could be because the logistic EuroSCORE variables did not include NYHA grade and preoperative renal function that were deemed to be significantly associated with delirium.

Figure 6.17. AUC of models predicting postoperative delirium, based on different number of preoperatively available variables, using original training data (left) or upsampled training data (right).



The Figure 6.18 shows that when comparing models amongst each other, the sensitivity and specificity varied visibly, especially for models developed with the original training data. The highest sensitivity was 0.832 (AB, 15 variables, original data) and the lowest was 0.206 (SVM, 15 variables, original data). For upsampled data, the variation was slightly lower, however the difference in sensitivity amongst models was noticeable, with the highest having been 0.748 (BCART, 15 variables, upsampled data), and lowest having been 0.435 (BCARTm 24 variables, upsampled data). The specificity varied from 0.771 (Stack RF, 24 variables, upsampled data) to 0.345 (BCART, 15 variables, upsampled data) for upsampling experiments and from 0.900 (SVM, 15 variables, original data) to 0.295 (AB, 15 variables, original data) for original data experiments.

Finally, the changes in PPV and NPV amongst models as the number of variables changed were not very visually noticeable (Figure 6.19). For all models the PPV stayed very low across the experiments, indicating that the models were not particularly confident at predicting patients with delirium correctly. NPV, however, stayed moderately high across the experiments with not much variation, ranging from 0.869 (RF, 15 variables, upsampled data) to 0.761 (Stack GLM, 24 variables, original data).

Figure 6.18. Sensitivity and specificity for models predicting postoperative delirium, based on different number of preoperatively available variables, using original training data (left) or upsampled training data (right).

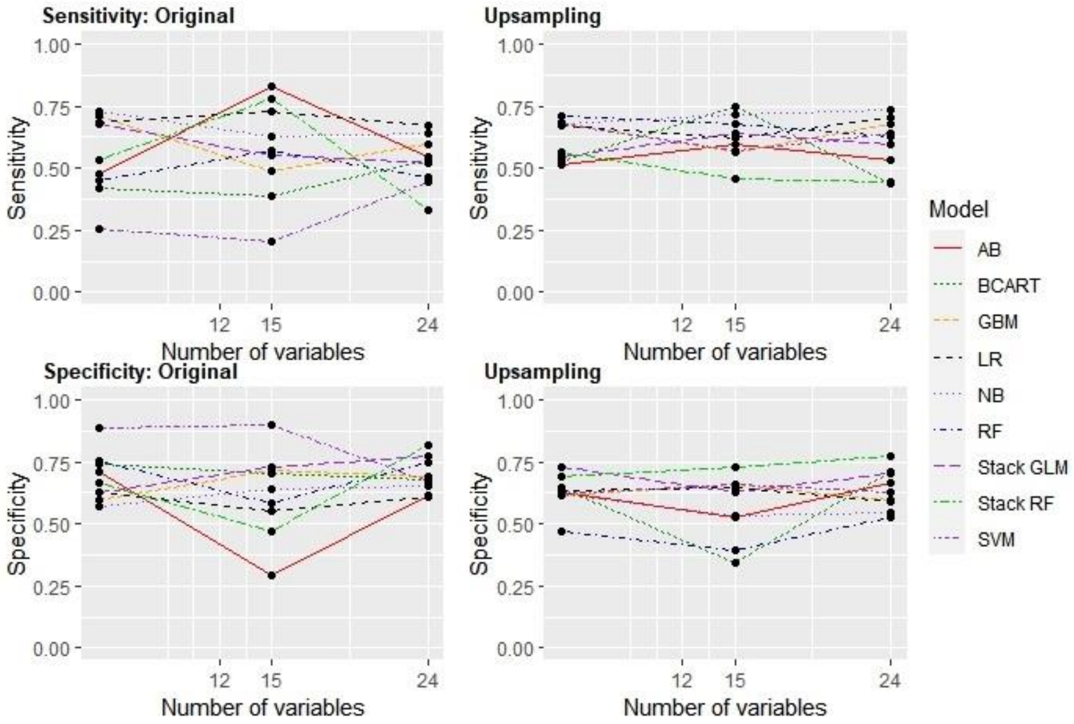
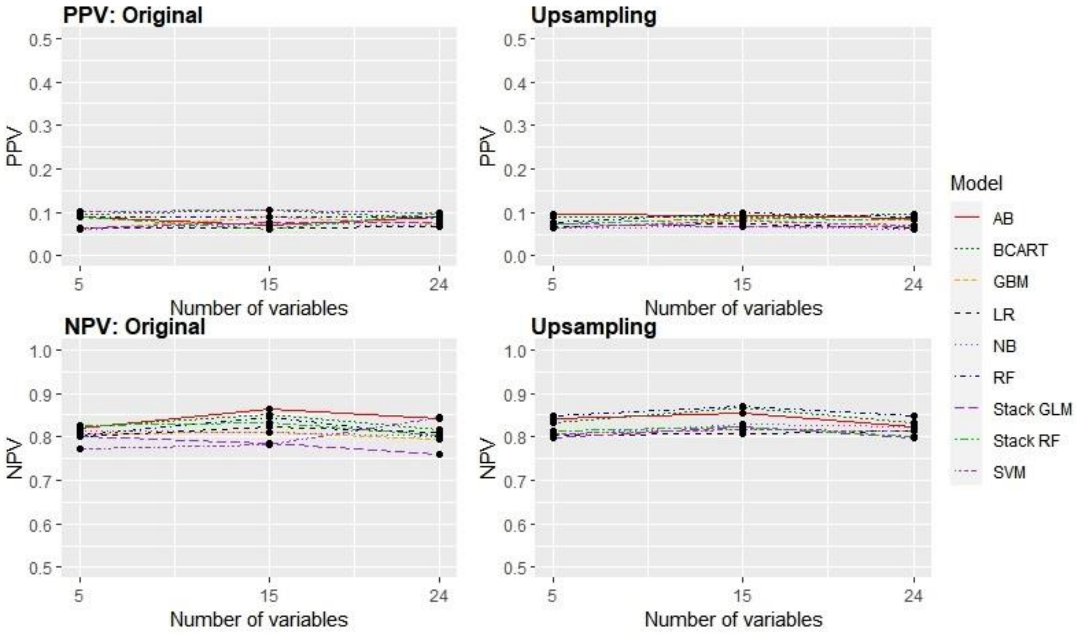


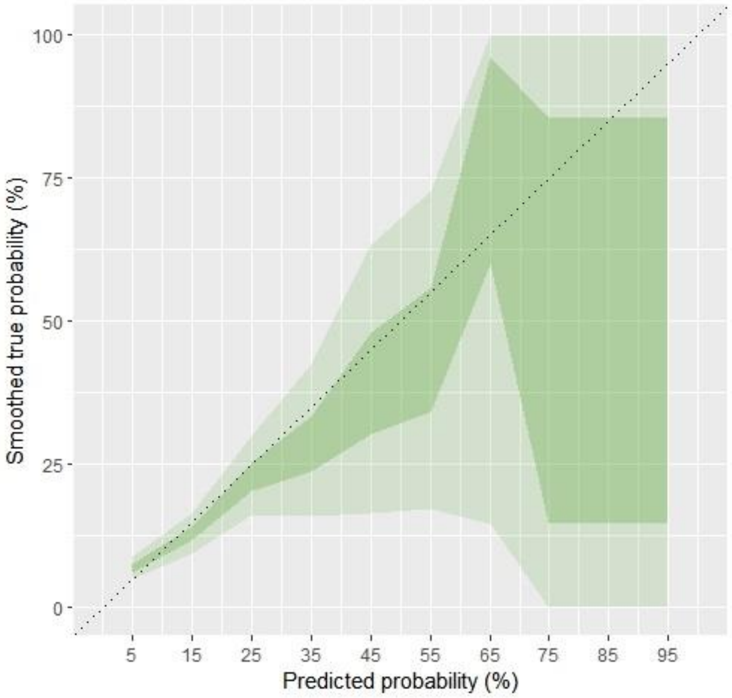
Figure 6.19. Positive and negative predictive values (PPV and NPV) for models predicting postoperative delirium, with different number of preoperatively available variables, using original training data (left) or upsampled training data (right).



6.4.2.2. Calibration and Variable Importance of the Best Performing Model

The logistic regression model that used 24 preoperatively available variables and original training data had the highest overall performance (AUC = 0.675). As seen from the calibration plot (Figure 6.20), the model was largely estimating the risk of delirium correctly for patients who had low true probability for the complication. As the true probability of delirium got higher than ca. 65%, the model's predicted probabilities become more uncertain, and the model largely overestimated risk for delirium for these patients. This can be explained by the very low positive predictive values of the model (NPV = 0.068). The mean predicted probability was 12.59% (SD = 9.27%), which is close to the prevalence of delirium (12.47%), however, the standard deviation also showed considerably low certainty, as was seen from the calibration plot.

Figure 6.20. Calibration plot for logistic regression model predicting delirium, using 24 preoperatively available variables and original training data. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.



In addition to the calibration, the model estimates with standard errors and p-values are shown (Table 6.10). Since this is a logistic regression model, to make estimates (log odds) easier to interpret, odds ratios with 95% confidence intervals are also shown. According to the p-values and odds ratios, based on this prediction model, patients with emergency surgery, combined CABG and valve surgery, having preoperatively severely impaired renal function and being older than 75 were significantly more likely to have postoperative delirium. It is worth noting that the confidence interval for odds ratios was quite wide for the emergency priority due to considerably small number of patients reported to undergo emergency surgery in this dataset (1.6%, Appendix 6.2). This means that this estimate should be interpreted with caution as the model was not very confident in this estimate.

Table 6.10. Model estimates with standard errors, p-values and odds ratios for the logistic regression model that used 24 preoperative variables and original training data.

Variable	Level	Estimate	Standard Error	P-value	OR (95% CI)
Intercept		-3.2866	0.34	<0.0001	0.04 (0.02 - 0.07)
Sex	Female	0.0816	0.16	0.5999	1.09 (0.80 - 1.47)
Priority	Emergency	1.4196	0.45	0.0014	4.14 (1.69 - 9.78)
	Priority	0.0718	0.16	0.6580	1.07 (0.78 - 1.47)
	Urgent	-0.0244	0.20	0.9050	0.98 (0.65 - 1.45)
Procedure	Valve	0.4187	0.22	0.0583	1.52 (0.99 - 2.35)
	CABG and valve	1.1182	0.20	<0.0001	3.06 (2.07 - 4.53)
LV Function	Moderate	-0.2191	0.18	0.2214	0.80 (0.56 - 1.13)
	Poor	-0.2115	0.36	0.5563	0.81 (0.39 - 1.59)
NYHA Grade	II	0.0760	0.20	0.6966	1.08 (0.74 - 1.59)
	III	0.3159	0.21	0.1331	1.37 (0.91 - 2.08)
	IV	0.5985	0.32	0.0637	1.82 (0.96 - 3.41)
Angina Status	I	-0.0348	0.23	0.8780	0.97 (0.61 - 1.49)
	II	0.0019	0.19	0.9921	1.00 (0.69 - 1.45)
	III	0.2583	0.24	0.2808	1.29 (0.80 - 2.06)
	IV	0.0097	0.35	0.9778	1.01 (0.50 - 1.96)
Renal Function	Moderately impaired	0.3292	0.18	0.0676	1.39 (0.98 - 1.98)
	Severely impaired	0.8675	0.28	0.0018	2.38 (1.38 - 4.11)
Rhythm	Abnormal	-0.0229	0.18	0.8998	0.98 (0.68 - 1.39)
	Unknown	-0.1493	0.36	0.6750	0.86 (0.41 - 1.67)
Previous Cardiac Surgery	Yes	0.4571	0.35	0.1899	1.58 (0.77 - 3.05)
Neurological Dysfunction	Yes	-0.2986	1.10	0.7855	0.74 (0.04 - 4.38)
Smoking Status	Ex-smoker	0.0048	0.16	0.9766	1.00 (0.73 - 1.39)
	Current smoker	0.2916	0.22	0.1862	1.34 (0.86 - 2.05)
	Unknown	0.1880	0.22	0.3921	1.21 (0.78 - 1.85)
Previous MI	Yes	0.2657	0.17	0.1198	1.30 (0.93 - 1.82)
Left Main Stem Disease	Yes	-0.1718	0.23	0.4593	0.84 (0.53 - 1.32)
	Unknown	-0.2872	0.17	0.0880	0.75 (0.54 - 1.04)
Pulmonary Disease	Yes	-0.0959	0.19	0.6155	0.91 (0.62 - 1.31)
Hypertension History	Yes	0.0775	0.16	0.6314	1.08 (0.79 - 1.49)
Congestive Cardiac Failure	Yes	0.0747	0.20	0.7094	1.08 (0.72 - 1.59)
Previous PCI	Yes	-0.1782	0.21	0.4012	0.84 (0.54 - 1.25)
Extracardiac Arteriopathy	Yes	0.0799	0.22	0.7116	1.08 (0.70 - 1.64)
Critical Preoperative State	Yes	0.2899	0.40	0.4630	1.34 (0.60 - 2.84)
Type II Diabetes	Yes	0.0706	0.16	0.6573	1.07 (0.78 - 1.46)
BMI Category	25.1-30.0	0.2264	0.20	0.2590	1.25 (0.85 - 1.87)
	Over 30.0	0.0712	0.18	0.6998	1.07 (0.75 - 1.55)
Age Group	61 to 67	0.0524	0.22	0.8158	1.05 (0.68 - 1.64)
	68 to 74	0.1788	0.22	0.4127	1.20 (0.78 - 1.84)
	75 to 99	0.5704	0.23	0.0119	1.77 (1.14 - 2.77)

Variable	Level	Estimate	Standard Error	P-value	OR (95% CI)
Preoperative Creatinine	≥ 100	0.1309	0.18	0.4686	1.14 (0.80 - 1.62)
Active Endocarditis	Yes	-0.3016	0.54	0.5744	0.74 (0.24 - 2.00)

Looking at the confusion matrices (Figures 6.21 to 6.23), the models were not very good at classifying patients with delirium. BCART using 15 variables and upsampling classified 75% of the patients with delirium correctly. Similarly when predicting AKI preoperatively, models were better at classifying non-delirium patients correctly than delirium patients. SVM, using original data and significant variables classified 93% of the non-delirium patients correctly. Just like when predicting AKI, naïve Bayes here also classified all patients to not have delirium when using 15 or 24 variables and original data.

Figure 6.21. Confusion matrices for models predicting postoperative delirium, based on variables significantly associated with the predicted outcome, using original training data (A) or upsampled training data (B).

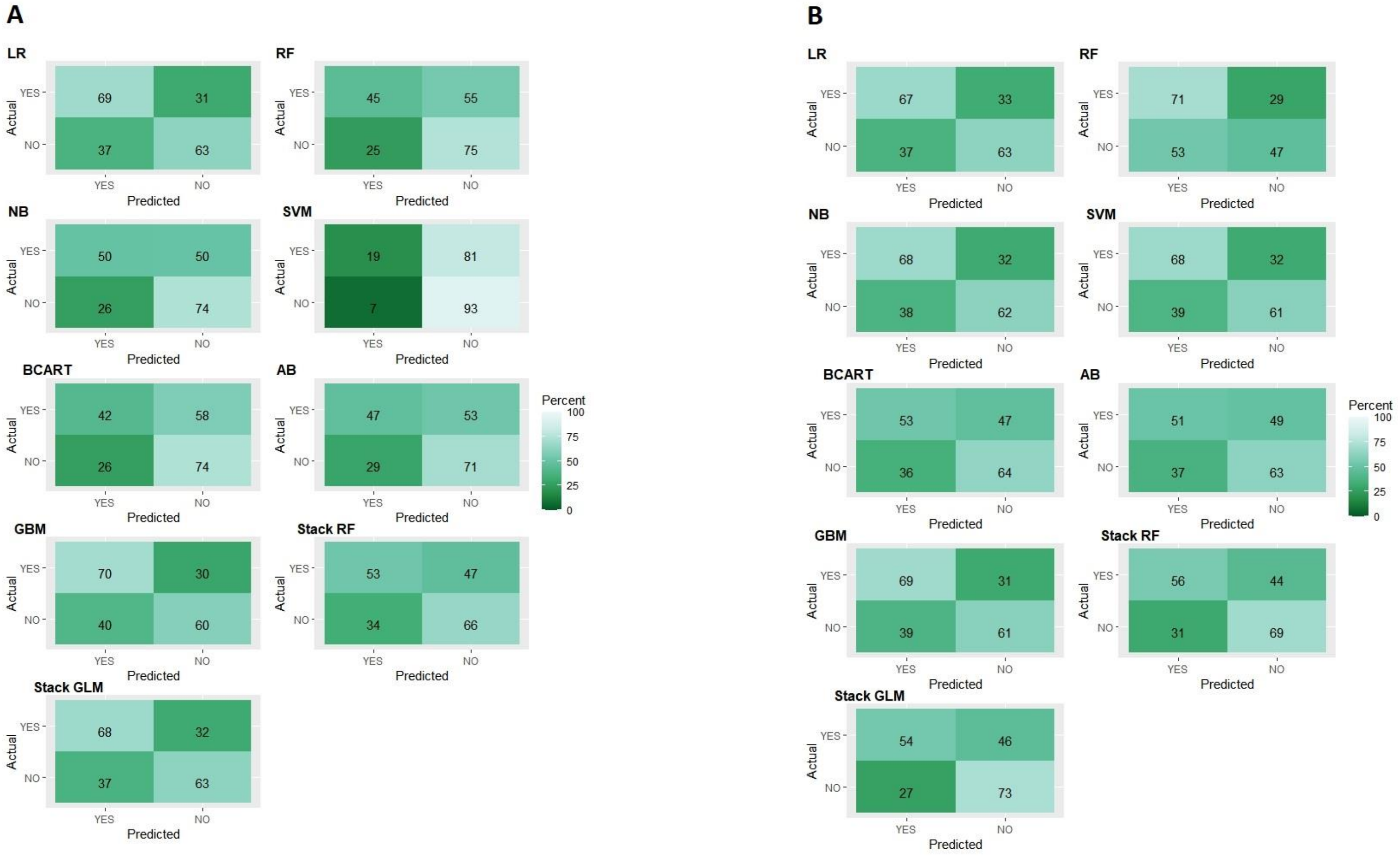


Figure 6.22. Confusion matrices for models predicting postoperative delirium, using logistic EuroSCORE variables and original training data (A) or upsampled training data (B).

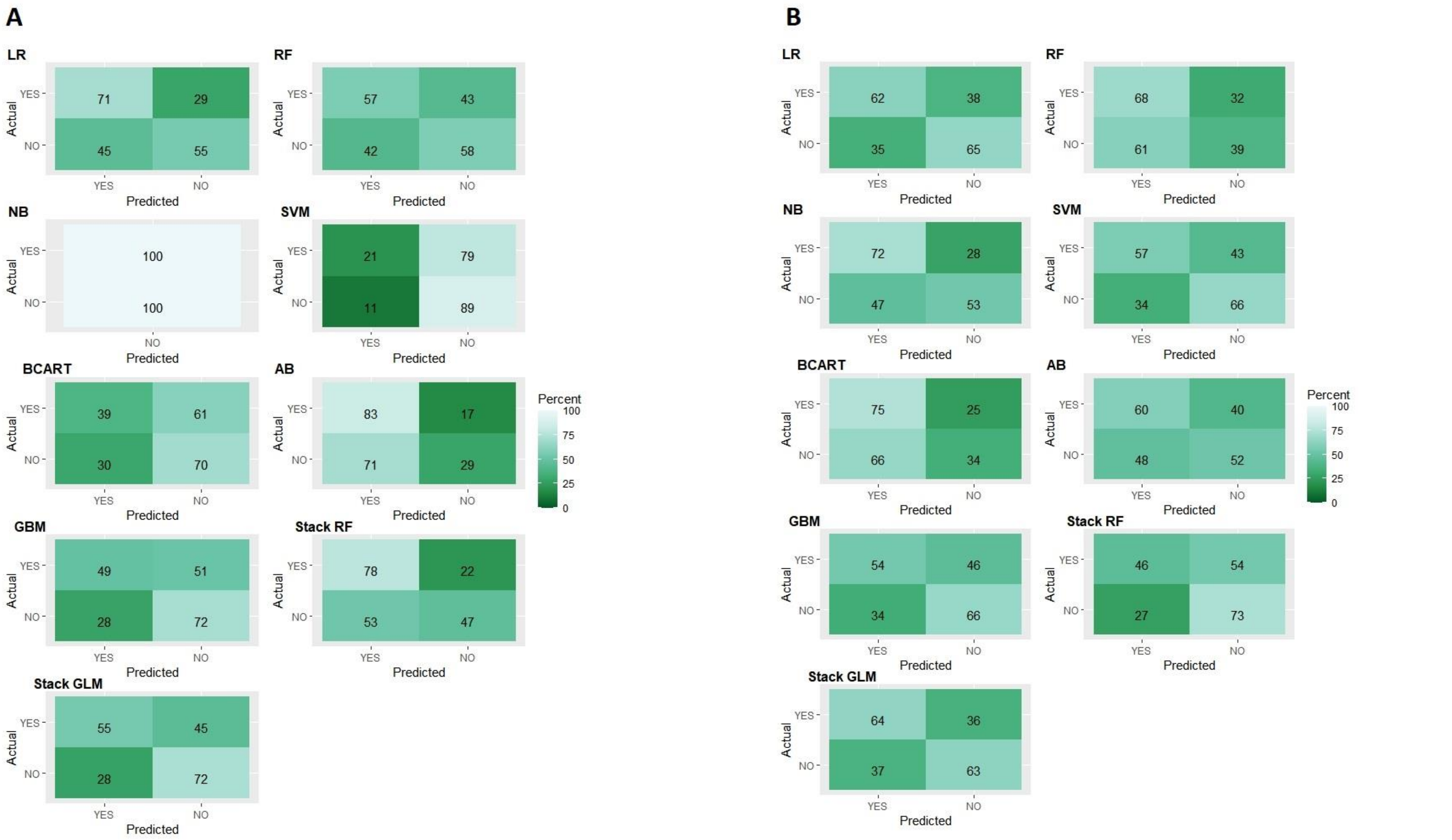
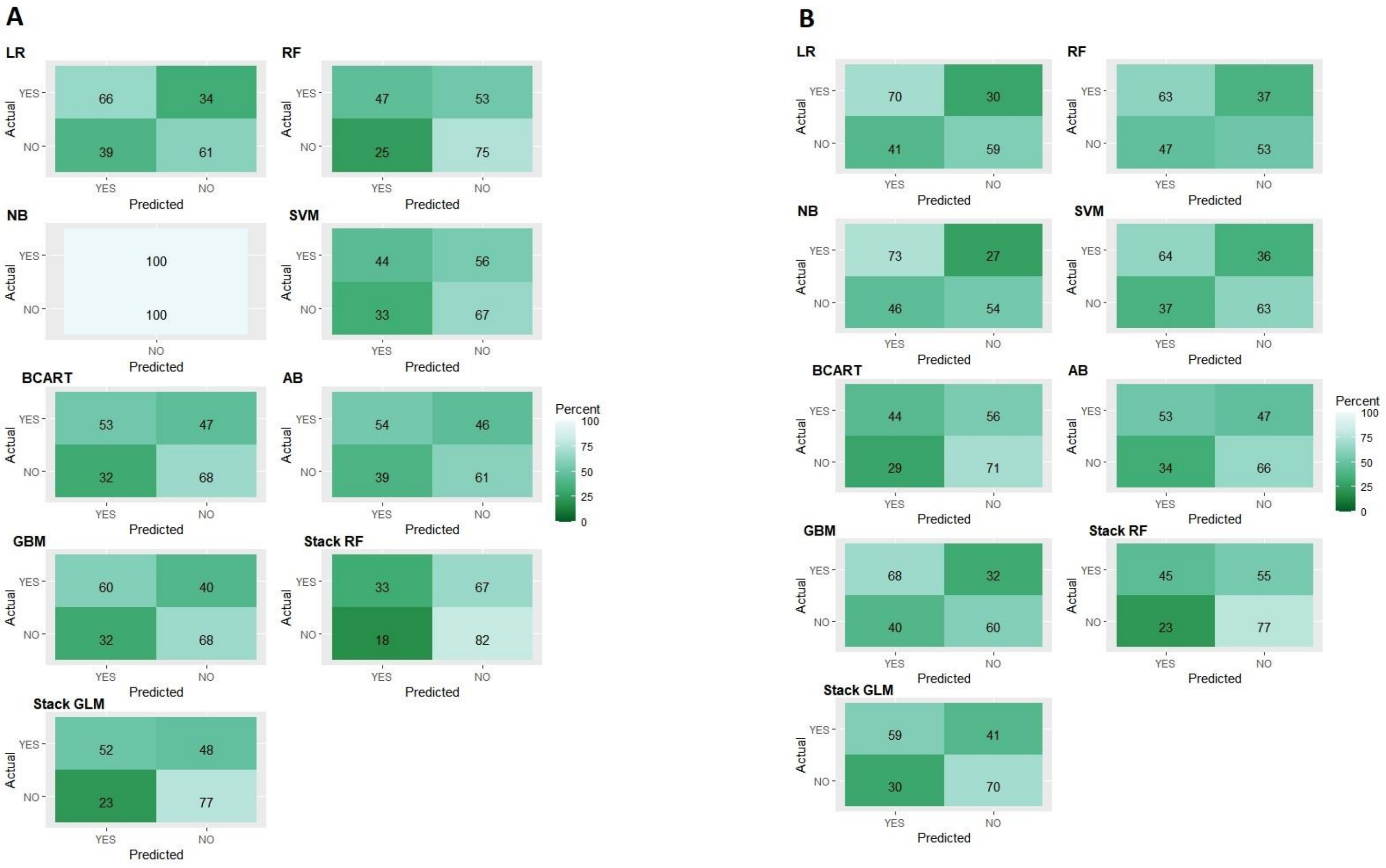


Figure 6.23. Confusion matrices for models predicting postoperative delirium, using all preoperatively available variables and original training data (A) or upsampled training data (B).



6.5. Discussion

6.5.1. Summary of Results

6.5.1.1. Classification Method with the Best Performance when Predicting Outcomes Following Cardiac Surgery, Based on Preoperative Data

When predicting severe postoperative complications, random forest using all preoperative variables (24 variables) and original training data achieved the highest overall performance (AUC = 0.713). Random forest using all variables, but upsampled training data had the highest sensitivity of 0.781, and support vector machine using logistic EuroSCORE variables (15 variables) and original training data had the highest specificity of 0.931. BCART using all variables and upsampled training data had the highest negative predictive value of 0.912.

When predicting AKI, the stacked model using generalised linear model, all variables and upsampled training data had the highest overall performance (AUC = 0.667). Random forest with only significant variables (10 variables) and upsampled training data had the highest sensitivity of 0.746. Support vector machine using significant variables and original training data had the highest specificity of 0.876. The highest negative predictive value belonged to random forest using 10 variables and upsampled training data (NPV = 0.774).

When predicting delirium, logistic regression using all variables and original training data had the highest AUC of 0.675. AdaBoost with 15 variables and original training data had the highest sensitivity of 0.832 and support vector machine using the same data had the highest specificity of 0.900. Random forest with 15 variables and upsampled training data had the highest NPV of 0.869.

In general, the models predicting severe complications and delirium had a slightly better performance than models predicting AKI, however the differences are negligible. None of the models had a particularly outstanding performance, meaning further work is needed to find a better performing model. That being said, the models with high negative predictive values give certainty that if the patient is not predicted to have delirium, then in reality they will not have delirium postoperatively, which can be helpful when making decisions about the patient's treatment plan before surgery.

Based on calibration plots, none of the models had perfect calibration, and all models tended to overestimate risk of the predicted outcome when the true probability was higher than 50%.

6.5.1.2. The Optimal Number of Preoperative Variables for Predicting the Outcomes

As per the results shown in this chapter, the models seemed to have a slightly higher performance when including all preoperative variables (24 variables), compared to when including only significant or logistic EuroSCORE variables. However, the differences in performance were very small, and depended more on the classification method than on the number of variables. Even though the performance measures of the models were modest, knowing that using a smaller number of variables can be beneficial in clinical practice – the fewer variables the model requires, the more user-friendly the model becomes by being quicker to gather the data and to calculate [335].

As shown in Table 6.11, age, procedure and priority were used in all experiments, regardless of whether only significant variables, logistic EuroSCORE variables or all variables were used in the model. Previous cardiac surgery, angina status and preoperative creatinine were also commonly used (all, apart from the significant variables for delirium). Three variables (BMI, smoking status and previous PCI) were not included in logistic EuroSCORE, nor were they significantly associated with any of the three predicted outcomes.

Logistic EuroSCORE variables are likely to be collected at most cardiac centres, especially in the UK, as logistic EuroSCORE is used to audit the performance of cardiac centres [6], [324]. This means that these data are readily available in most cardiac centres, meaning the models using these variables are easier to put into use in practice. There were some variables that were significantly associated with outcomes but were not part of logistic EuroSCORE. These variables were type II diabetes, rhythm, congestive cardiac failure, renal function, left main stem disease and NYHA grade. Even though these are commonly recorded data about cardiac patients, it is evident from the earlier data recorded in CaTHI dataset, that these variables were not recorded very well, especially as seen from Appendix 6.2, that rhythm, left main stem and renal function have several “unknowns”. For the patients recorded in the dataset from 2016, however, there were no “unknowns” for the preoperative renal function.

Table 6.11. Comparison of which variables were included in each model.

Variable **24 variables**	**Logistic EuroSCORE** **15 variables**	**Severe Complications** **12 variables**	**Acute Kidney Injury** **10 variables**	**Delirium** **5 variables**
Age	x	x	x	x
Sex	x	x		
BMI				
Type II Diabetes		x		
Smoking Status				
Procedure	x	x	x	x
Priority	x	x	x	x
Critical Pre-op. State	x			
Previous Cardiac Surgery	x	x	x	
Previous PCI				
Extracardiac Arteriopathy	x		x	
LV Function	x			
NYHA Grade		x		x
Angina Status	x	x	x	
Rhythm		x		
Renal Function			x	x
Preoperative Creatinine	x	x	x	
Neurological Dysfunction	x			
Previous MI	x			
Left Main Stem Disease			x	
Pulmonary Disease	x			
Hypertension History	x		x	
Congestive Cardiac Failure		x		
Active Endocarditis	x	x		

6.5.1.3. The Effect of Balancing Methods on the Performance of Preoperative Models

Through developing predictive models of postoperative outcomes, this chapter was faced with an imbalanced classification problem. The prevalence of severe complications was found to be 5.91%, and the prevalence of AKI and delirium was found to be 18.93% and 12.47%, respectively. Even though, the prevalence of severe complications was especially low, the performance measures for the models appeared to be similar, regardless of whether original or upsampled training data were used. This result has also been shown elsewhere [135], and shows that the classification methods are robust enough to handle class imbalance, as shown in the literature [147], [148] . In addition, when using performance measures, such as AUC, sensitivity, specificity, PPV and NPV, the accuracy paradox is avoided, meaning that class imbalance is taken into account by these performance measures [136]. In general, upsampling is not widely encouraged when developing clinical prediction models because the classification threshold for predicted probabilities can be misleading [135]. Upsampling can lead to poor calibration, meaning that the model can either over- or underestimate patients’ risk for the predicted outcome, if applied to a population that the model was not trained on [135]. This is also evident in this chapter, where the model with the highest discriminatory ability predicting acute kidney injury showed very poor discrimination, especially for patients who had true probability for AKI above 50%. Hence, as the results were similar to the models using original training data, in the next chapters, balancing methods will not be used.

6.5.2. Comparison with the Literature

6.5.2.1. Prediction of Severe Postoperative Complications

There are various papers developing prediction models for certain postoperative complications, such as renal complications [336], postoperative bleeding [337], and cardiac complications [338], to name a few, however the commonly used preoperative risk stratification tools were developed, and are still used, to predominantly predict mortality [14]. These known risk prediction tools include logistic EuroSCORE, [164] EuroSCORE II [193], the Initial Parsonnet Score [214], the Society of Thoracic Surgeons score [329], [339], and the Cleveland Clinic Score [328]. The first three were developed to predict 30-day mortality, and the latter two were developed to predict mortality as well as some complications. Even though these models are used to mostly predict mortality, some studies have assessed the use of these scoring systems to predict combinations of postoperative complications [340]–[347]. The findings of these studies were discussed in previous work as part of the MPhil [17], however the AUC ranged from 0.590 to 0.730.

These studies should be compared with the study undertaken in this chapter with caution. Firstly, all of these studies predict different complications as a combined outcome, ranging from prolonged hospital stay to stroke, and from a combination of four complications to a combination of twelve complications. As stated in previous work, the rationale behind choosing these particular complications as the predicted outcome is unclear and subjective, as studies have a different definition for "morbidity". In this chapter, severe postoperative complications were defined based on the results of Chapter 4, where postoperative complications in cardiac surgery were defined and classified, using the Delphi method. Hence, the outcome of "severe complications" consisted of reported complications that fitted the criteria of a "severe" complication based on the Chapter 4. Having a specific criterion to group complications makes studies more comparable and objective.

Furthermore, all these aforementioned models were developed using logistic regression. Even though logistic regression has been shown to perform competitively, compared to various machine learning methods [150]–[152], the findings from this chapter show that logistic regression has the 15th place in terms of overall performance (AUC = 0.670), whereas random forest achieved the top performance of AUC = 0.713, followed by AdaBoost (AUC = 0.706) and BCART (AUC = 0.704). These results show that, even though logistic regression can have a competitive performance, it is still important to test different classification methods to achieve the best possible results.

6.5.2.2. Prediction of Acute Kidney Injury

The risk factors of female sex, advanced age, renal impairment, previous cardiac surgery, pulmonary problems, diabetes, hypertension history, congestive cardiac failure, reduced left ventricular function are known risk factors for AKI following cardiac surgery [348]. Also, patients undergoing emergency surgery [349] and more complicated surgical procedures than CABG [281], [350] are also associated with AKI following cardiac surgery. The variables found to be associated with AKI in this chapter are in accordance with these findings.

There are other risk prediction models developed to predict AKI based on information about patient demographics and co-morbidities. Well-known validated models are the

Cleveland Clinic Score [351], Mehta score [352] and Simplified Renal Index Score [353]. These three scores, however, predict AKI specifically requiring dialysis, which is considerably rare (1%-7% of patients [354]), and hence the usability of these prediction scores can be limited.

A newer prediction model, developed by Birnie et al., however uses KDIGO guidelines to predict different stages of AKI in a cardiac surgery population [279]. The model by Birnie et al. achieved AUC of 0.74 when predicting AKI in general, which is noticeably higher than the highest AUC achieved to predict AKI in this chapter (AUC = 0.667, Stack GLM, 24 variables, upsampled training data). The study by Birnie et al, however, was a multi-centre study (3 hospitals), using just under 40,000 patient records. In their data, in addition to the variables reported in the CaTHI database, haemoglobin, glomerular filtration rate, administration of heparin or nitrates and catheter to surgery were also reported [279]. These variables are shown to be associated with AKI [279], however were not available in the CaTHI dataset, explaining why Birnie et al. achieved a higher performance.

There are also other prediction models developed for AKI following cardiac surgery [355]–[357], but the number of models developed that use preoperative data only (no intraoperative or postoperative data) is limited. The reason why only preoperative data were included in this chapter, is that knowing the risk of AKI before surgery, in preoperative clinic could help clinicians to make an informed decision about the treatments, and also to communicate risk to the patient appropriately.

6.5.2.3. Prediction of Delirium

Delirium has been attempted to be predicted for decades [358]. Various studies have been undertaken to understand the risk factors associated with delirium [359]–[365], however only one study at the time of writing had been published that attempt predicting postoperative delirium, using preoperative data only [331]. Studies often include pre- and intraoperative data, including types of anaesthesia and length of surgery [359], [361], which could be beneficial when predicting delirium once the patient is admitted to the ICU. However, in order to make decisions about the medicines given (e.g., preoperative administration of beta blockers, statins [366] and benzodiazepine [367]) before, during and right after the surgery, predicting delirium, using preoperatively available variables could be beneficial. This could be especially useful in cardiac surgery patients, where the prevalence of delirium in this study was reported 12.47%, but has been reported to be as high as 52% [286].

Rudolph et al.'s preoperative prediction rule for delirium following cardiac surgery achieved AUC of 0.750 [331], which is considerably higher than the highest overall performance in this chapter (AUC = 0.675). However, it is worth noting that the methods of participant recruitment, data collection and analysis are very different from the study undertaken in this chapter, making it difficult to compare Rudolph et al.'s results with those found in this chapter. Firstly, Rudolph et al. carried out a prospective study, where the derivation and validation cohort patients were recruited prospectively. This means that the data that were collected was more likely to be relevant to delirium, unlike the data used in this study, which is commonly collected information in preoperative assessment clinic and auditing database. In addition, Rudolph et al.'s patient cohort had a very large number of patients with delirium - 52% in derivation and 44% in validation cohort – which is a considerably higher prevalence than it was in this chapter's patient population. The high prevalence in Rudolph et al's work with prospective data collection can also indicate some possibility for selection bias. Finally, Rudolph et al. used multiple imputation in both training and testing data, which can affect the overall results and applicability of their model in practice [368].

In this study, five preoperatively available variables were found to be significantly associated with postoperative delirium: *age, surgical procedure, preoperative renal function, surgical priority,* and *NYHA grade*. Age is a common risk factor for delirium [369], especially affecting patients who are over the age of 65 years [285]. Since most patients in this patient population are older than 65, age is an expected risk factor. The type of surgery has been shown to be associated with postoperative delirium in other studies [366], [370]. Surgical priority also matters if a patient is likely to have delirium, especially if undergoing emergency surgery [371], [372]. Abnormal preoperative renal function has also been shown to be associated with delirium elsewhere [367]. NYHA grade has been also shown to be associated with delirium in other studies [373], [374].

Other studies assessing delirium risk factors have also included information on patient's mental health, such as whether a patient has depression [366], [375], [376] or is on psychoactive medications [377]. This information is currently not recorded in the CaTHI database.

6.5.3. Limitations

Even though the predicted outcome “severe” postoperative complications was defined based on the study shown in Chapter 4, there are limitations associated with this outcome due to the reporting of complications in medical databases. As mentioned in Section 6.3, 5.22% of patients were reported to have renal complications in CaTHI database, however, when using the KDIGO classification to diagnose AKI alone, 18.93% of the patient population had postoperative AKI. This shows that renal complications are heavily underreported in the CaTHI database. In addition, as seen from Appendix 6.1, delirium has been reported for only 1 patient in the CaTHI database (0.01%). This is because in the CaTHI database delirium as a postoperative complication is not generally reported. This is a strong limitation to the “severe” postoperative complications as the outcome and shows that the reporting of postoperative complications needs to be improved for future analysis of complications. As stated in Chapter 4 “*If we do not know what is going wrong, we cannot work out how to stop it happening*”, meaning that audit databases especially, such as CaTHI should include higher quality reporting of complications to help research complications, and therefore mitigate the risks of adverse outcomes and to treat and recognise the complications early.

A reason why the models had only moderate performance when predicting the three outcomes could be missing data. A strategy to avoid AKI before surgery is to optimise preoperative haemoglobin levels. Low preoperative haemoglobin has been shown to be associated with postoperative AKI [378]. A limitation in this study, however, is that preoperative haemoglobin has not been collected very well in our patient population, and due to a large number of missing values, was removed from the analysis in this chapter.

Another variable that had many “unknowns” in patients who had surgery in earlier years (2012-2015), is renal impairment. It has been shown that knowing renal function preoperatively is an indicator whether a patient has postoperative acute kidney injury

[379]. However, since delirium analysis was done using a slightly newer dataset (2016-2018), there are no "unknowns" in the preoperative renal function variable. This shows that since 2016, the recording of data has slightly improved.

6.6. Conclusion

Overall, the performance of the models was very similar, however neither of these had remarkably high performance, most probably due to delirium and AKI being difficult to predict based on the available preoperative data. There was no notable difference in performance based on whether the models were developed using the original training data or upsampled training data. This shows that the classification methods were quite robust at handling imbalanced classification problems. Hence, going forward, only the original training data, without any balancing methods, are used in Chapters 7 and 8.

To improve the models' predictive ability for AKI and delirium, more granular data about the patient's condition, i.e., laboratory variables measured in the ICU, are included to predict AKI (Chapter 7) and delirium (Chapter 8).

Chapter 7. Study 2.1: Predicting the Onset of Acute Kidney Injury on an Hourly Basis in Intensive Care

7.1. Introduction

This chapter aims to carry out experiments to predict acute kidney injury (AKI) following cardiac surgery within 25 hours since ICU admission on an hourly basis, using static preoperatively recorded and dynamic ICU data.

In this chapter a binary classification problem is investigated: “Does a patient have an AKI or not?”

Three main experiments were undertaken to find the most optimal model for predicting AKI:

1. Predicting AKI using complete data only.
2. Predicting AKI, using complete training data and missing values in testing data. Patient records missing more than 40% of the variables were excluded from the analysis, as done elsewhere [300], [302].
3. Predicting AKI, using complete training data and imputation methods to replace missing values in testing data. Again, patient records with more than 40% of the variables missing were excluded from the analysis.

All model development methods are described in detail in Chapter 5.

7.2. Related Work

Even though AKI is a persistent and wide-spread problem in cardiac surgery, there are numerous preoperative prediction models for AKI [281], however not many dynamic models, suitable for ICU use, have been developed. In the literature review, presented in Chapter 2, only one paper developing dynamic classification prediction models for renal complications was found – paper by Meyer et al. [74]

Meyer et al. predicted various postoperative complications, such as postoperative bleeding, renal failure requiring renal replacement therapy and in-hospital mortality. The study achieved AUC of 0.96, sensitivity of 0.94 and negative predictive value of 0.94 when predicting renal failure in the original cohort. In external validation on MIMIC-III dataset, the study achieved AUC of 0.91, Sensitivity of 0.79 and NPV of 0.80. It is also important to note that the study used a balanced dataset for both developing and testing their model [74], which was further discussed in Chapter 2.

Overall, the study used a very large sample size: 47,559 admissions. Even though Meyer et al. used the KDIGO guidelines to define renal failure, the study lacked in specific information about when AKI occurred. If the time of complication occurrence is unknown, there is a high risk that the complication that is predicted has already happened, meaning the prediction task undertaken is irrelevant. This issue is especially likely to happen in Meyer et al.'s paper where AKI is predicted within the first 24 hours of ICU admission. It is, however, unrealistic that all patients have AKI occurring in the same time window.

Therefore, to improve upon the currently existing dynamic classification prediction model for AKI, the following was undertaken in this chapter:

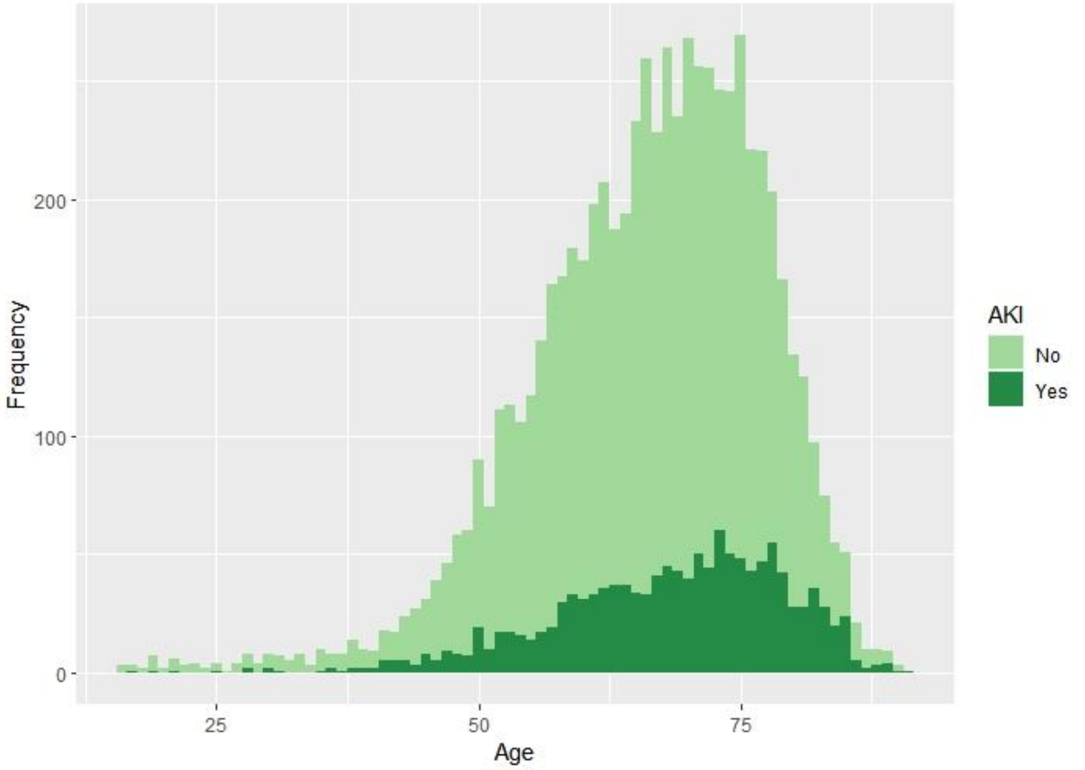
- AKI was defined using KDIGO criteria, using preoperative creatinine as the baseline and subsequent creatinine measurements taken in ICU.
- Experiments were undertaken to predict AKI on an hourly basis within 25 hours since ICU admission.
- Experiments were undertaken using complete data, missing data and imputation methods to achieve the best possible prediction model to predict AKI.

7.3. Patient Population and Acute Kidney Injury

7.3.1. Patient Demographics

Of all patients included in the analysis, 750 had AKI, with the prevalence of 11.92% (95% CI 11.14-12.74%). As shown in Appendix 7.1, Table 7.1.A, overall, the mean age for the total population was 66.09 years, the majority being male (72.34%). The most common procedure was coronary artery bypass graft (CABG) (56.93%). The mean logistic EuroSCORE for patients was 5.16. Of all patients, 7.10% had a severe renal impairment and 28.28% had moderate renal impairment before surgery. For 23.68% of the patients the renal function was not recorded at the preoperative clinic. The mean preoperative creatinine was 91.18, which is in the higher side of the normal range. The mean hospital stay was 11.20 days, and the mean ICU hours was 44.05. Overall, 0.91% of the patients died in the hospital.

Figure 7.1. Histogram of age distribution among patients with no acute kidney injury (light-green) vs with acute kidney injury (dark-green).



When comparing the patients with AKI and patients without AKI, these two groups of patients were significantly different from one another based on the variables shown in Table 7.1.A. The patients who had AKI in the ICU had a higher mean age (Figure 7.1) (69.52 years vs 65.74 years), a higher proportion was female (33.73% vs 26.84%) and a higher proportion of patients had a more complicated surgery (40.13% vs 28.14% for valve surgery and 20.80% vs 12.52% for combined CABG and valve surgery) than patients who did not develop AKI in the ICU. Interestingly, patients with AKI postoperatively had a higher mean logistic EuroSCORE (Figure 7.2) (8.10 vs 4.76), but a higher proportion of patients had preoperative renal complications than patients with no AKI (16.80% vs 5.79% with severe preoperative renal function).

Patients with AKI had higher mean preoperative creatinine levels (Figure 7.3) (101.58 vs 89.77), and stayed in the ICU for approximately 3 days longer (Figure 7.4) (112.26 hours vs 34.82) and hence also stayed in the hospital for longer (Figure 7.5) (16.12 days vs 10.54 days) than patients without AKI. Almost 4% more patients died in hospital if they had AKI than if they did not (4.40% died vs 0.43% died).

Figure 7.2. Histogram of logistic EuroSCORE distribution among patients without acute kidney injury (light-green) vs with acute kidney injury (dark-green).

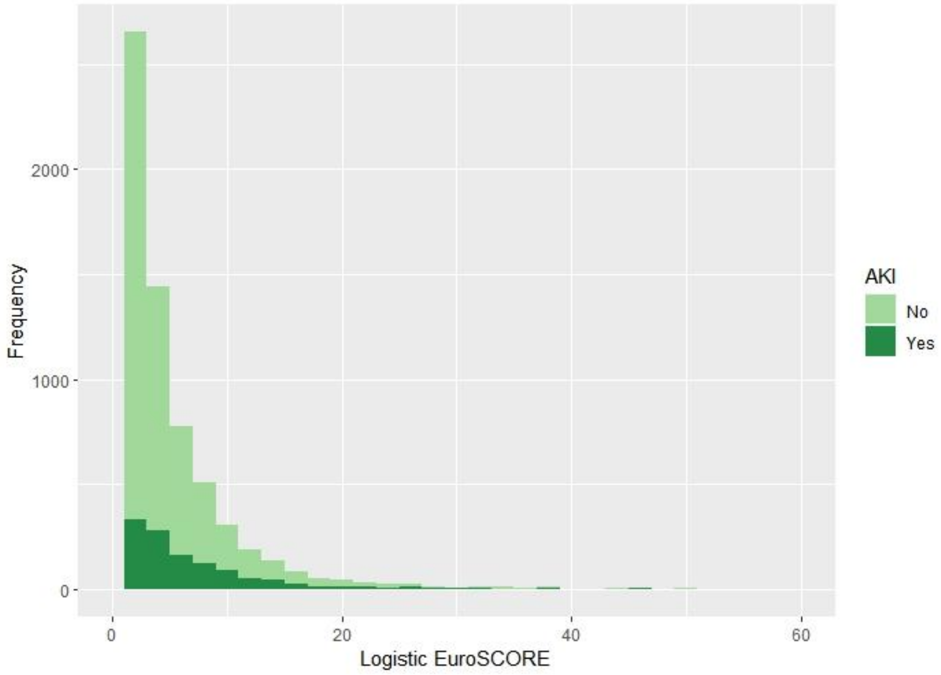


Figure 7.3. Histogram of preoperative creatinine distribution among patients without acute kidney injury (light-green) vs with acute kidney injury (dark-green).

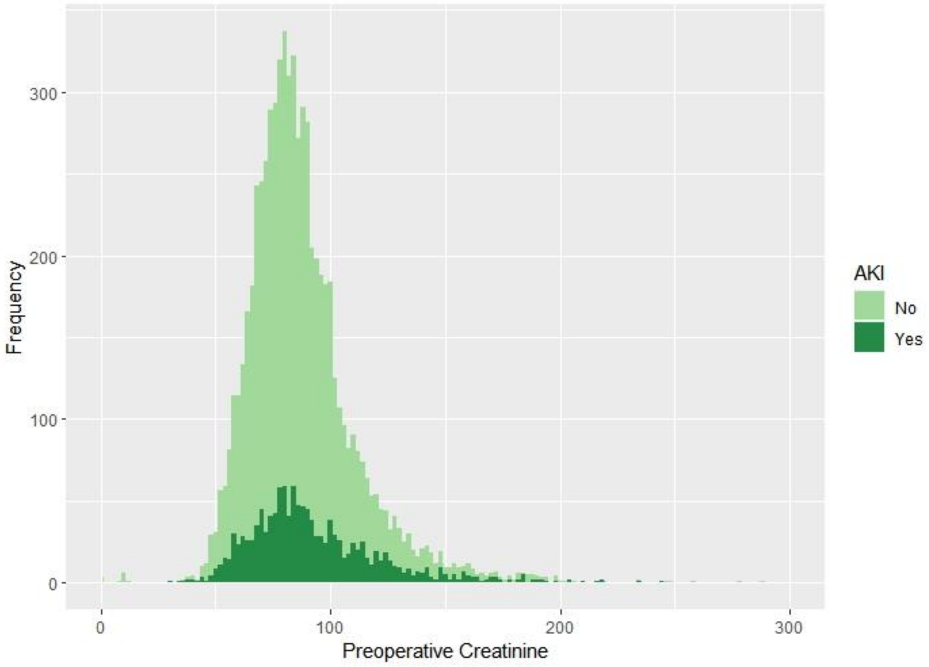


Figure 7.4. Histogram of hours spent in intensive care unit for patients without acute kidney injury (light-green) vs with acute kidney injury (dark-green).

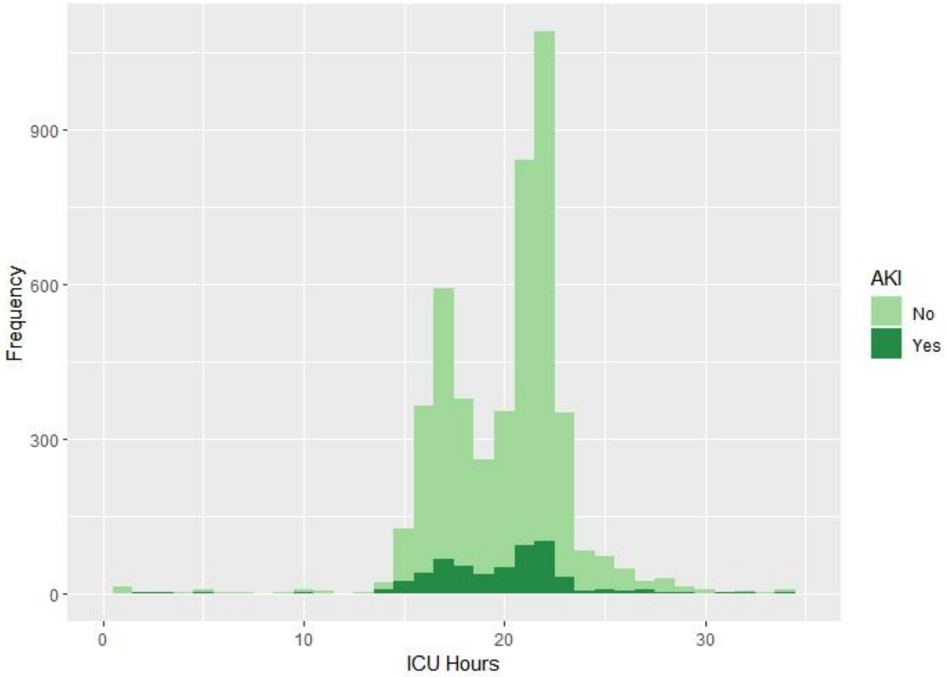


Figure 7.5. Histogram of total days spent in hospital for patients without acute kidney injury (light-green) vs with acute kidney injury (dark-green).

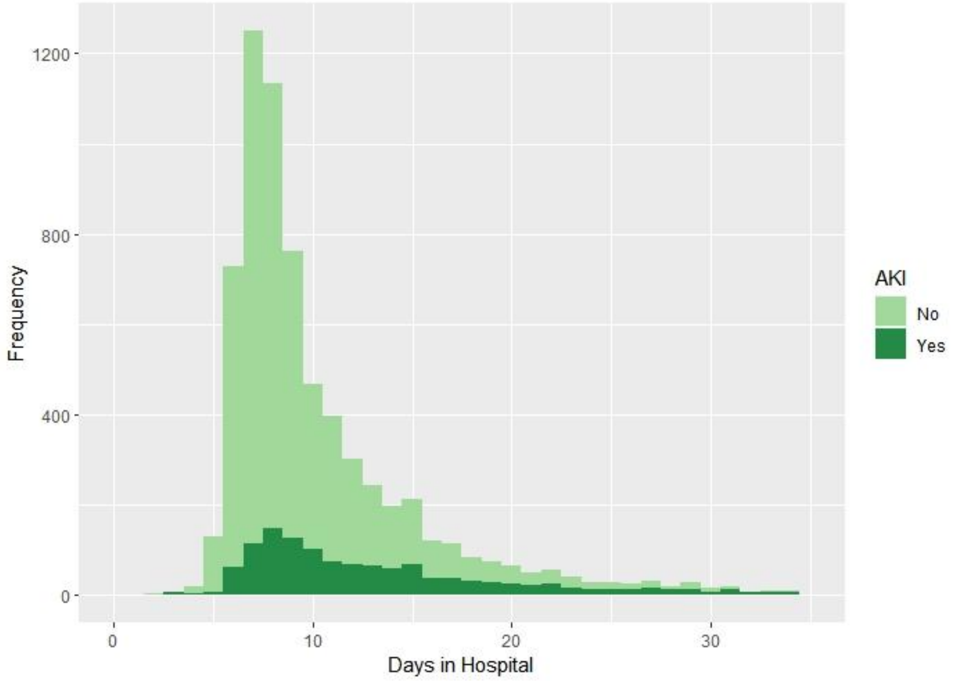
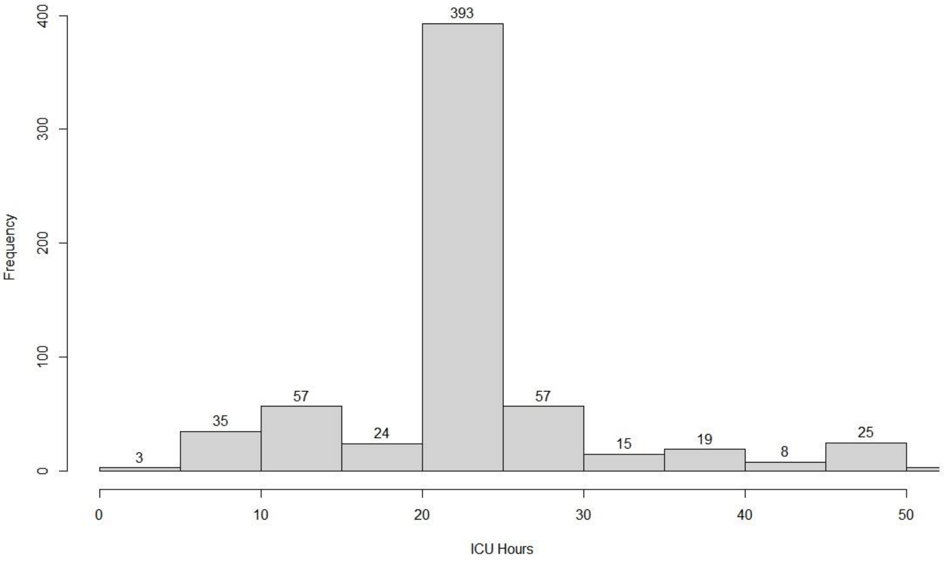


Figure 7.6. Histogram of the time when acute kidney injury occurs in ICU based on KDIGO criteria.



As shown in Figure 7.6, most patients had AKI between 20 and 25 hours since ICU admission, more specifically at median hours of 24.28, IQR 4.21. The reason for this timing of AKI was most likely due to when creatinine was measured in ICU, as shown in Chapter 5, which was in median of 23.67 hours since ICU admission. Hence, the creatinine changes were also captured in these median hours, showing the majority of patients to have AKI at this time window.

7.3.2. Descriptive Statistics

7.3.2.1. Preoperative Data

As shown in Appendix 7.1, Table 7.1.B, of all patients, the majority had an elective surgery (64.08%). Only 1.02% of patients had an emergency surgery. A very small number of patients were in a critical preoperative state (1.75%) and had had a previous cardiac surgery (2.56%). Slightly over a tenth of patients had had a previous percutaneous coronary intervention (13.35%) and had extracardiac arteriopathy (11.49%). About a fifth of the patients had either a moderate (17.43%) or poor (2.83%) left ventricular function. In terms of rhythm, 11.38% of the patients had an abnormal rhythm and for 5.78% it was unknown. Only 1.38% of the patients had a neurological dysfunction before surgery. Overall, 37.05% of the patients had had a previous myocardial infarction. Just over a tenth of the patients had a recorded left main stem disease (13.98%). For 38.91% of the patients, however, the left main stem status is unknown. Pulmonary disease was recorded for 16.40% of the patients. The large majority of patients (72.18%) had a hypertension history. Just under ten percent of the patients (9.72%) had congestive cardiac failure and only 1.10% of the patients had active endocarditis before surgery.

The table also shows these aforementioned characteristics for patients based on whether they had postoperative acute kidney injury. There was a statistically significant difference in these two patient populations preoperatively based on all variables, apart from previous percutaneous coronary intervention, neurological dysfunction, previous myocardial infarction and pulmonary disease. The population with postoperative AKI had higher amounts of emergency surgery (2.93% vs 0.76%) than patients without AKI. More patients who developed AKI had also factors that increased the overall surgical risk, such as critical preoperative state (3.47% vs 1.52%), previous cardiac surgery (5.07% vs 2.22%), extracardiac arteriopathy (14.67% vs 11.06%), poor left ventricular function (4.40% vs 2.62%), higher NYHA grade (7.47% vs 2.83% for grade IV), abnormal rhythm (18.00% vs 10.48%), hypertension history

(76.67% vs 71.57%), congestive cardiac failure (17.07% vs 8.73%) and active endocarditis (2.53% vs 0.90%). Interestingly, the patient population with postoperative AKI had better preoperative angina status and left main stem status than patients without AKI.

7.3.2.2. Laboratory Data

The Appendix 7.1, Table 7.1.C shows overall mean, standard deviation, median and range for each laboratory variable. The p-values are calculated based on t-tests for numerical variables and Chi-Squared test of independence for categorical variables (medicines only).

All variables, apart from dopamine have statistically significantly different levels for patients with AKI and without AKI. When looking at the dopamine dose, there is no statistically significant difference between patients with and without AKI. The literature shows dopamine administration to be controversial. Historically, dopamine has been used as a vasopressor to avoid AKI, and many countries still use it in practice. However, in the early 2000's, the literature has shown that dopamine does not benefit in terms of preventing mortality or dialysis [380], [381] and the evidence of the benefits of dopamine is inconclusive [382].

Patients with postoperative AKI had notably lower mean arterial base excess (-0.43 vs -0.20) than patients without AKI. They also had lower arterial haematocrit (27.71 vs 29.34), daily fluid balance (255.50 vs 322.60), haemoglobin (94.19 vs 99.64) and urine output (85.91 vs 97.50) levels. The patients with AKI had clearly higher mean C-reactive protein (140.40 vs 135.00), creatinine (143.60 vs 89.54), and urea (13.37 vs 6.78) levels.

7.4. Experiment 1 Results: Models Predicting Acute Kidney Injury in ICU on an Hourly Basis, Using Complete Data

7.4.1. Data Preparation

As shown in Table 7.1, for each lead time for AKI within 25 hours since ICU admission, the number of patients varied between 3606 and 3723, with the mean number of patients across each lead time of 3676 (SD = 38.43). The prevalence of AKI was considerably low, staying between 6.49% and 8.97%, (mean = 8.02%, SD = 0.84) with more patients with AKI at lead times closer to the event and less patients further away from the event. This was because, as shown in Section 7.3.1, most patients had AKI between 20-25 hours since ICU admission. More specifically, the median time of AKI occurrence was 24.28 hours since the admission to ICU.

As explained in the Methods chapter (Chapter 5), the training set consisted of 2/3 of the included patient records, and the testing data consisted of 1/3 of the records. This is reflected in Table 7.1. If the patient did not have AKI at all during the ICU stay, their time window was marked as the time window of the experiment. This means that in this experiment, these patients had the end time window marked as 25 hours.

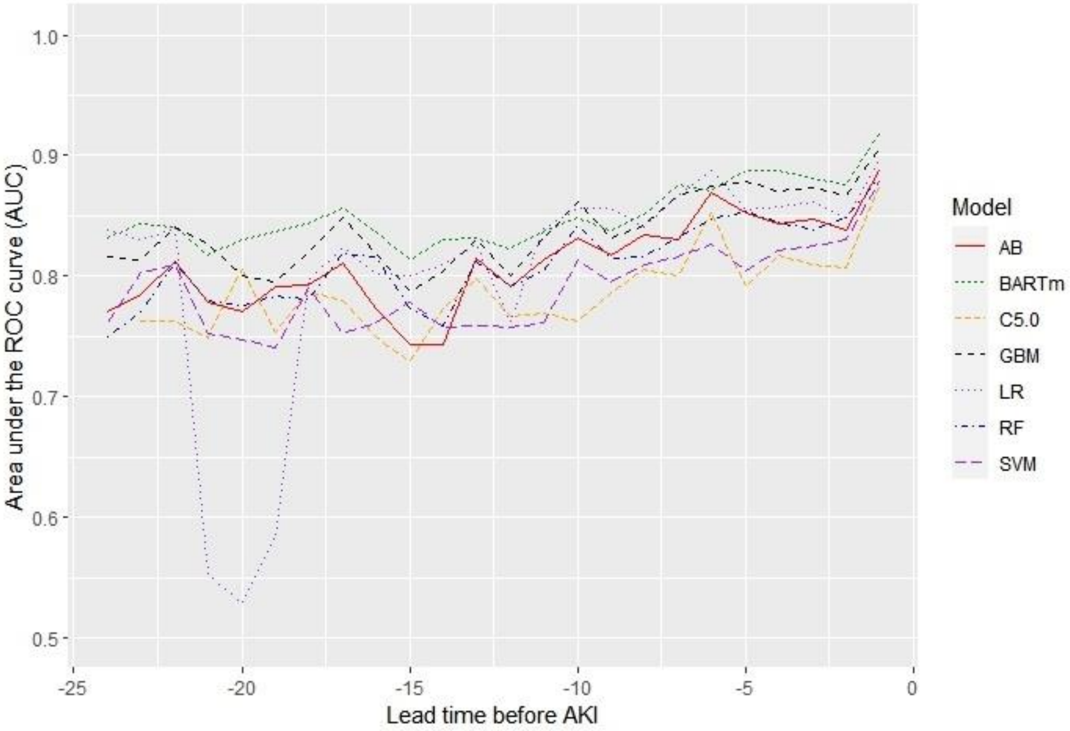
Table 7.1. Number of patients and proportion of patients with AKI in each training and testing data, depending on lead time, if predicting AKI within 25 hours of ICU stay.

Lead Time	Total Data Number of Patients	AKI (%)	Training Data Number of Patients	AKI (%)	Testing Data Number of Patients	AKI (%)
-24	3606	6.49	2417	6.33	1189	6.81
-23	3622	6.79	2427	6.88	1195	6.61
-22	3628	6.95	2431	7.03	1197	6.77
-21	3630	7.00	2433	6.86	1197	7.27
-20	3636	7.12	2437	7.18	1199	7.01
-19	3638	7.17	2438	6.97	1200	7.58
-18	3640	7.22	2439	7.18	1201	7.33
-17	3645	7.35	2443	7.29	1202	7.49
-16	3648	7.43	2445	7.69	1203	6.90
-15	3656	7.63	2450	7.55	1206	7.79
-14	3665	7.86	2456	7.9	1209	7.78
-13	3674	8.03	2462	7.88	1212	8.33
-12	3683	8.20	2468	8.14	1215	8.31
-11	3693	8.42	2475	8.36	1218	8.54
-10	3703	8.67	2482	8.78	1221	8.44
-9	3708	8.74	2485	9.01	1223	8.18
-8	3712	8.84	2488	8.68	1224	9.15
-7	3718	8.93	2492	8.87	1226	9.05
-6	3718	8.93	2492	8.87	1226	9.05
-5	3719	8.95	2492	8.79	1227	9.29
-4	3719	8.95	2492	8.79	1227	9.29
-3	3719	8.95	2492	8.79	1227	9.29
-2	3720	8.98	2493	8.78	1227	9.37
-1	3723	8.97	2495	9.26	1228	8.39
Mean ± SD	3676 ± 38.43	8.02 ± 0.84	2464 ± 25.75	7.99 ± 0.85	1212 ± 12.69	8.08 ± 0.91

7.4.2. Models’ Discriminative Performance

As seen from Figure 7.7, most models, apart from logistic regression at lead times -21 to -19, had a similar pattern of how the overall performance changed as the lead time changed. All models tended to have a slightly better performance as the lead time got closer to AKI. From the figure it can also be seen that BARTm had the highest performance at most times. The differences between the models’ overall performance, however, did not seem to be visually very large. The models had a relatively good performance, staying above 0.750 at most lead times.

Figure 7.7. AUC for each model for each lead time when predicting acute kidney injury on an hourly basis, using complete data.



The Table 7.2 shows the mean and standard deviation performance measures across each lead time for each model. BARTm had the highest overall performance based on mean AUC of 0.850. Most models, apart from C5.0 and SVM had a mean AUC higher than 0.800, meaning most models had moderately high overall performance.

In terms of sensitivity, BARTm also had the highest mean sensitivity across all lead times (Sens = 0.821). Logistic regression had the highest mean specificity of 0.824. All models had very low positive predictive values, which was expected due to the relatively low proportion of patients having AKI in each lead time. Negative predictive values, however, were moderately high, the mean NPV staying above 0.700 for all models. C5.0 and random forest had the highest mean NPV of 0.793 across all lead times.

The sensitivity, specificity, positive and negative predictive values for each lead time are visualised in Figures 7.8 and 7.9, respectively. The changes in sensitivity and specificity for models were quite volatile as the lead times changed. The positive and negative predictive values were quite similar for all models. Interestingly, logistic regression had slightly higher positive predictive values than other models at lead times -20 and -10.

Table 7.2. Mean and standard deviation model performance measures for each model across each lead time before delirium when predicting AKI within 25h since ICU admission, using complete data. The highest result for each performance measure is marked in bold.

Model	AUC Mean ± SD	Sensitivity Mean ± SD	Specificity Mean ± SD	PPV Mean ± SD	NPV Mean ± SD
AB	0.810 ± 0.037	0.752 ± 0.072	0.745 ± 0.069	0.028 ± 0.005	0.786 ± 0.052
BARTm	**0.850 ± 0.026**	**0.821 ± 0.053**	0.741 ± 0.057	0.021 ± 0.006	0.775 ± 0.054
C5.0	0.787 ± 0.034	0.742 ± 0.052	0.738 ± 0.058	0.030 ± 0.005	**0.793 ± 0.045**
GBM	0.838 ± 0.031	0.786 ± 0.067	0.755 ± 0.072	0.024 ± 0.006	0.770 ± 0.056
LR	0.802 ± 0.100	0.668 ± 0.216	**0.824 ± 0.080**	**0.038 ± 0.037**	0.742 ± 0.076
RF	0.810 ± 0.034	0.754 ± 0.079	0.739 ± 0.057	0.034 ± 0.028	**0.793 ± 0.040**
SVM	0.790 ± 0.035	0.719 ± 0.071	0.761 ± 0.071	0.031 ± 0.006	0.781 ± 0.054

Figure 7.8. Sensitivity and specificity, positive and negative predictive values (PPV and NPV) for all models for each lead time when predicting acute kidney injury on an hourly basis in the ICU, using complete data.

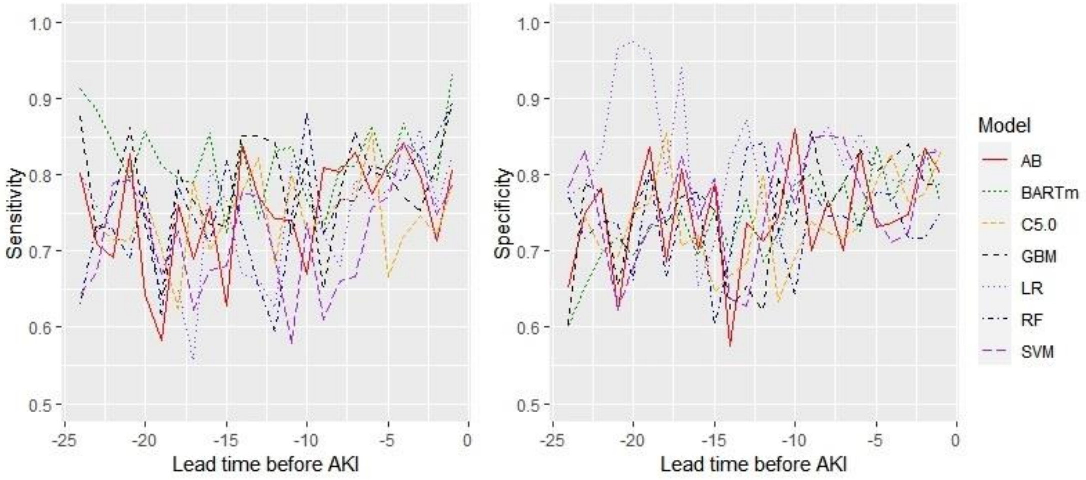
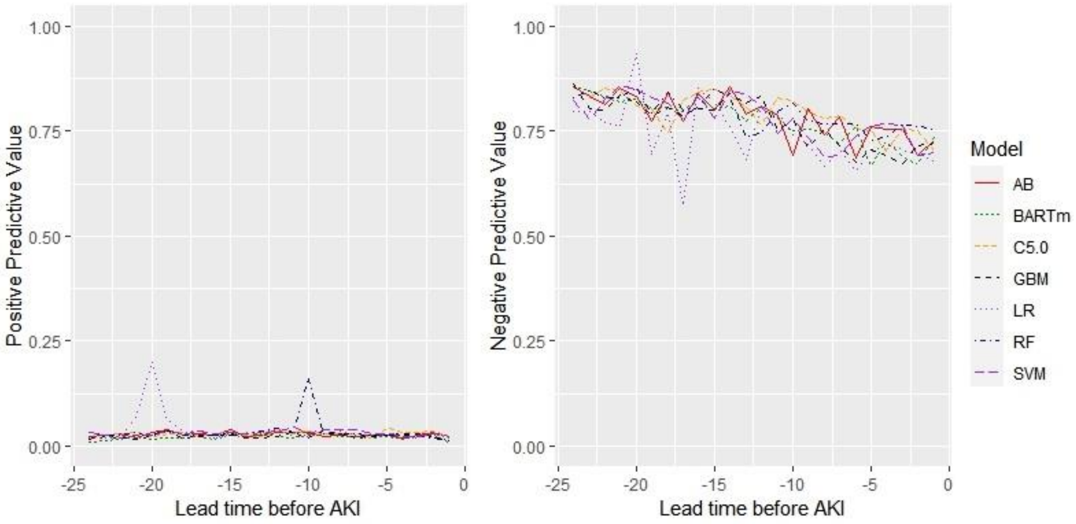


Figure 7.9. Positive and negative predictive values (PPV and NPV) for all models for each lead time when predicting acute kidney injury on an hourly basis in the ICU, using complete data.

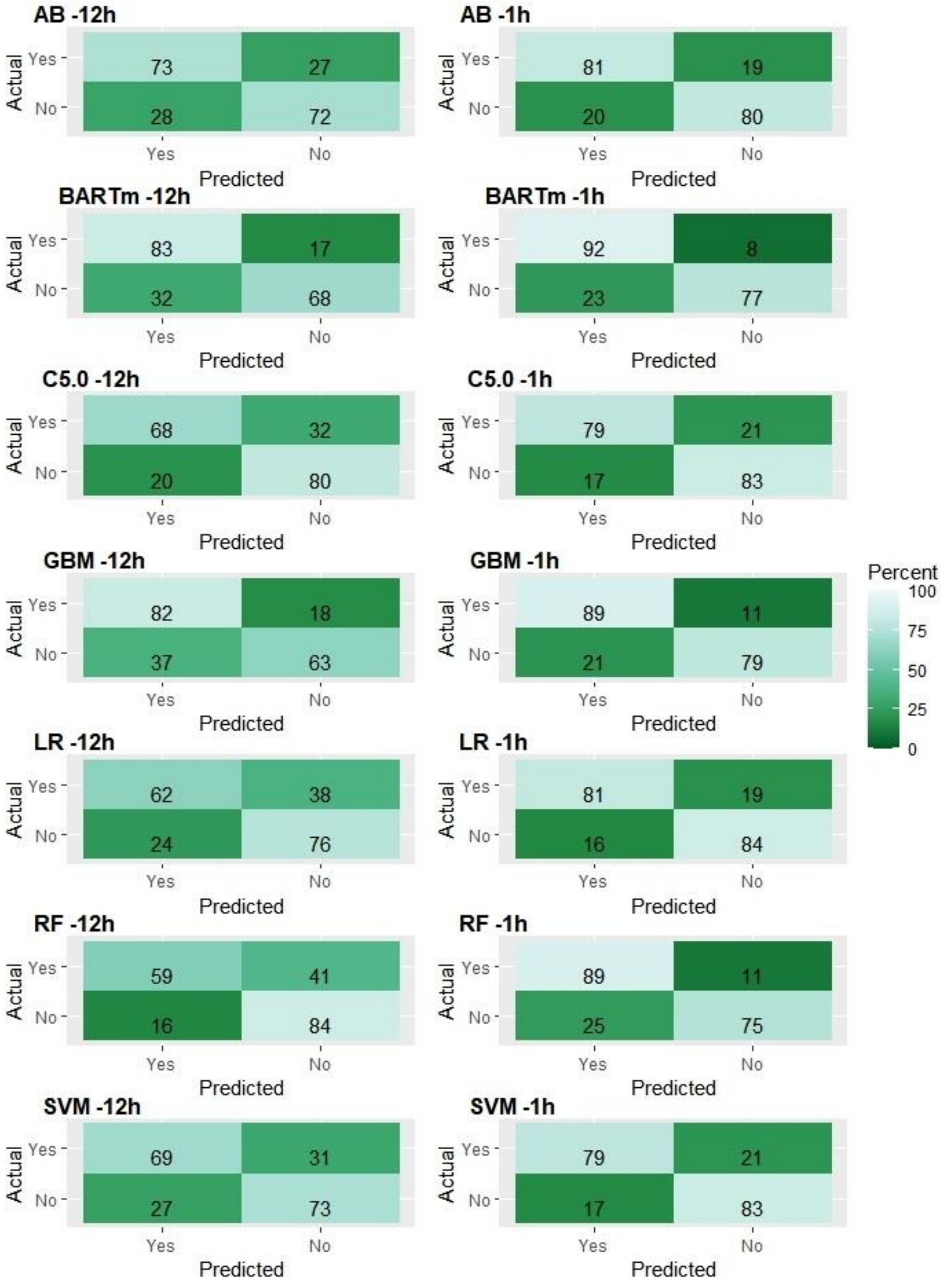


The exact performance measures for each lead time and their respective cut-off values can be found from Appendix 7.3. The Table 7.3.A in Appendix 7.3 shows that when predicting AKI 24 hours in advance, logistic regression had the highest AUC of 0.838. BARTm had the highest sensitivity of 0.914 and support vector machine had the highest specificity of 0.783.

At 12 hours before AKI, BARTm had the highest AUC of 0.823. Gradient boosting model had the highest sensitivity of 0.842 and random forest had the highest specificity of 0.843. When predicting AKI 4 hours in advance, BARTm again had the highest AUC of 0.887, and also the highest sensitivity of 0.868. C5.0, however, had the highest specificity if 0.827.

When looking at the confusion matrices (Figure 7.10), all models were better at predicting AKI 1 hour in advance than 12 hours in advance. Random forest had the worst performance, predicting only 59% of the cases of AKI correctly at 12 hours before AKI. BARTm, however, had particularly good performance at 1 hour before AKI as it predicted AKI correctly for 92% of the cases.

Figure 7.10. Confusion matrices for all models predicting acute kidney injury 12 hours in advance and 1 hour in advance, using complete data.



7.4.3. Models' Calibration and Variable Importance

As BARTm model using complete data achieved the highest mean AUC of 0.850 across all lead times, calibration for this model for -24-, -12- and -1-hour lead times is shown (Figure 7.11). The model had a better calibration when it predicted AKI fewer hours in advance, i.e., the calibration was overall better when predicting AKI 1 hour in advance, as opposed to 24 hours in advance. That being said, for all three cases, the predicted probability estimations were more certain for patients who had lower true probability of having AKI. The certainty, as shown by confidence intervals, was especially low when predicting AKI 24 hours in advance due to small number of patients having had AKI at this lead time.

According to the BARTm model, the mean predicted probabilities (Table 7.3) were in general slightly lower than the proportion of patients with AKI in each respective lead time dataset. This means that in general, the model slightly underestimated the risk of patients having AKI. The difference in the mean predicted probability and proportion of AKI was especially noticeable when predicting AKI 12 hours in advance. The table also shows that the proportion of patients with AKI was considerably lower in -24-hour lead time, which explains the higher uncertainty in predicted probabilities in Figure 7.11 at this lead time.

Figure 7.11. Calibration plots for BARTm model predicting acute kidney injury 1 hours (top-left), 12 hours (top-right), and 24 hours (bottom-centre) in advance. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.

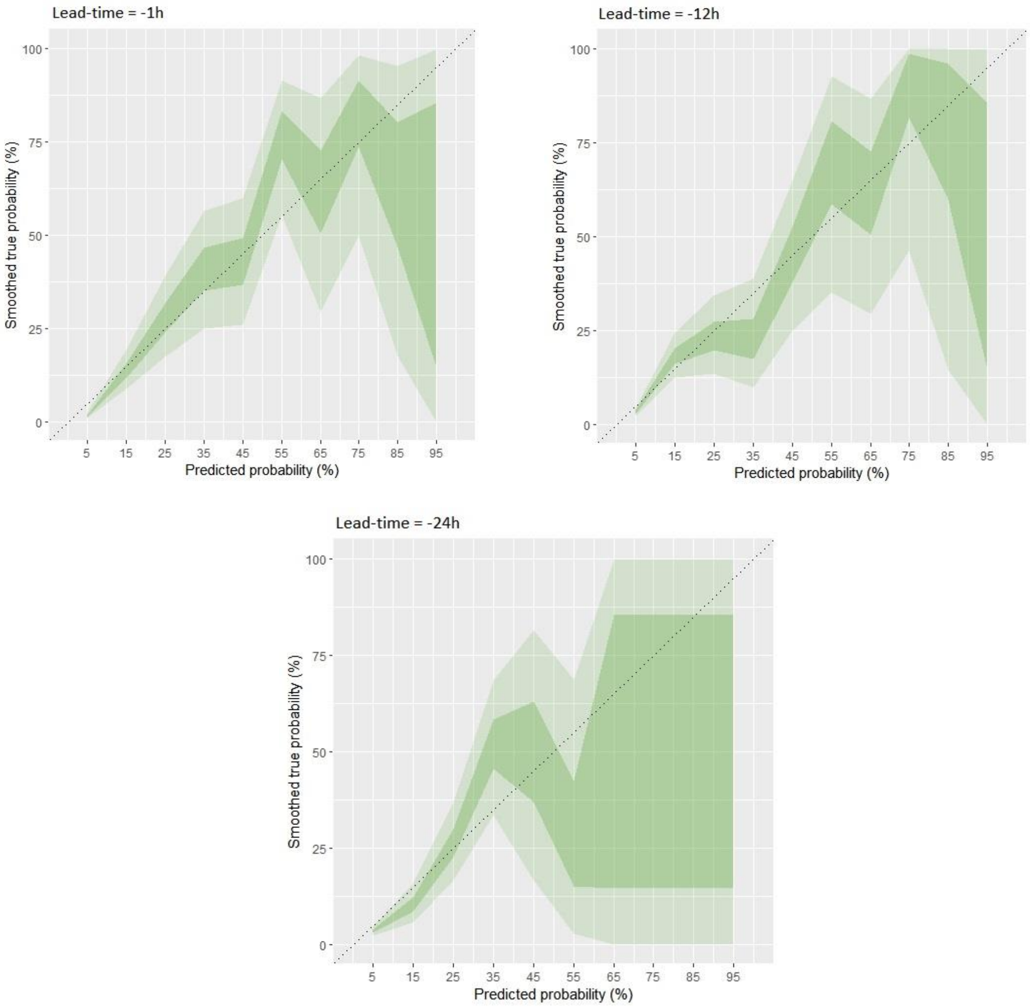
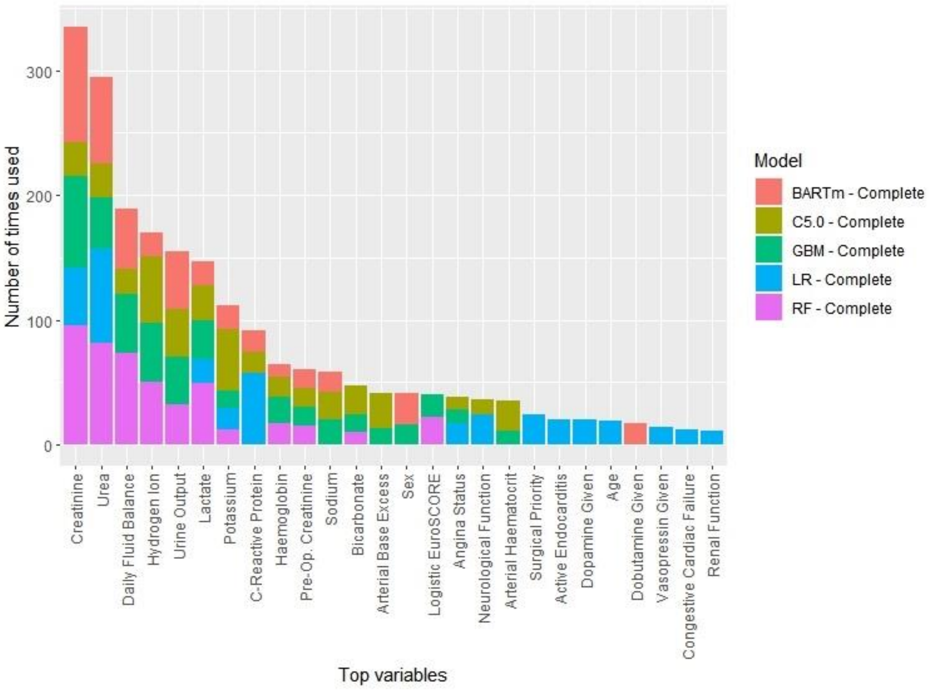


Table 7.3. Number of patients, the proportion of patients with AKI and mean predicted probability with standard deviation (SD) for lead times of -1, -12 and -24.

Lead Time	Number of Patients	AKI (%)	Mean Predicted Probability (%) ± SD (%)
-24	3606	6.49	6.06 ± 8.55
-12	3683	8.2	7.32 ± 11.46
-1	3723	8.97	8.60 ± 13.62

To investigate variable importance, for each model developed for each lead time, the top 20 variables were extracted from models. In Figure 7.12, the most commonly used variables are shown that were used by models for at least 10 times. All five models used creatinine, urea, lactate and potassium in their models. Daily fluid balance, hydrogen ion, urine output and c-reactive protein were also very commonly used. Creatinine and urea, however, were by far the most used variables by the five models.

Figure 7.12. Most important variables used at least 10 times in prediction models.



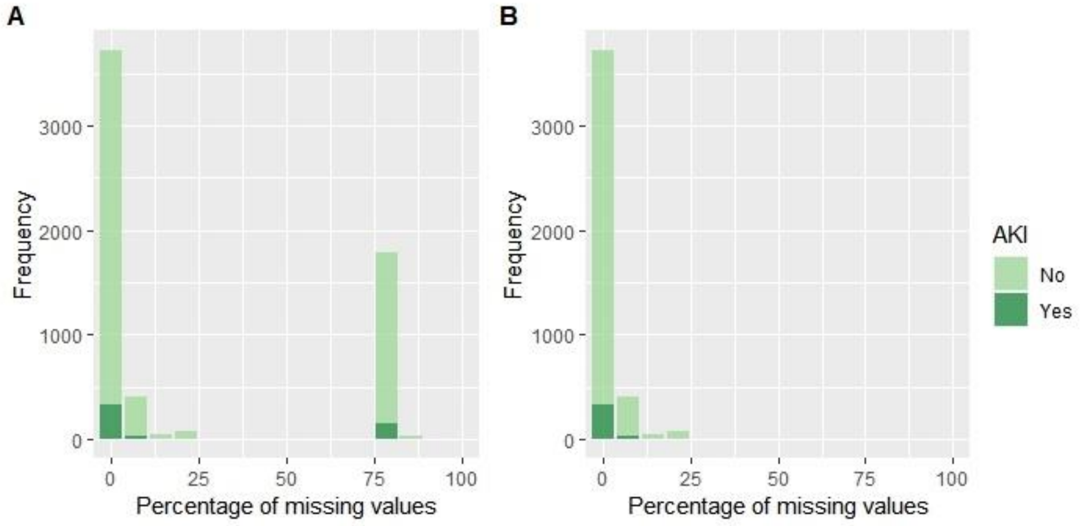
7.5. Experiment 2 and 3 Results: Predicting Acute Kidney Injury in ICU on an Hourly Basis, Using Complete Training Data and Incomplete Testing Data

7.5.1. Data Preparation

7.5.1.1. Missing Data

As found in the literature review in this thesis (Chapter 2), some studies, while dealing with missing values overall, removed patients with more than a certain number of missing features. Following the studies by Hug et al. [300] and Ho et al. [302], further explained in the Chapter 5, patients with more than 40% missing features were removed from this study.

Figure 7.13. Distribution of proportion of missing data in patient records for full data (A) vs when records with >40% missing values have been removed (B), where patients with acute kidney injury (AKI) are marked as dark-green and without AKI are light-green.



When predicting AKI within 25 hours in ICU, including all missing data, for 6056 patients the mean percentage of missing data was 24.40% (SD=35.63%). After removing patients with 40% or more missing data, there were 4244 patient records available for analysis (ca. 30% of patient records removed), where the mean percentage of missing data was 1.21% (SD=3.78%). The histograms in Figure 7.13 show the distribution for the percentage of missing data for the total population (A) and after records with more than 40% of missing variables are removed (B). When removing patients with more than 40% of missing values, 144 patients with AKI were removed from the analysis.

7.5.1.2. Descriptive Statistics

The Table 7.2.A from Appendix 7.2 shows the descriptive statistics for the laboratory values used in the prediction models, and how these statistics changed when using different imputation methods. The comparisons were made with each imputation method and the original data, using Welch Two-Sample t-tests [291].

When comparing the median imputation experiment with original data, there was no statistically significant difference in the distribution of the original dataset and imputed dataset. When applying 0 imputation, there was a significant different between the original dataset and the imputed dataset in terms of minimum, maximum, first and last creatinine, daily fluid balance, bicarbonate, hydrogen ion, and lactate. When applying missForest imputation to the original dataset, there was a significant difference only between maximum daily fluid balance.

7.5.1.3. Training and Testing Data

As seen from Table 7.4, similarly to previous experiments, when predicting AKI within 25 hours of ICU stay, as the lead time got closer to the event of AKI, the percentage of patients with AKI and the number of patients in the total dataset increased. The mean percentage of patients with AKI in total dataset was 7.80% (SD = 0.78). The mean number of patients in the training set was 2802 (SD = 23.43) and testing set was 1393 (SD = 14.78). The completeness in testing datasets varied from 60.66% to 62.54%.

Table 7.4. Number of patients in each training and testing data based on the lead time until AKI within 25 hours in ICU, completeness of data and percentage of patients with AKI.

	Total Data			**Training data (100% complete)**		**Testing data**		
Lead Time	**Number of patients**	**Completeness (%)**	**AKI (%)**	**Number of Patients**	**AKI (%)**	**Number of Patients**	**Completeness (%)**	**AKI (%)**
-24	4139	87.12	6.35	2758	6.74	1355	60.66	5.39
-23	4154	87.20	6.69	2769	6.90	1385	61.60	6.28
-22	4159	87.20	6.80	2772	7.11	1387	61.70	6.20
-21	4162	87.20	6.87	2774	7.14	1388	61.70	6.34
-20	4166	87.28	6.96	2776	7.49	1368	61.26	5.85
-19	4167	87.30	6.98	2777	7.20	1390	61.90	6.55
-18	4171	87.30	7.07	2780	7.55	1391	61.80	6.11
-17	4177	87.30	7.21	2784	7.18	1393	61.80	7.25
-16	4180	87.27	7.27	2786	7.21	1375	61.62	7.27
-15	4188	87.30	7.45	2791	7.70	1397	61.90	6.94
-14	4197	87.30	7.65	2797	7.94	1400	62.00	7.07
-13	4204	87.40	7.80	2802	8.07	1402	62.20	7.28
-12	4211	87.46	7.95	2806	8.27	1380	61.74	7.25
-11	4220	87.50	8.15	2812	8.25	1408	62.60	7.95
-10	4231	87.50	8.39	2820	8.69	1411	62.60	7.80
-9	4234	87.60	8.46	2822	8.50	1412	62.70	8.36
-8	4238	87.59	8.54	2824	8.99	1392	62.21	7.69
-7	4242	87.60	8.63	2827	9.02	1415	63.00	7.84
-6	4242	87.60	8.63	2827	9.02	1415	63.00	7.84
-5	4243	87.70	8.65	2828	8.70	1415	63.00	8.55
-4	4243	87.65	8.65	2828	8.70	1389	62.27	8.42
-3	4243	87.65	8.65	2828	8.70	1389	62.27	8.42
-2	4244	87.65	8.67	2827	8.74	1392	62.36	8.48
-1	4244	87.72	8.67	2829	8.77	1391	62.54	8.34
Mean ± SD	4204 ± 35.22	87.43 ± 0.18	7.80 ± 0.78	2802 ± 23.43	8.02 ± 0.75	1393 ± 14.78	62.10 ± 0.56	7.31 ± 0.92

7.5.2. Experiment 2: Models' Performance

7.5.2.1. Discriminative Performance

Based on AUC, shown in Figure 7.14, BARTm seemed to have a better performance overall than C5.0. The pattern of how the AUC changed with lead time was very similar for both models. According to the Table 7.8, BARTm had a considerably higher mean AUC than C5.0 (0.830 vs 0.794).

Figure 7.14. AUC for both models predicting acute kidney injury for each lead time, using complete training data and missing values in testing data.

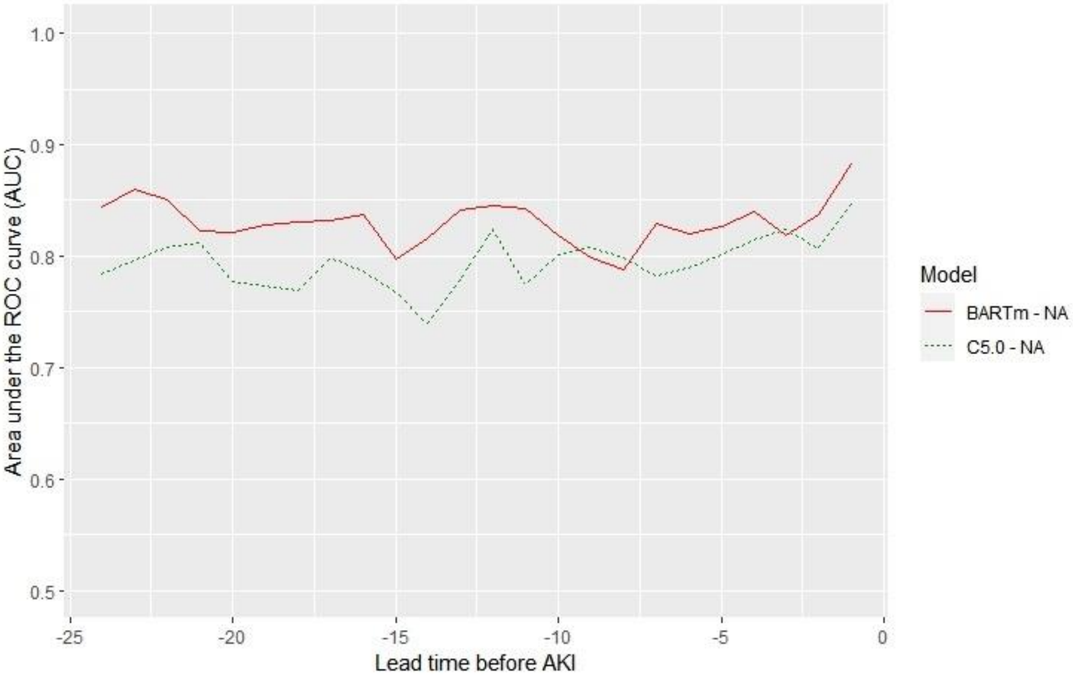
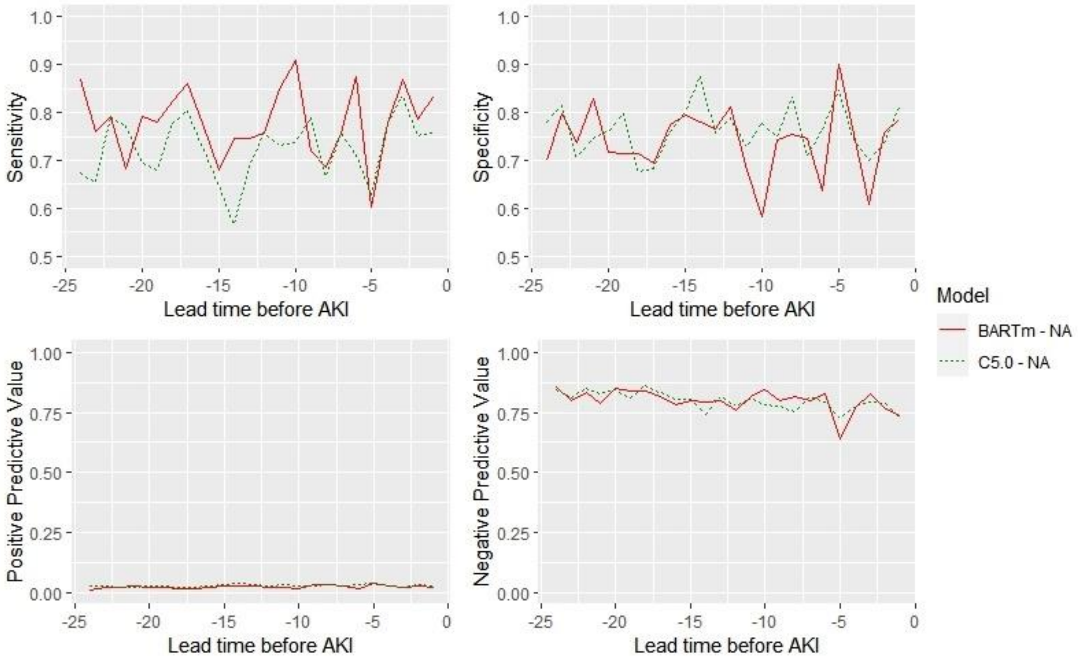


Table 7.5. Mean and standard deviation model performance measures for each model across each lead time before AKI when predicting AKI within 25h since ICU admission, using complete training data and missing values in testing data. The highest result for each performance measure is marked in bold.

	AUC	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Model	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
BARTm - NA	**0.830 ± 0.020**	**0.780 ± 0.074**	0.741 ± 0.070	0.023 ± 0.007	**0.800 ± 0.046**
C5.0 - NA	0.794 ± 0.023	0.724 ± 0.064	**0.764 ± 0.051**	**0.027 ± 0.005**	**0.800 ± 0.037**

According to Table 7.5, BARTm also had a higher mean sensitivity of 0.780 (vs 0.724), however, C5.0 had a higher mean specificity of 0.764 (vs 0.741). Similarly to Experiment 1, the positive predictive values for both models were very low due to low prevalence of AKI in each lead time dataset. The negative predictive values, however, were moderately high, both models having had equal mean NPV of 0.800. C5.0 had slightly less variation in its NPV values than BARTm.

Figure 7.15. Sensitivity and specificity, positive and negative predictive values (PPV and NPV) for both models for each lead time when predicting acute kidney injury on an hourly basis in the ICU, using complete training data and missing values in testing data.



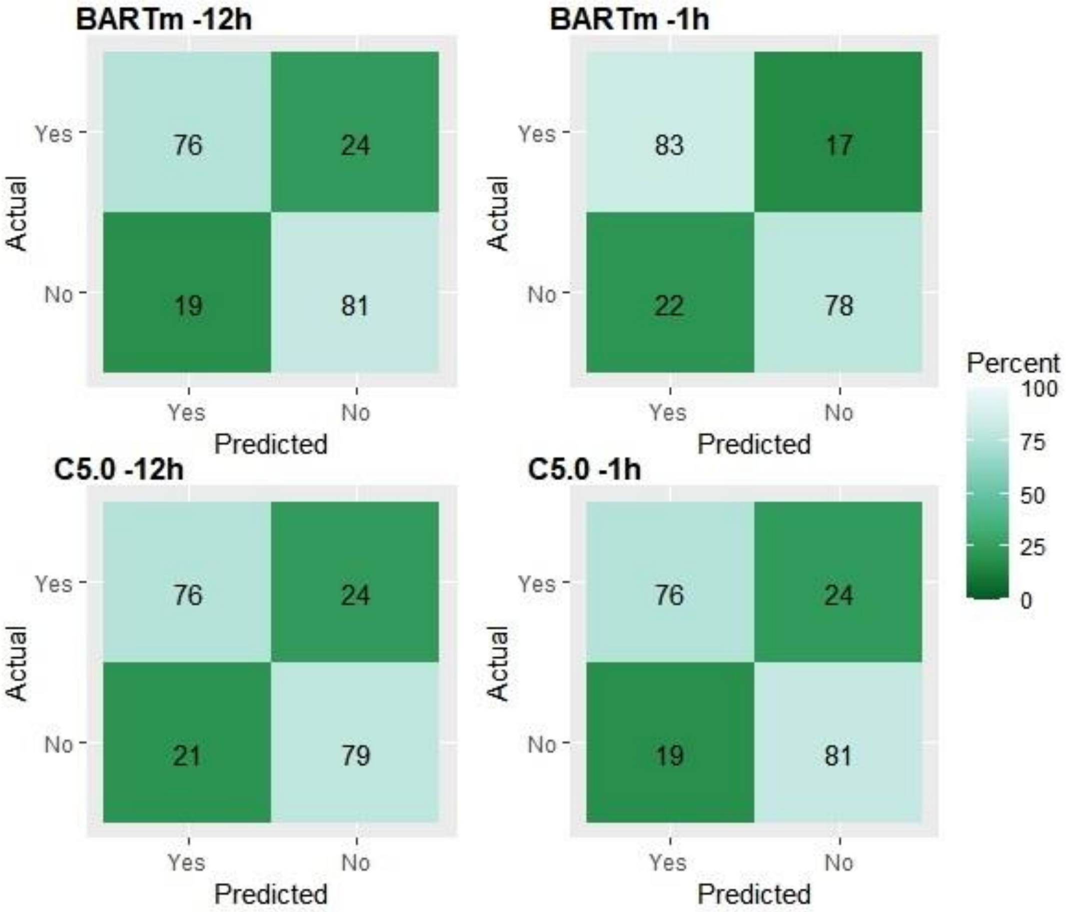
The plots in Figure 7.15 reflect the mean sensitivity, specificity, PPV and NPV in Table 7.8, where BARTm had visibly higher sensitivity at most lead times. There appeared to be more variation in BARTm specificity values than in C5.0 values. The NPV values were moderately high, staying above 0.750 at most lead times, apart from 5 hours before AKI.

The exact performance measures for each lead time for both models can be found from Appendix 7.3, Table 7.3.B. When predicting AKI 24 hours in advance, using missing values in testing set, BARTm had a higher AUC than C5.0 (0.844 vs 0.784) and also higher sensitivity (0.87 vs 0.675). C5.0, however, had considerably higher specificity (0.781 vs 0.699) than BARTm. At 12 hours in advance, the AUC values were closer, BARTm performing slightly better than C5.0 (0.845 vs 0.824). Both models had an equal sensitivity of 0.757 and BARTm had a higher specificity of 0.811 (vs 0.790).

When predicting AKI 4 hours in advance, BARTm again had a higher AUC of 0.840 (vs 0.814). Both models again had the same sensitivity of 0.777, and BARTm had a higher specificity (0.753 vs 0.743).

The confusion matrices (Figure 7.16) show that the models recognised patients with and without AKI correctly similarly when predicting AKI 12 hours in advance. At 1 hour in advance, BARTm classified patients with AKI correctly slightly more than C5.0 (83% vs 76%).

Figure 7.16. Confusion matrices for the two models predicting acute kidney injury, using complete training data and missing values in testing data, 12 hours vs 1 hour in advance.



7.5.2.2. Calibration and Variable Importance

As the BARTm model achieved the highest mean AUC of 0.830, the calibration plots for this model for the lead times of -1, -12 and -24 hours were examined (Figure 7.17).

Similarly to the Experiment 1, here also the models had better calibration if the prediction was made closer to the onset of AKI, i.e., 1 hour before AKI, as opposed to 24 hours before AKI. This was most likely due to the number of patients with AKI being higher when making the -1-hour lead time prediction than when making the prediction 24 hours in advance, as shown in Table 7.6.

The mean predicted probability (Table 7.6) shows that when predicting AKI 24 or 12 hours in advance, the model tended to overestimate risk of AKI in this patient population, however, slightly underestimate risk of AKI when predicting it 1 hour in advance.

Figure 7.17. Calibration plots for BARTm model predicting acute kidney injury 1 hour (top-left), 12 hours (top-right), and 24 hours (bottom-centre) in advance, developed with complete training data and evaluated with incomplete testing data. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.

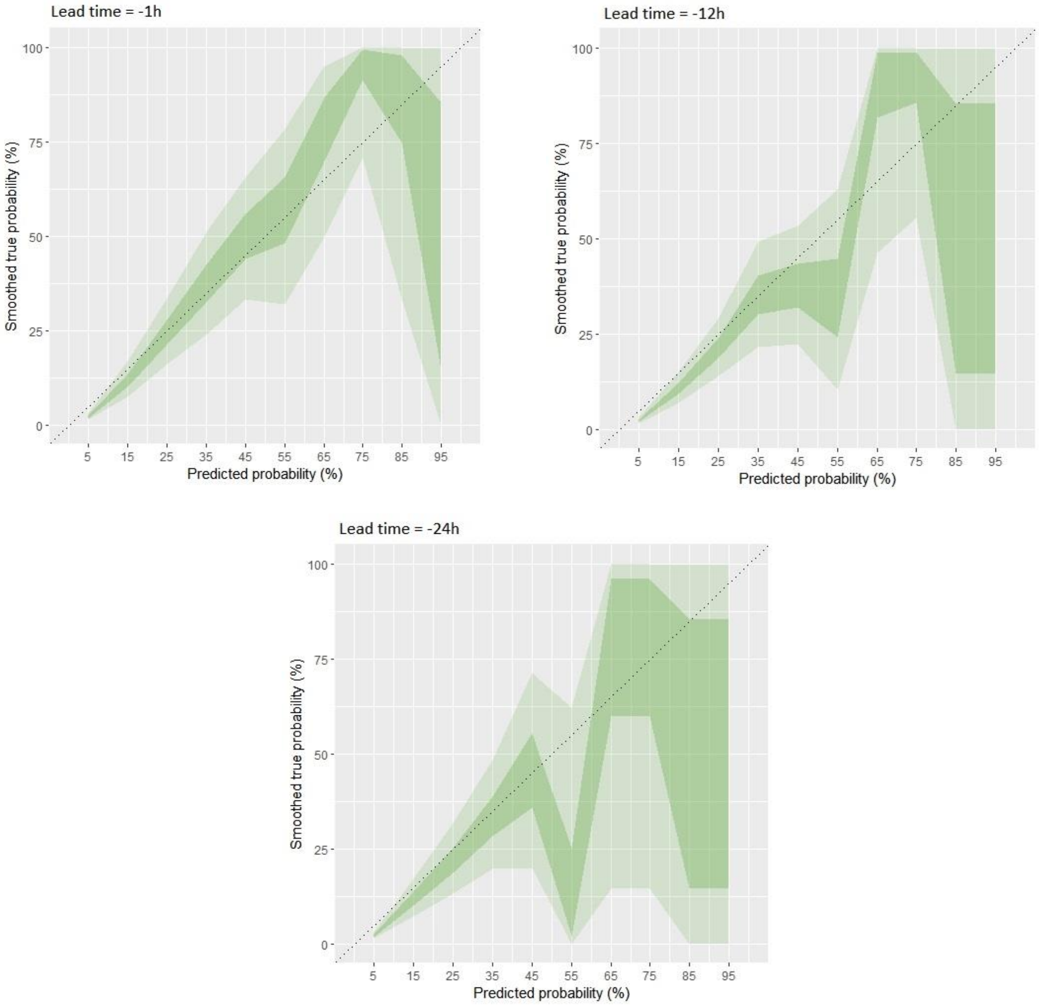
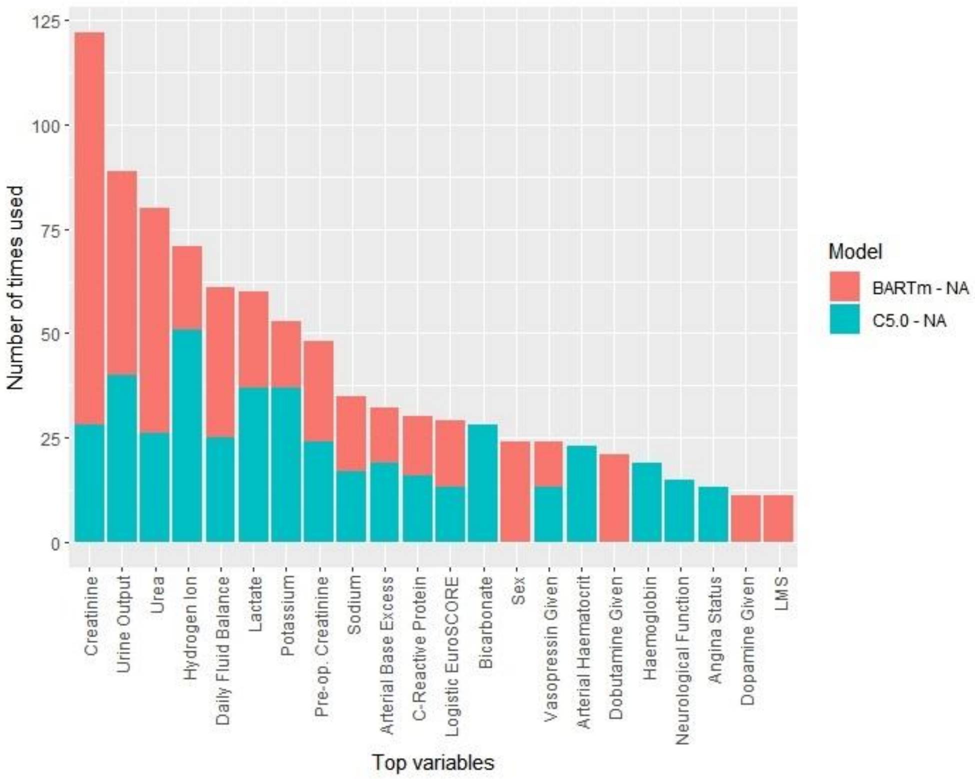


Table 7.6. Number of patients, the proportion of patients with AKI and mean predicted probability with standard deviation (SD) for lead times of -1, -12 and -24.

Lead Time	Number of patients	Completeness (%)	AKI (%)	Mean Predicted Probability (%) ± SD (%)
-24	4139	87.12	6.35	6.38 ± 8.83
-12	4211	87.46	7.95	8.71 ± 11.53
-1	4244	87.72	8.67	8.46 ± 13.37

Figure 7.18. Most important variables used by both models for at least 10 times.



Looking at the variable importance, similarly to the previous experiment, from each model developed for each lead time, the top 20 variables were extracted. The Figure 7.18 shows the variables that were most used by models. Most variables were used by both models, especially the ones that are at the top. As expected, creatinine was by far the most used variable, followed by urine output, urea and hydrogen ion. Daily fluid balance, lactate, potassium, and preoperative creatinine were also deemed important by the models.

7.5.3. Experiment 3: Models' Performance

7.5.3.1. Discriminative Performance

The Figure 7.19 shows how AUC changed for each model as the lead time got close to AKI. In general, the AdaBoost, BARTm, C5.0 and gradient boosting models with all three imputation methods performed better than logistic regression, random forest and support vector machine models. At 15 hours before AKI the AdaBoost, BARTm, C5.0 and gradient boosting with median imputation had a noticeably lower performance than at other lead times. This could be because, as shown in Appendix 7.4, there was a rise in creatinine for patients with AKI from lead time of -15h onwards. In addition, for daily fluid balance, from -15h onwards, there was a higher variability in mean values for both patients with AKI and without AKI. Furthermore, for lactate levels, there was a jump in values and also increased in variation for both maximum and last lactate value for AKI and non-AKI patients. These three variables were amongst top 5 most important variables used by models, as shown later in Figure 7.27.

The Table 7.7 shows the mean and standard deviation of each performance measure across all lead times. BARTm models had the highest mean AUC, the model with 0 imputation in test set having the highest overall mean AUC of 0.849. BARTm with 0 imputation also had the highest mean sensitivity of 0.807. Gradient boosting with median imputation had the highest mean specificity of 0.792. Overall, as seen in previous experiments, all models had very low mean positive predictive values. This was due to low percentage of patients with AKI at each lead time. The negative predictive values, however, were considerably high. Support vector machine with 0 imputation had the highest negative predictive value of 0.838.

In general, BARTm models did very well, regardless of the imputation method. The C5.0, logistic regression and support vector machine models with all imputation methods did the least well in terms of overall performance.

Figure 7.19. AUC for all models predicting acute kidney injury for each lead time, using complete training data and 0, median and missForest imputation methods in testing set.

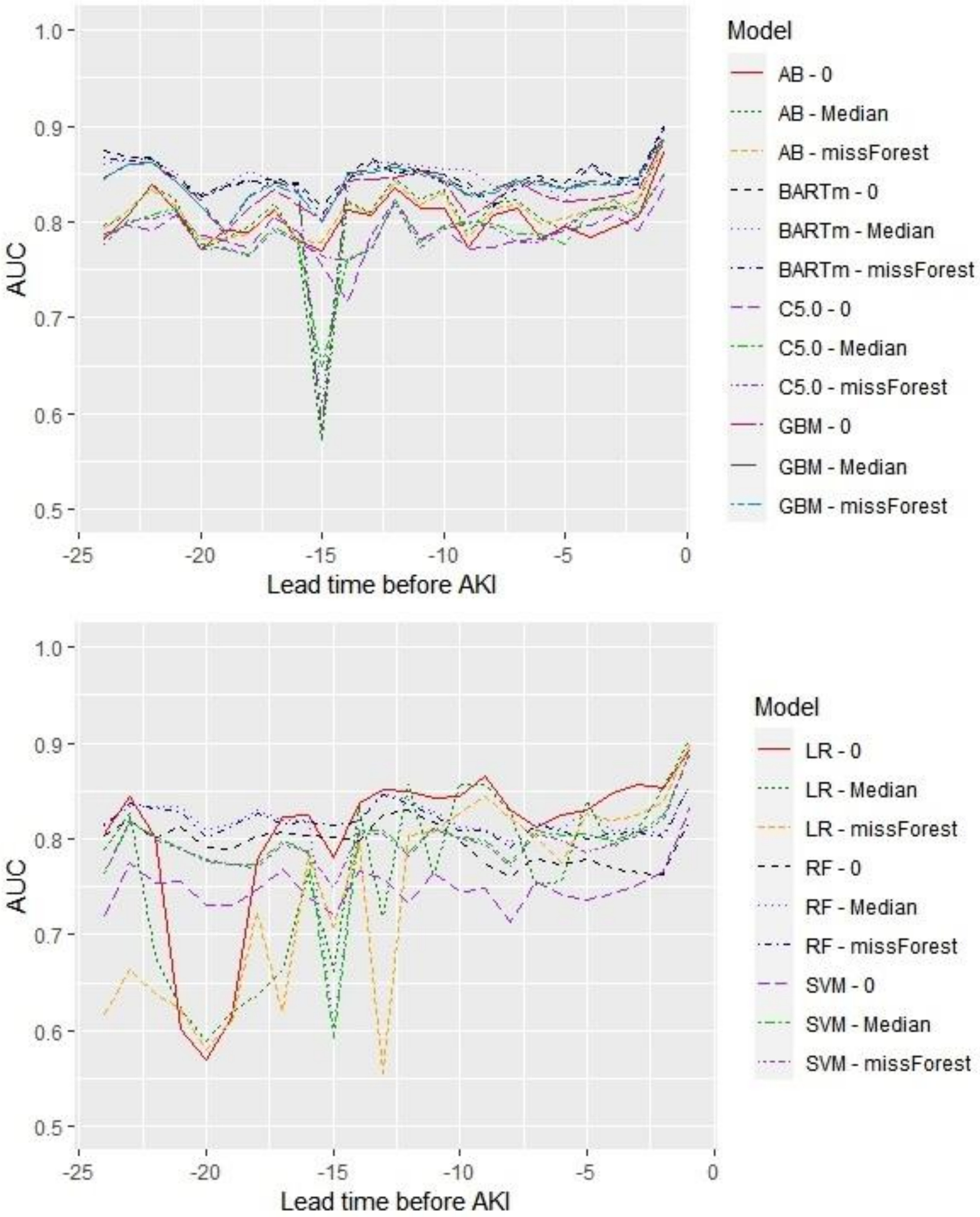


Table 7.7. Mean and standard deviation model performance measures for each model across each lead time before AKI when predicting AKI within 25h since ICU admission, using complete training data and imputation methods to replace missing values in testing data. The highest result for each performance measure is marked in bold.

	AUC	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Model	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
BARTm - 0	**0.849 ± 0.018**	**0.807 ± 0.048**	0.758 ± 0.047	0.020 ± 0.005	0.787 ± 0.037
BARTm - missForest	0.845 ± 0.018	0.785 ± 0.065	0.760 ± 0.057	0.022 ± 0.006	0.789 ± 0.040
BARTm - Median	0.840 ± 0.052	0.760 ± 0.103	0.777 ± 0.076	0.024 ± 0.011	0.777 ± 0.048
GBM - missForest	0.840 ± 0.020	0.766 ± 0.070	0.774 ± 0.069	0.023 ± 0.007	0.779 ± 0.052
GBM - 0	0.833 ± 0.021	0.760 ± 0.074	0.774 ± 0.074	0.024 ± 0.008	0.775 ± 0.055
GBM - Median	0.833 ± 0.056	0.741 ± 0.107	**0.792 ± 0.055**	0.025 ± 0.011	0.773 ± 0.045
RF - missForest	0.818 ± 0.015	0.738 ± 0.059	0.765 ± 0.051	0.026 ± 0.006	0.796 ± 0.035
RF - Median	0.813 ± 0.047	0.744 ± 0.088	0.752 ± 0.053	0.026 ± 0.011	0.804 ± 0.039
AB - missForest	0.811 ± 0.024	0.740 ± 0.081	0.760 ± 0.076	0.026 ± 0.008	0.794 ± 0.052
AB - 0	0.803 ± 0.024	0.722 ± 0.069	0.759 ± 0.053	0.028 ± 0.008	0.803 ± 0.038
AB - Median	0.803 ± 0.055	0.732 ± 0.104	0.761 ± 0.066	0.027 ± 0.011	0.798 ± 0.046
LR - 0	0.803 ± 0.085	0.729 ± 0.199	0.777 ± 0.090	0.025 ± 0.011	0.777 ± 0.062
RF - 0	0.795 ± 0.020	0.747 ± 0.064	0.724 ± 0.063	0.027 ± 0.007	0.819 ± 0.034
SVM - missForest	0.795 ± 0.026	0.709 ± 0.075	0.765 ± 0.078	0.029 ± 0.007	0.797 ± 0.051
SVM - Median	0.792 ± 0.049	0.707 ± 0.078	0.761 ± 0.072	**0.030 ± 0.010**	0.800 ± 0.053
C5.0 - missForest	0.791 ± 0.021	0.717 ± 0.066	0.765 ± 0.052	0.028 ± 0.005	0.801 ± 0.034
C5.0 - Median	0.789 ± 0.036	0.719 ± 0.069	0.762 ± 0.046	0.028 ± 0.008	0.803 ± 0.033
C5.0 - 0	0.788 ± 0.023	0.703 ± 0.079	0.776 ± 0.073	0.029 ± 0.006	0.792 ± 0.039
LR - Median	0.761 ± 0.091	0.692 ± 0.200	0.767 ± 0.122	0.028 ± 0.012	0.782 ± 0.071
SVM - 0	0.750 ± 0.024	0.740 ± 0.070	0.670 ± 0.063	0.029 ± 0.007	**0.838 ± 0.048**
LR - missForest	0.742 ± 0.101	0.683 ± 0.185	0.754 ± 0.100	**0.030 ± 0.010**	0.797 ± 0.072

The Figures 7.20 and 7.21 show how sensitivity and specificity changed as the lead time got closer to AKI. With each lead time the models' performance varied remarkably, making it difficult to point out visually which models had the best performance. C5.0 with 0 and median imputation had a noticeably lower sensitivity at lead times around -15 and -14 than in other lead times. The logistic regression models with all three imputation methods had noticeably higher specificities from lead times of -21 to -19. SVM with 0 imputation and missForest had a substantial drop in specificity at lead times of -6 and -7.

The Figures 7.22 and 7.23 show the changes in positive and negative predictive values for each model as the lead time got closer to AKI. As seen in previous experiments, here also the positive predictive values were very low for all models at all lead times. The negative predictive values tended to stay above 0.700 for all models at most lead times.

The specific performance measures for each lead time for all models are shown in Appendix 7.3 Tables 7.3.C, 7.3.D and 7.3.E. When predicting AKI 24 hours in advance, using imputation methods in testing set, BARTm with 0 imputation had the highest AUC of 0.874. The same model also had the highest sensitivity of 0.870. C5.0 with 0 imputation had the highest specificity of 0.854.

When predicting AKI 12 hours in advance, BARTm with median imputation had the highest AUC of 0.863, followed by gradient boosting model with median imputation (AUC = 0.860). Random forest with 0 imputation had the highest sensitivity of 0.864 and BARTm with median imputation had the highest specificity of 0.869.

When predicting AKI 4 hours in advance, the three BARTm models took the first three places in terms of highest AUC of 0.862 for 0 imputation, and 0.859 equally for median and missForest imputation. The BARTm model with 0 imputation also had the highest sensitivity of 0.843 and the gradient boosting model with 0 imputation had the highest specificity of 0.861.

Figure 7.20. Sensitivity and specificity for AdaBoost, BARTm, C5.0 and gradient boosting models for each lead time when predicting acute kidney injury on an hourly basis in the ICU, using complete training data and imputation methods in testing data.

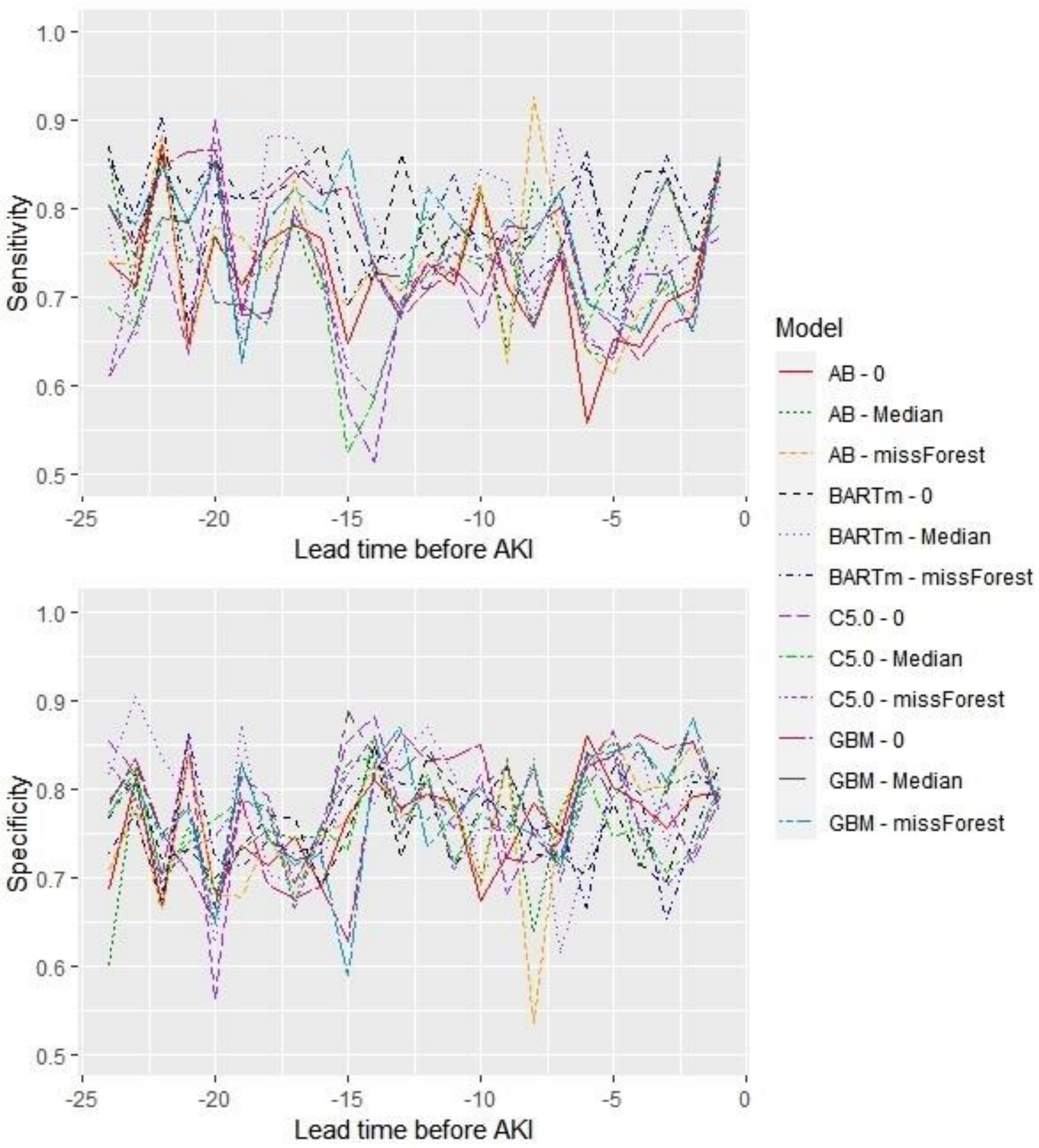


Figure 7.21. Sensitivity and specificity for logistic regression, random forest and support vector machine models for each lead time when predicting acute kidney injury on an hourly basis in the ICU, using complete training data and imputation methods in testing data.

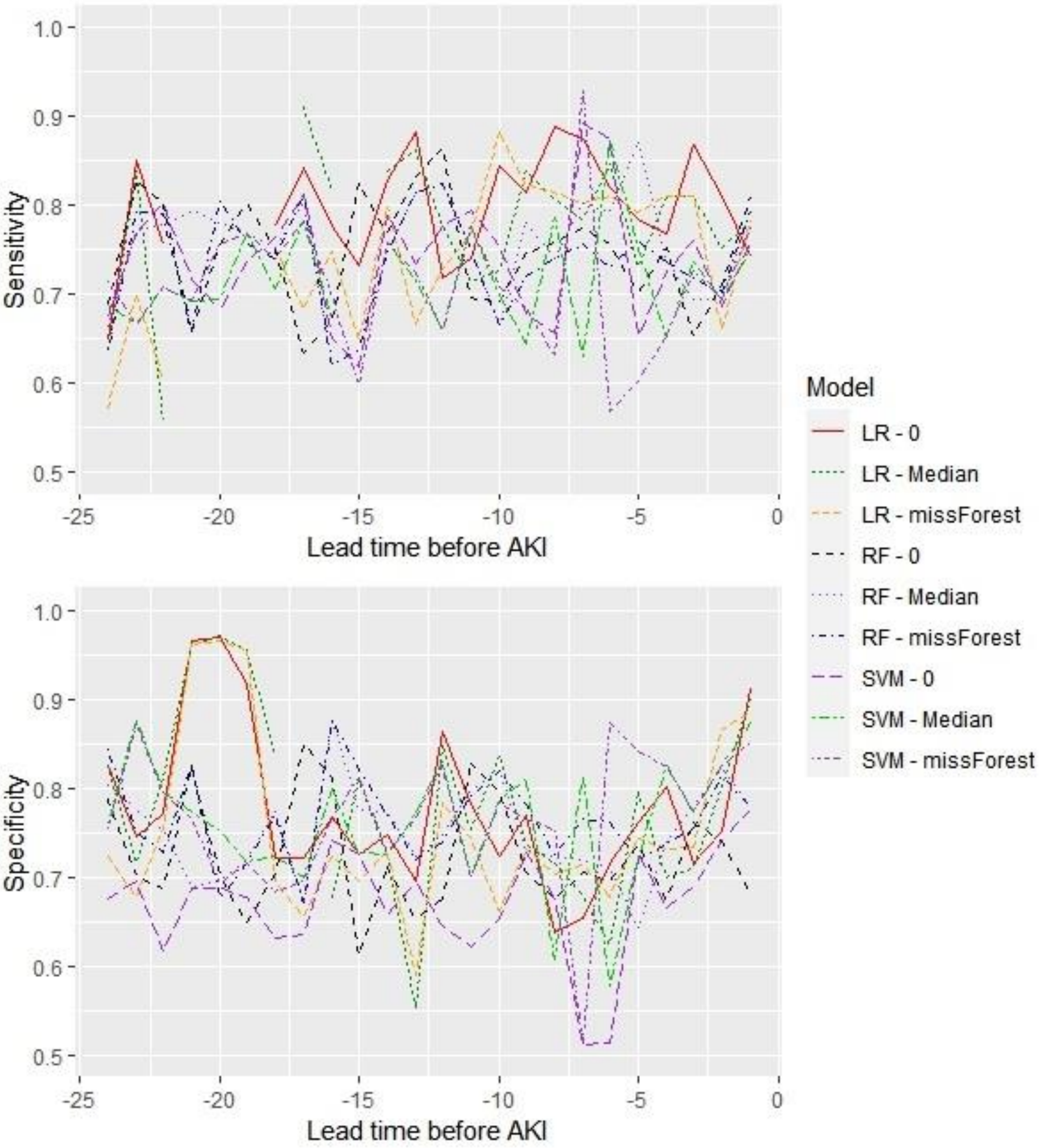


Figure 7.22. Positive predictive values (PPV) for all models for each lead time when predicting acute kidney injury on an hourly basis in the ICU, using complete training data and imputation methods in testing data.

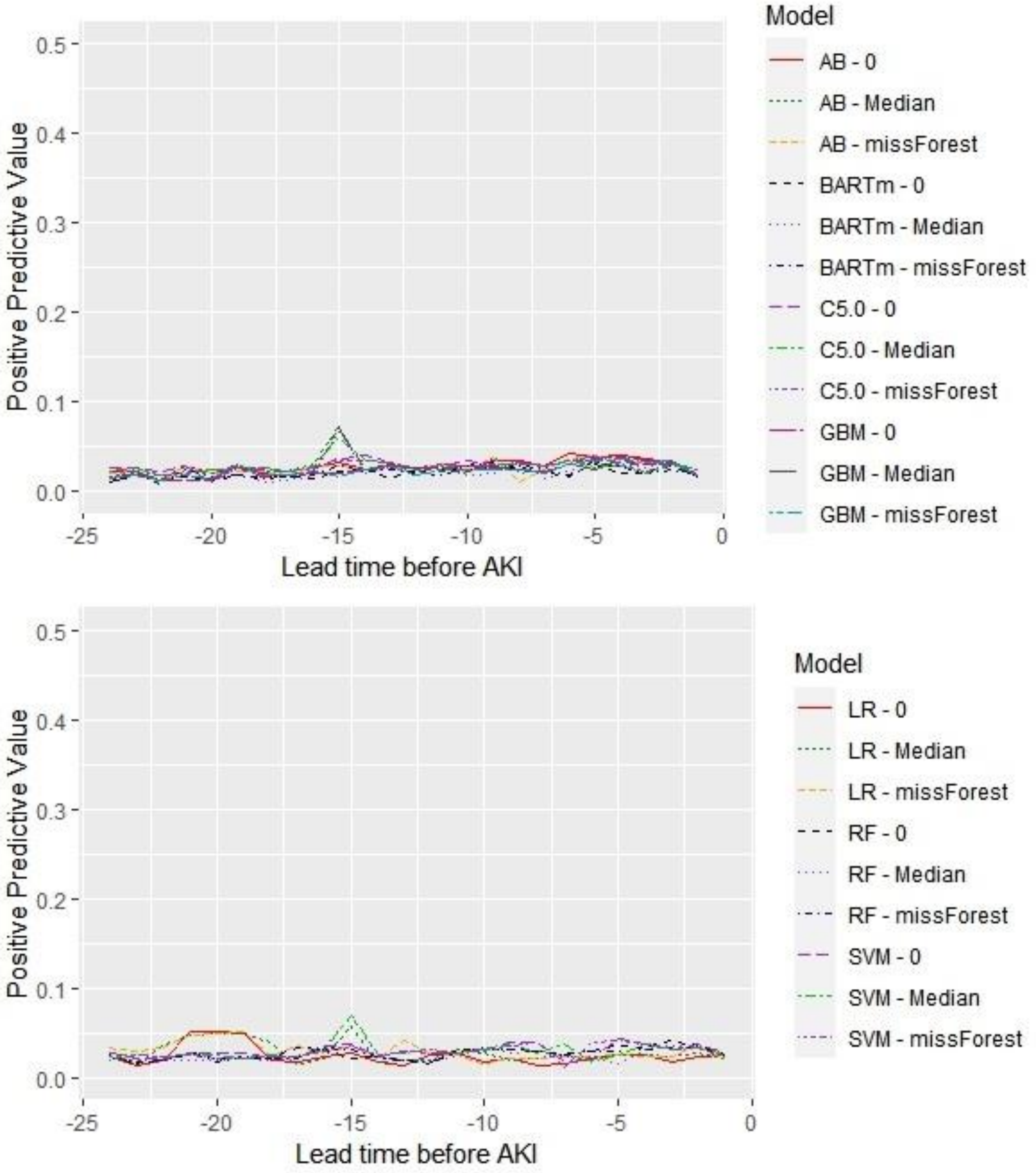


Figure 7.23. Negative predictive values (NPV) for all models for each lead time when predicting acute kidney injury on an hourly basis in the ICU, using complete training data and imputation methods in testing data.

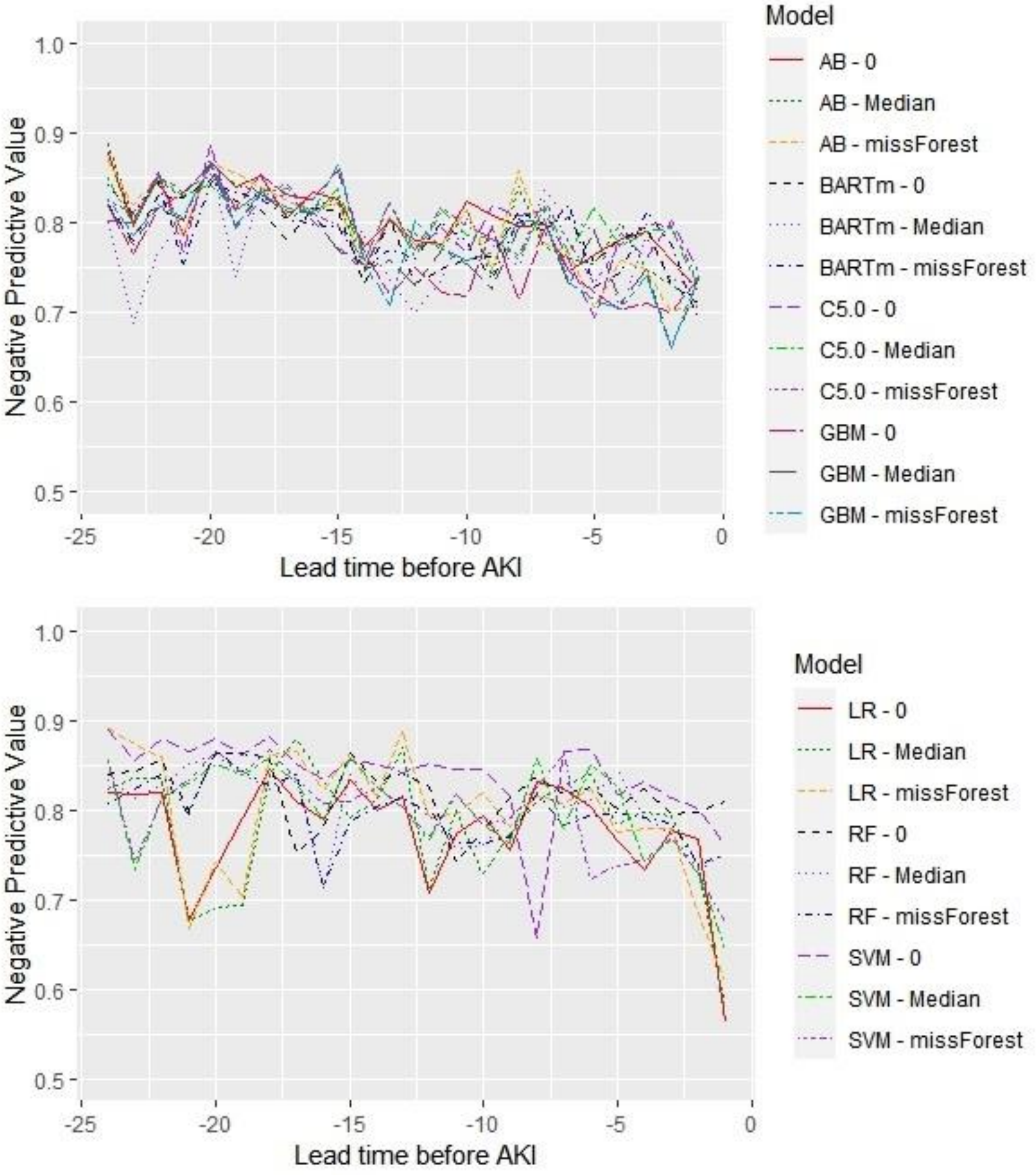


Figure 7.24. Confusion matrices for all models predicting acute kidney injury 12 hours vs 1 hour in advance, using median imputation (A) and 0 imputation (B) in testing data.

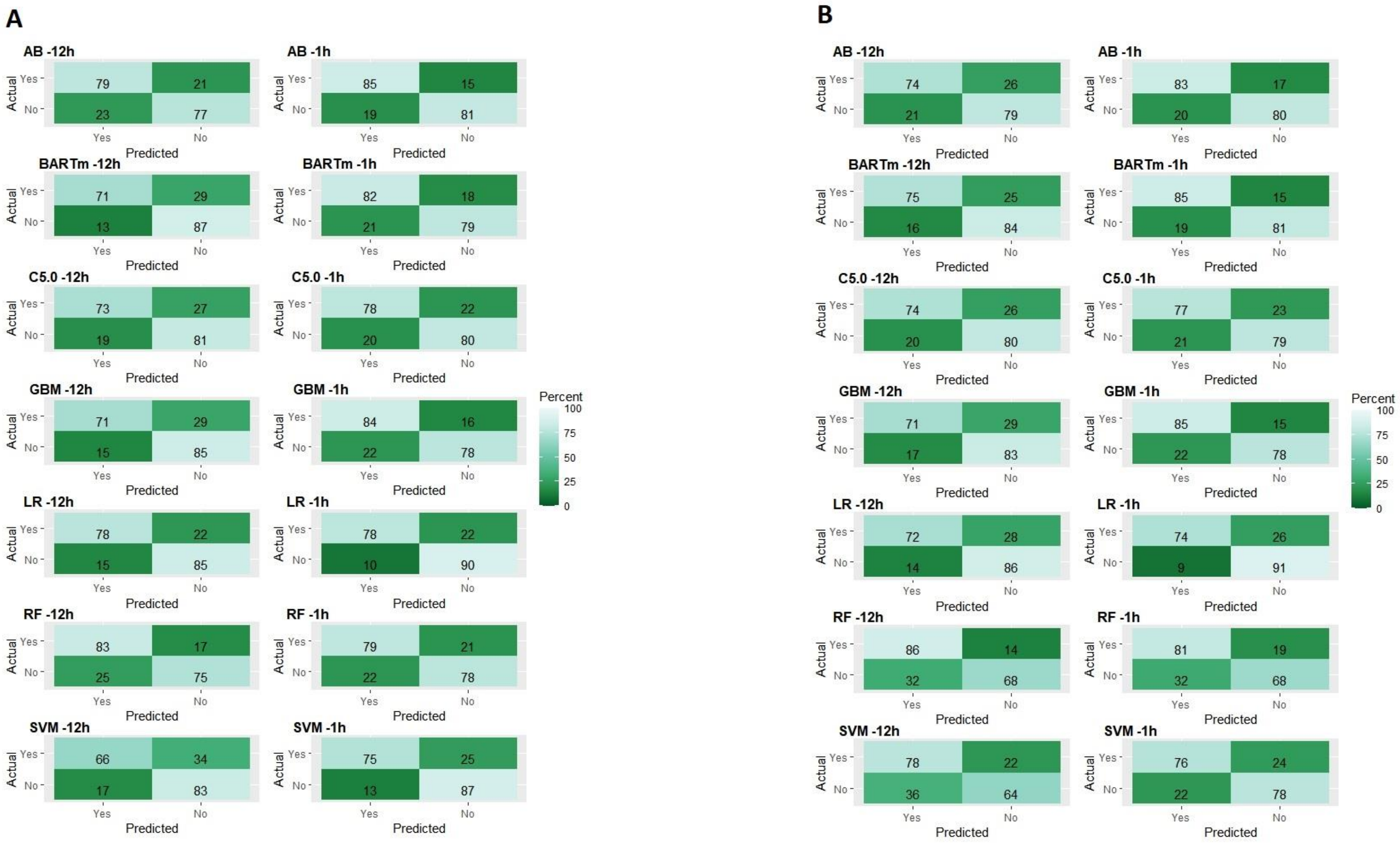
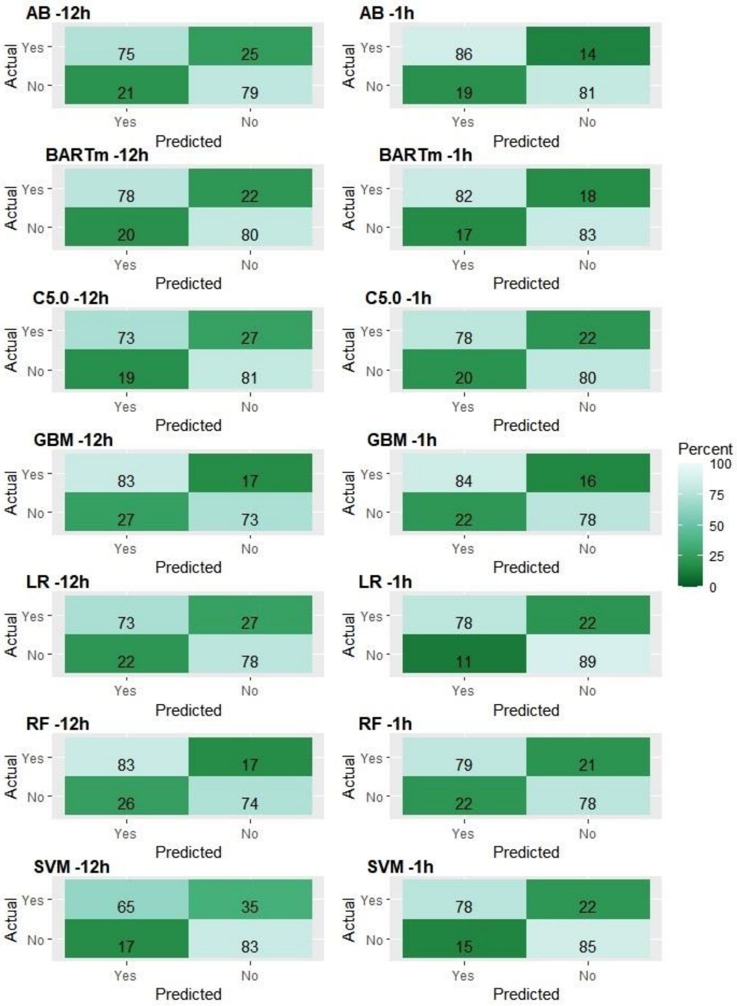


Figure 7.25. Confusion matrices for all models predicting acute kidney injury 12 hours vs 1 hour in advance, using missForest imputation in testing data.



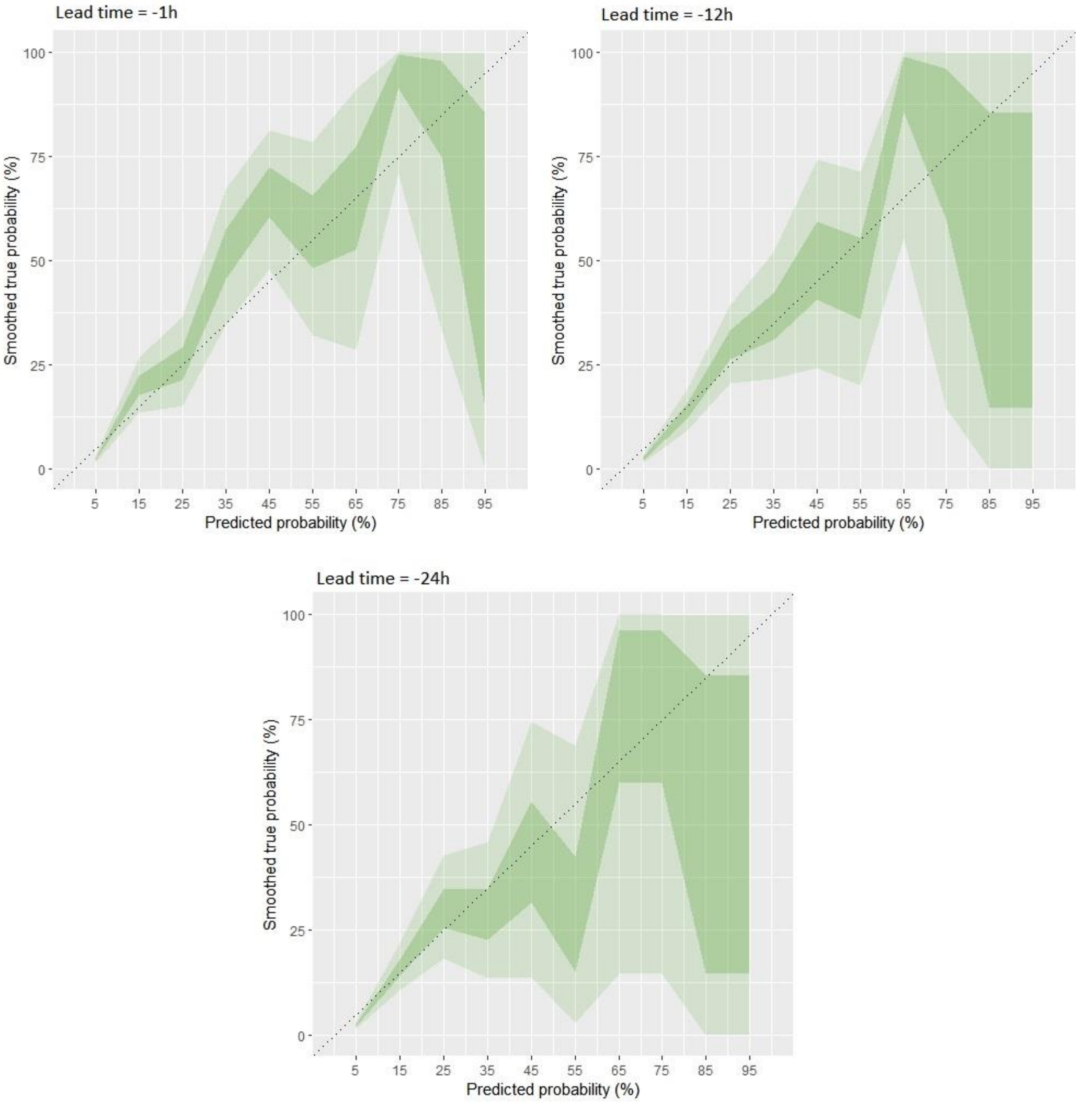
The Figures 7.24 and 7.25. show the confusion matrices for each model developed to predict AKI 12 hours and 1 hour in advance. In general, the models predicted a similar number of patients with AKI correctly for both lead times. Support vector machine, however, tended to be better at predicting patients without AKI with the three imputation methods than patients with AKI.

7.5.3.2. Calibration and Variable Importance

Since the BARTm model evaluated with 0 imputation in test set had the best mean performance (AUC = 0.849), the calibration for this model was assessed. As seen from Figure 7.26, the model calibration when predicting AKI 1 hour in advance largely overestimated risk of AKI for patients who had true probability for AKI lower than ca.

75%. Similarly to previous experiments, due to considerably small number of patients with AKI in the patient population, the confidence intervals for the estimated predicted probabilities were very wide for patients who actually had AKI. This indicates the model's uncertainty about recognising patients with AKI.

Figure 7.26. Calibration plots for BARTm model predicting acute kidney injury 1 hour (top-left), 12 hours (top-right), and 24 hours (bottom-centre) in advance, developed with complete training data and evaluated with testing data, where missing values were replaced with 0. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.



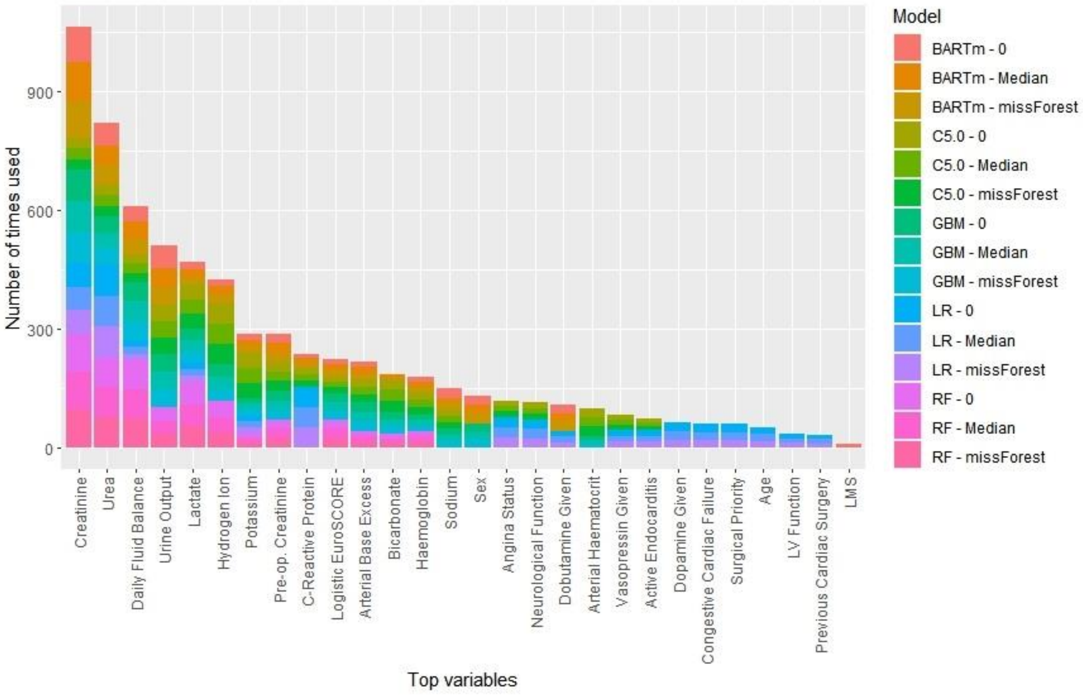
However, according to the mean predicted probability (Table 7.8), in all cases, the model underestimated risk of AKI based on the comparison between the mean probability and the proportion of AKI patients in each respective dataset for each lead time. The underestimation of risk for AKI was especially clear for the lead times of -1 and -24 hours.

Table 7.8. Number of patients, the proportion of patients with AKI and mean predicted probability with standard deviation (SD) for lead times of -1, -12 and -24.

Lead Time	**Number of patients**	**Completeness (%)**	**AKI (%)**	**Mean Predicted Probability (%) +- SD (%)**
-24	4139	87.12	6.35	5.53 +- 8.35
-12	4211	87.46	7.95	7.28 +- 10.28
-1	4244	87.72	8.67	6.89 +- 12.35

When looking at variable importance (Figure 7.27), the results were very similar to the previous experiments, where creatinine, urea, daily fluid balance, urine output, lactate and hydrogen ion were the most commonly used variables by models.

Figure 7.27. Top variables used by models for at least 10 times.



7.6. Discussion

7.6.1. Summary of Results

As shown in Table 7.9, based on AUC, BARTm had the highest performance at most lead times before AKI. There was less clarity in terms of sensitivity and negative predictive values to say which model was the best at predicting AKI within 25 hours of ICU stay. However, all models seemed to have similarly moderately high performance.

Table 7.9. Highest performance measures for lead times in every 4 hours when predicting AKI within 25 hours of ICU stay.

Lead Time	AUC (Model)	Sensitivity (Model)	Specificity (Model)
-24	0.874 (BARTm – 0)	0.914 (BARTm)	0.854 (C5.0 – 0)
-20	0.831 (BARTm)	0.902 (C5.0 – 0)	0.975 (LR)
-16	0.841 (BARTm – 0)	0.874 (BARTm – 0)	0.878 (RF – Median)
-12	0.863 (BARTm – Median)	0.864 (RF – 0)	0.869 (BARTm – Median)
-8	0.849 (LR)	0.926 (AB – missForest)	0.863 (LR)
-4	0.887 (BARTm)	0.868 (BARTm)	0.861 (GBM – 0)
-1	0.918 (BARTm)	0.932 (BARTm)	0.911 (LR – 0)

When looking at the overall mean performance measures for each model across all lead times (Table 7.10), BARTm with complete data had the highest overall mean AUC (0.850) and sensitivity (0.821). This was lower than Meyer et al.'s AUC of 0.91, however higher than their sensitivity of 0.79 [74]. The highest specificity belonged to the logistic regression model developed with complete data (Spec = 0.824), which is slightly lower than the specificity achieved by Meyer et al. (Spec = 0.86) [74].

Although Meyer et al.'s study was the only one found in the literature review to develop dynamic prediction models for AKI, the results by Meyer et al., however, should be taken with a pinch of salt as their model was developed and evaluated on a balanced test set. In case of an imbalanced classification problem, if any balancing methods are used, these methods should be applied to training data only, and never to testing data [134]. This is, because, when evaluating models' performance, the testing data should be as close to the real-world data as possible. When using a balanced testing data, the model results are not applicable to the real-world cardiac patients.

In general, all models in this chapter showed a moderately high performance based on AUC, sensitivity and specificity, many staying above 0.700 in terms of mean AUC.

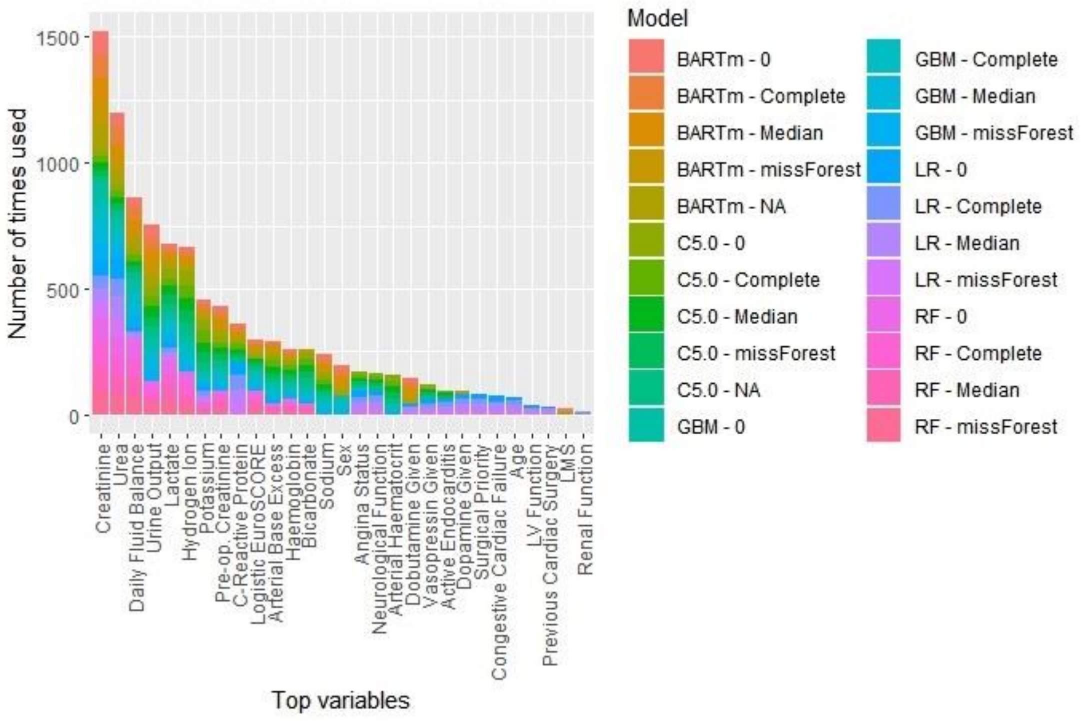
Table 7.10. Overall mean and standard deviation performance measures across all lead times for all models and experiments when predicting AKI within 25 hours in the ICU.

Model	**AUC** **Mean ± SD**	**Sensitivity** **Mean ± SD**	**Specificity** **Mean ± SD**
BARTm	**0.850 ± 0.026**	**0.821 ± 0.053**	0.741 ± 0.057
BARTm - 0	0.849 ± 0.018	0.807 ± 0.048	0.758 ± 0.047
BARTm - missForest	0.845 ± 0.018	0.785 ± 0.065	0.760 ± 0.057
GBM - missForest	0.840 ± 0.020	0.766 ± 0.070	0.774 ± 0.069
BARTm - Median	0.840 ± 0.052	0.760 ± 0.103	0.777 ± 0.076
GBM	0.838 ± 0.031	0.786 ± 0.067	0.755 ± 0.072
GBM - 0	0.833 ± 0.021	0.760 ± 0.074	0.774 ± 0.074
GBM - Median	0.833 ± 0.056	0.741 ± 0.107	0.792 ± 0.055
BARTm - NA	0.830 ± 0.020	0.780 ± 0.074	0.741 ± 0.070
RF - missForest	0.818 ± 0.015	0.738 ± 0.059	0.765 ± 0.051
RF - Median	0.813 ± 0.047	0.744 ± 0.088	0.752 ± 0.053
AB - missForest	0.811 ± 0.024	0.740 ± 0.081	0.760 ± 0.076
AB	0.810 ± 0.037	0.752 ± 0.072	0.745 ± 0.069
RF	0.810 ± 0.034	0.754 ± 0.079	0.739 ± 0.057
AB - 0	0.803 ± 0.024	0.722 ± 0.069	0.759 ± 0.053
AB - Median	0.803 ± 0.055	0.732 ± 0.104	0.761 ± 0.066
LR - 0	0.803 ± 0.085	0.729 ± 0.199	0.777 ± 0.090
LR	0.802 ± 0.100	0.668 ± 0.216	**0.824 ± 0.080**
RF - 0	0.795 ± 0.020	0.747 ± 0.064	0.724 ± 0.063
SVM - missForest	0.795 ± 0.026	0.709 ± 0.075	0.765 ± 0.078
C5.0 - NA	0.794 ± 0.023	0.724 ± 0.064	0.764 ± 0.051
SVM - Median	0.792 ± 0.049	0.707 ± 0.078	0.761 ± 0.072
C5.0 - missForest	0.791 ± 0.021	0.717 ± 0.066	0.765 ± 0.052
SVM	0.790 ± 0.035	0.719 ± 0.071	0.761 ± 0.071
C5.0 - Median	0.789 ± 0.036	0.719 ± 0.069	0.762 ± 0.046
C5.0 - 0	0.788 ± 0.023	0.703 ± 0.079	0.776 ± 0.073
C5.0	0.787 ± 0.034	0.742 ± 0.052	0.738 ± 0.058
LR - Median	0.761 ± 0.091	0.692 ± 0.200	0.767 ± 0.122
SVM - 0	0.750 ± 0.024	0.740 ± 0.070	0.670 ± 0.063
LR - missForest	0.742 ± 0.101	0.683 ± 0.185	0.754 ± 0.100

Interestingly, all BARTm and gradient boosting model experiments took the top places based on mean AUC. In general, the models with imputation methods had higher performance overall. However, BARTm with missing values in test set was at the 9th place in terms of mean AUC, and 5th place in terms of sensitivity, and the AUC values compared to the top models were not noticeably different (AUC of 0.830 for BARTm with NA in test vs 0.850 for the top model). The BARTm model with missing values also had one of the highest NPV of 0.800 (Appendix 7.3 Table 7.3.B) This performance is promising as missing data in electronic health records is a major obstacle predictive modelling in healthcare [110]. Being able to make a prediction for a patient who does not have all the necessary data available allows for clinicians to make a decision with the aid of prediction model for all patients.

The top-performing models for each experiment had similar calibration, where the certainty about predicted probabilities was high when true probabilities for AKI were low, and the certainty about predicted probabilities was low when true probabilities for AKI were high. This is due to the number of patients with AKI being very low compared to the number of patients without AKI, also reflected in very low PPV for all models. Based on mean predicted probabilities for each top-performing model in the three experiments, the models tended to slightly underestimate the risk of AKI.

Figure 7.28. Most important variables that are used by models for at least 10 times overall.



The Figure 7.28 shows the most commonly used variables across all lead times for all experiments in this chapter. As seen from the individual experiments, the models tended to use the same variables as top variables. These variables are creatinine, urea, daily fluid balance, urine output, lactate and hydrogen ion, all of which have been shown to be associated with AKI in other studies [270], [271], [383], [384].

The variable importance is not often reported in the other paper predicting AKI in a dynamic manner. Meyer et al. do not report variable importance [74] presumably due to using recurrent deep neural network, which is a complex method that does not allow for measuring variable importance. Knowing variable importance, however, can be a key aspect to adopting the model in clinical practice. If the clinician knows which factors are associated to the probability of AKI to be high, the clinicians can pay closer attention to these factors and apply relevant clinical interventions.

To sum up, the models predicting AKI, developed in this chapter, had in general moderately high performance in terms of AUC, sensitivity, and specificity. It is also promising that the models developed in this chapter achieved all moderately high negative predictive values, which offers certainty that if a patient is predicted to not have AKI, then in reality the patient actually will not have AKI (shown by NPV).

7.7. Conclusion

In this chapter the onset of AKI was predicted on an hourly basis, using preoperatively variables and laboratory variables available in the ICU. Overall, most models had considerably good performance based on AUC. The overall best performance was achieved by BARTm model that used complete training and complete testing data. However, the models that were evaluated on testing sets with missing values or where missing values were replaced, also achieved comparable results.

Chapter 8. Study 2.2: Predicting the Onset of Delirium on an Hourly Basis in Intensive Care

8.1. Introduction

This chapter aims to experiment with different predictive modelling methods to predict the onset of delirium following cardiac surgery within 21 hours since ICU admission on an hourly basis, using static preoperatively recorded and dynamic ICU data.

In this chapter a binary classification problem was investigated: “Does a patient have delirium or not?”

As done in Chapter 7, three main experiments were undertaken to find the most optimal model for predicting delirium:

1. Predicting delirium using complete data only.
2. Predicting delirium, using complete training data and missing values in testing data. Patient records missing more than 40% of the variables were excluded from the analysis, as done elsewhere [300], [302].
3. Predicting delirium, using complete training data and imputation methods to replace missing values in testing data. Again, patient records with more than 40% of the variables missing were excluded from the analysis.

Detailed information about the methods and experiments in this chapter can be found from Chapter 5.

8.2. Related Work

There are various models developed to predict delirium following cardiac surgery. Three recent systematic reviews found 26 unique prediction models for predicting delirium in ICU [362]–[364]. Of these models, 4 were identified by the review papers to be “dynamic” models: DYNAMIC-ICU [385], Auto-DelRAS [386], ABD-pm [387], and a model developed by Oh et al. [388]. As explained in Chapter 2, a model is considered dynamic if its prediction is updated as the time changes, based on the renewed input information [389]. Predicting clinical outcomes in a dynamic manner helps clinicians to be informed of patient’s risk for delirium on near real-time basis.

It is worth noting that none of these studies were included in the literature review (Chapter 2) due to not meeting the eligibility criteria. Namely, while Oh et al. used patients’ heart rate for their analysis, they did not include any laboratory data in their model [388]. The ABD-pm [387] and the DYNAMIC-ICU did not include any variables that are repeatedly measured in the ICU, like laboratory variables or vital signs [385]. Auto-DelRAS included blood urea nitrogen as a dynamic variable [386], however, it was not specified how exactly this variable was treated to make a prediction as laboratory variables are recorded in the ICU several times for each patient. Hence, it is also unclear how exactly this model is deemed to be a “dynamic” prediction model.

For DYNAMIC-ICU, it is unknown how often the model is calculated and how much time in advance delirium is predicted [385]. For Auto-DelRAS, it is known that the model is calculated once a day, however, again it is not reported at what time point before delirium the prediction is made [386]. ABD-pm is calculated once a day to predict the next day probability for delirium (and other outcomes), however, it is unknown at what time in the next day delirium could happen [387]. As delirium can manifest itself in a matter of hours [390], a daily prediction is too infrequent.

Out of the four models, based on the information published by papers, the model developed by Oh et al, is the only truly dynamic prediction model for delirium in the ICU. It reports CAM-ICU assessment in every 8 hours, and it uses heart rate variability to predict delirium every 3 hours. By predicting delirium on a real-time basis using heart-rate variability only, Oh et al. achieve the highest performance when predicting delirium using linear extreme learning machine (Accuracy = 0.6389, Sens = 0.8797, Spec = 0.2776, PPV = 0.6485 and NPV = 0.5286). AUC is not reported. Even though this model shows some promising results, it was developed using 94 patient records only. [388]

Overall, none of the 26 studies report at what time specifically delirium occurs in their patient population. Pisani et al., however, state that 70.4% of their patient population had delirium within 48 hours of ICU admission [391]. This means that these models could attempt to predict delirium when it has already happened, which is a serious limitation to these models.

To improve upon delirium prediction in ICU, in this chapter various steps were taken:

- Experiments were undertaken to predict delirium, using preoperative data. This was done to understand which preoperative variables were associated with postoperative delirium.
- Experiments were undertaken to predict delirium on an hourly basis, using CAM-ICU status with their timestamps.
- As opposed to static variables used in models found in the review papers [362]–[364], dynamic laboratory variables were used to use more up-to-date data for prediction.
- The analysis was undertaken using data recorded within the first 21 hours since ICU admission. This was due to delirium happening within the first 21 hours of ICU admission in this dataset, as shown in Section 8.3.1.
- Experiments were undertaken using complete data, missing data and imputation methods in order to understand whether the models using only complete data could be improved upon.

As said by the developers of the CAM-ICU [285],

> *"Predictive models for delirium are useful to identify high risk patients for proactive implementation of preventive strategies, for identifying patients who need closer monitoring, for identifying vulnerability factors for intervention, for prognostic decision-making, and for determining clinical trial eligibility. The ability to stratify risk can assist physicians in explaining risks to patients and families and can help families to better understand the recovery process and potential outcomes."*

8.3. Patient Population and Delirium

Delirium was recorded at GJNH for patients from 2016 onwards. Hence, patients' data who had a procedure from 2016 to 2018 were included in the analysis. The patients who had delirium within the first hour since being admitted to the ICU were excluded from the analysis. This is because the event at time = 0 would be impossible to predict in the ICU, and more information would be needed about the surgery itself. Hence, the total number of patients included in the analysis was 3322.

8.3.1. Patient Demographics

The prevalence of delirium was found to be 12.47% (95% CI 11.39% – 13.63%) out of 3322 patients. As seen from Appendix 8.1. Table 8.1.A, the overall mean age for patients was 65.82 years. The majority of patients were male (71.44%). Most patients did not have type II diabetes (74.85%) and about a third of the patients had never smoked (34.00%) or were ex-smokers (34.27%).

The most common surgery was CABG (51.97%), followed by valve surgery (33.13%). The mean logistic EuroSCORE calculated for patients was 5.58. Patients usually stayed in the ICU for slightly more than two days (mean hours = 51.77) and in the hospital for just under 12 days. Following surgery, 1.20% of the patients died.

When comparing these demographics between patients with delirium and without delirium, there were statistically significant differences between these two populations based on age, sex, type of procedure, logistic EuroSCORE, ICU hours, and total days in hospital. Patients with delirium were significantly older than patients without delirium. More patients were also females and more patients had higher risk surgery, such as valve, or combined CABG and valve surgery. Patients with delirium had higher preoperatively calculated logistic EuroSCORE, which shows there is a potential that logistic EuroSCORE could also indicate risk of postoperative delirium. On average, patients without delirium stayed in the ICU for slightly less than 2 days (mean = 38.01 hours), whereas patients with delirium stayed in the ICU for almost four days (mean = 148.36 hours). Patients with delirium stayed in the hospital, on average, 7 days longer than patients without delirium.

The Figures 8.1 to 8.4 show the distribution of age, logistic EuroSCORE, ICU hours and total days in hospital for patients without delirium and with delirium.

Figure 8.1. Histogram of age for patients without delirium (light-green) vs with delirium (dark-green).

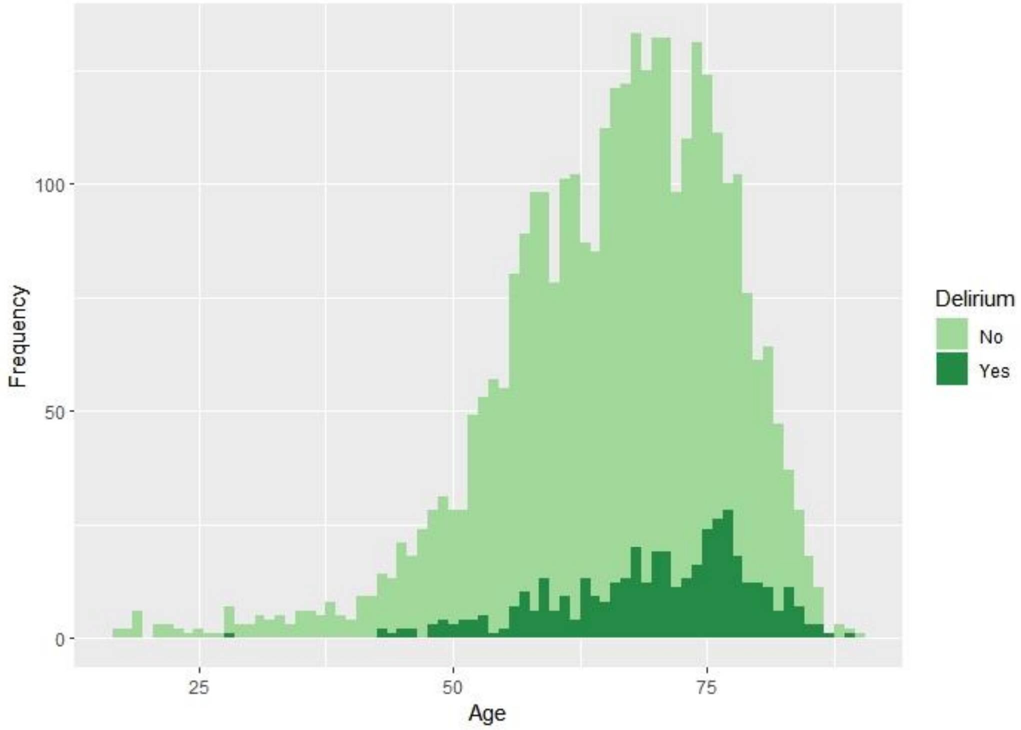


Figure 8.2. Histogram of logistic EuroSCORE for patients without delirium (light-green) vs with delirium (dark-green).

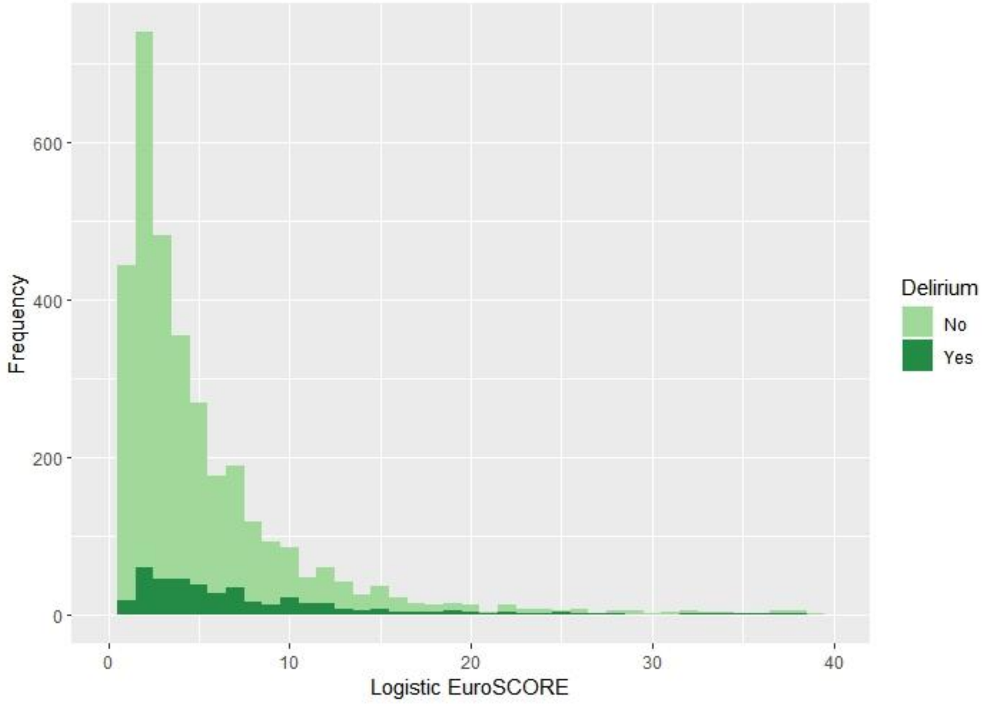


Figure 8.3. Histogram of intensive care unit hours for patients without delirium (light-green) vs with delirium (dark-green).

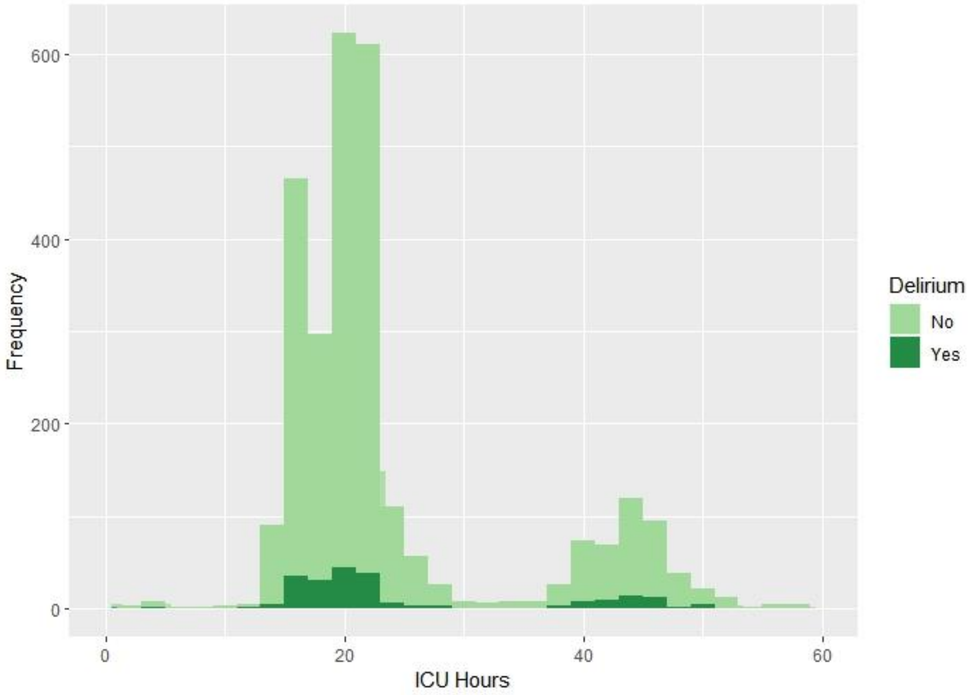
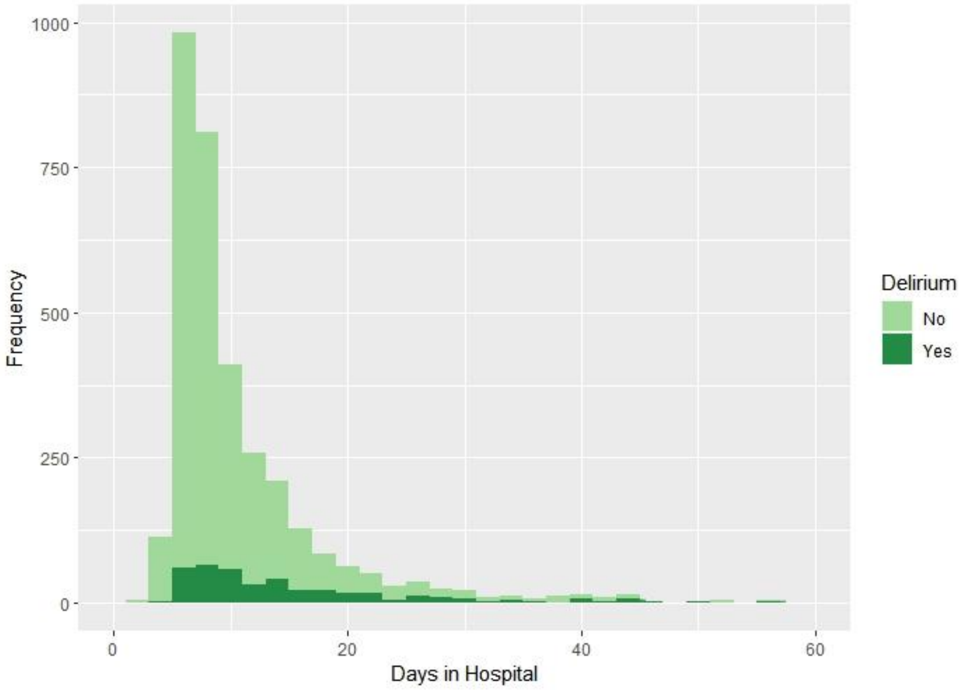
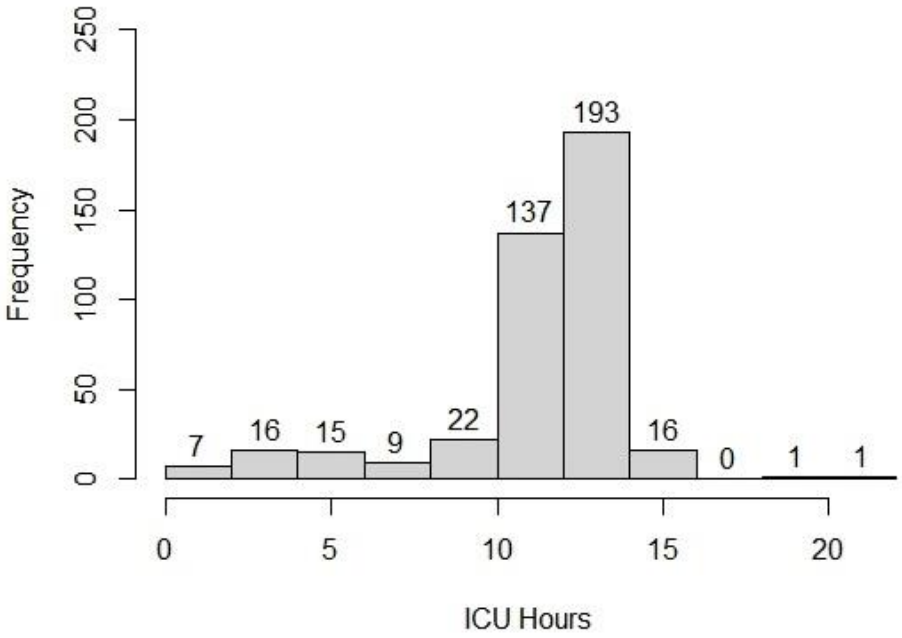


Figure 8.4. Histogram of total days in hospital for patients without delirium (light-green) vs with delirium (dark-green).



As seen from the Figure 8.5, the majority of patients had delirium between 10 – 13 hours since the admission to ICU. More specifically, the mean time of delirium occurrence was 11.21 hours (SD = 2.84), and median time was 12.02 (IQR = 1.1). The maximum time of the first delirium occurrence was at 20.6 hours since ICU admission. The distribution of the time of delirium onset determines how often and how many hours in advance delirium can be predicted.

Figure 8.5. Time of delirium onset in ICU, based on CAM-ICU assessments.



It is also important to note that CAM-ICU score was measured in mean time of every 10.52 hours (SD = 4.02), with median time of every 11.90 hours (IQR = 0.95). This also explains why suddenly such a large number of patients had delirium diagnosis at these hours. This will be further discussed as a limitation in Section 9.2 in Chapter 9.

8.3.2. Descriptive Statistics

Along with the demographic variables presented in Section 8.3.1, other preoperative variables and laboratory variables (collected in ICU) were used in the analysis. These comprise of six demographic variables (age, sex, BMI, type II diabetes, smoking status and procedure), 19 preoperative variables (priority, critical preoperative state, previous cardiac surgery, previous percutaneous coronary intervention, extracardiac arteriopathy, left ventricular function, New York Heart Association (NYHA) grade, angina status, rhythm of the heart, preoperative renal function, preoperative creatinine, neurological dysfunction, previous myocardial infarction, left main stem disease, pulmonary disease, hypertension history, congestive cardiac failure and active endocarditis), and 17 laboratory variables (arterial base excess, arterial haematocrit, bicarbonate, c-reactive protein, creatinine, daily fluid balance, haemoglobin, hydrogen ion, lactate, potassium, sodium, urea, urine output, dobutamine, dopamine, noradrenaline and vasopressin).

8.3.2.1. Preoperative Data

The Appendix 8.1, Table 8.1.B shows patient characteristics for the preoperatively recorded data. Overall, the most common surgery was elective surgery (48.53%). The majority of patients were not at critical preoperative state (97.34%) and had not had a previous cardiac surgery (96.56%) or percutaneous coronary intervention (86.21%). Most patients did not have extracardiac arteriopathy (89.77%), and the majority of patients had also good left ventricular function (76.61%). The most common NYHA grade level was II (47.82%). Most patients had normal heart rhythm (80.95%).

Just over a half of the patients had normal renal function before surgery (53.11%) and normal creatinine levels (mean = 90.79). The majority of patients had no neurological problems (99.28%), previous myocardial infarction (64.21%) and left main stem disease (53.74%). It was also common not to have pulmonary disease (85.02%), congestive cardiac failure (86.93%) or active endocarditis (98.32%). The majority of patients, however, had hypertension history (72.40%). This was expected, as these patients were going to have an open-heart surgery.

When comparing patients with delirium versus without delirium, there were statistically significant differences in these two patient populations in terms of surgical priority, critical preoperative state, NYHA grade, heart rhythm, renal function and preoperative creatinine, left main stem disease and congestive cardiac failure.

Patients with delirium had a higher proportion of emergency surgery (4.08% vs 1.20%) than patients without delirium. More patients with delirium were in a critical preoperative state (5.99% vs 2.19%). With delirium, more patients had higher levels of NYHA grades, especially level III (33.57% vs 24.94%) and level IV (9.35% vs 4.07%). Preoperative abnormal rhythm was more common for patients who had postoperative delirium (20.14% vs 13.87%). In addition, the renal function was significantly worse for patients with delirium (20.14% vs 7.93% for severely impaired function) and preoperative creatinine levels were above normal (mean = 101.17). Fewer patients, however, had left main stem disease if they belonged to the population with delirium (65.47% vs 52.92%). More patients, however, had congestive cardiac failure (20.14% vs 12.06%) if they had delirium.

8.3.2.2. Laboratory Data

As seen from Appendix 8.1, Table 8.1.C, the patient population with delirium was significantly different from the patient population without delirium based on all laboratory variables, apart from the vasopressin dose. This indicates that using these laboratory variables can help with predicting delirium in the ICU.

8.4. Experiment 1 Results: Models Predicting Delirium in ICU on a Real Time Basis, Using Complete Data

8.4.1. Data Preparation

As seen from Table 8.1, the number of patients in each lead time dataset ranged from 2149 to 2307 in total. The percentage of patients with delirium ranged from 6.51% to 12.60%. There were remarkably fewer patients with delirium in the -13-hour lead time dataset than there were in the other lead time datasets. This was, because most patients had delirium for the first time between 10 and 13 hours (mean = 11.21 ± 2.84). The mean number of patients across all lead times was 2266 (SD = 40.74) and the mean proportion of patients with delirium was 11.08% (SD = 1.59). This was slightly lower than the overall prevalence of delirium for patients who experienced it within the first 21 hours since ICU admission, without accounting for patients who had the onset of delirium within the first hour. Even though there was quite a large difference in the proportion of patients with delirium in testing and training dataset at the lead time of -

10, the mean proportion of patients with delirium across all lead times for these datasets were quite similar (10.87% ± 1.82 for training and 11.73% ± 1.89 for testing).

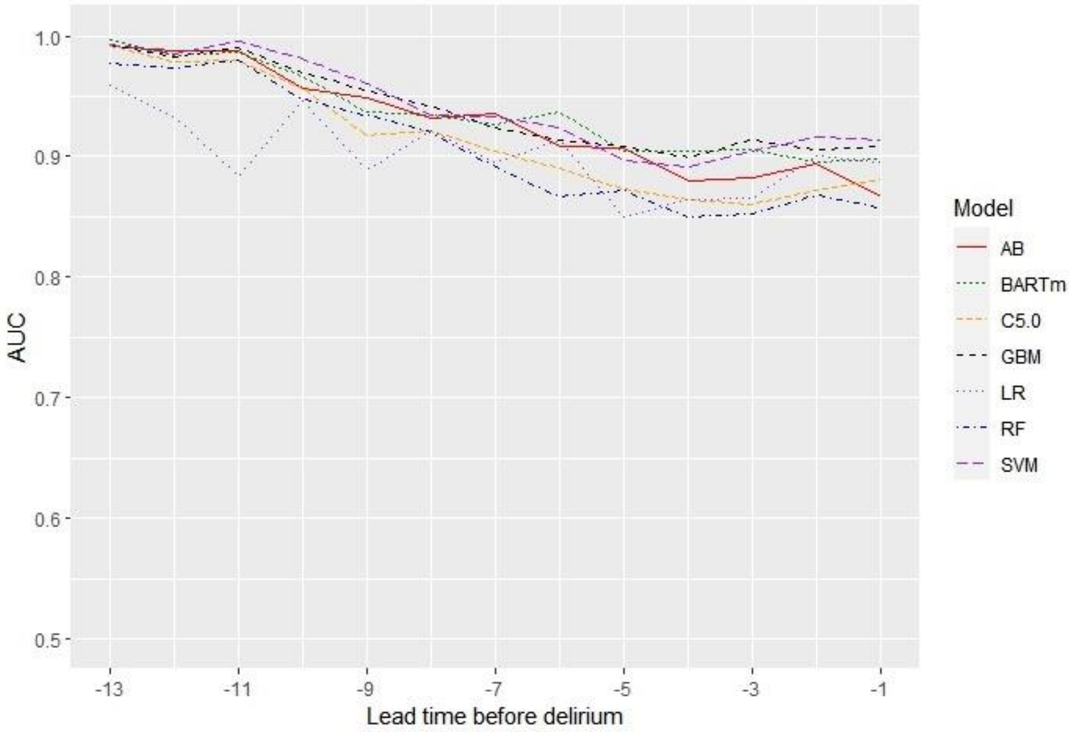
Table 8.1. Number of patients and proportion of patients with delirium in each training and testing data, depending on lead time, if predicting delirium within 21 hours of ICU stay.

	Total Data		**Training Data**		**Testing Data**	
Lead Time	**Number of Patients**	**Delirium (%)**	**Number of Patients**	**Delirium (%)**	**Number of Patients**	**Delirium (%)**
-13	2149	6.51	1440	6.39	709	6.77
-12	2233	10.00	1497	9.22	736	11.70
-11	2245	10.05	1505	10.00	740	11.50
-10	2250	10.07	1508	8.75	742	14.70
-9	2263	11.20	1517	10.50	746	12.60
-8	2269	11.40	1521	11.80	748	10.40
-7	2273	11.50	1523	11.40	750	11.60
-6	2281	11.70	1529	10.90	752	13.40
-5	2287	11.90	1533	12.10	754	11.70
-4	2294	12.20	1537	13.10	757	10.20
-3	2301	12.30	1542	12.80	759	11.30
-2	2307	12.60	1546	12.20	761	13.30
-1	2307	12.60	1546	12.20	761	13.30
Mean ± SD	2266 ± 40.74	11.08 ± 1.59	1519 ± 27.28	10.87 ± 1.82	747 ± 13.47	11.73 ± 1.89

8.4.2. Models' Discriminative Performance

Figure 8.6 shows that all models had moderately good to very good performance at all lead times. Most models stayed above AUC of 0.850 throughout. Visually, logistic regression, random forest and C5.0 had the lowest overall performance. The other models had visually similar performance. Interestingly, all models' performance decreased slightly as the lead time got closer to the predicted event. This is further discussed in Section 8.6.

Figure 8.6. AUC for all models for each lead time when predicting delirium on an hourly basis in the ICU, using complete data.



The Table 8.2 shows that overall, all models had mean AUC above 0.900. Support vector machine, however, had the highest mean AUC across all lead times (mean AUC = 0.941, SD = 0.038) and also the highest mean sensitivity (mean Sens = 0.907, SD = 0.048). C5.0 had the highest mean specificity of 0.885 (SD = 0.067). Overall, all models had moderately high to very high mean sensitivity and specificity, all staying above 0.800. All models had very low positive predictive values and also quite low negative predictive values, random forest having had the highest mean NPV of 0.563 (SD = 0.133).

The Figures 8.7 and 8.8 show how sensitivity, specificity, positive and negative predictive values changed for each model as the lead time changed. When looking at sensitivity, all models decreased in performance as the lead time got closer to delirium. However, sensitivity appeared to be moderately high at most lead times, staying above 0.700 for most models. Specificity appeared to be in general slightly higher than sensitivity at all lead times. All models stay above 0.700 at all lead times.

Table 8.2. Mean and standard deviation model performance measures for each model across each lead time before delirium when predicting delirium within 21h since ICU admission, using complete data. The highest result for each performance measure is marked in bold.

	AUC	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Model	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
AB	0.929 ± 0.043	0.863 ± 0.060	0.877 ± 0.078	**0.027 ± 0.015**	0.453 ± 0.186
BARTm	0.937 ± 0.036	0.875 ± 0.062	0.884 ± 0.069	0.019 ± 0.010	0.449 ± 0.188
C5.0	0.915 ± 0.047	0.822 ± 0.097	**0.885 ± 0.067**	**0.027 ± 0.015**	0.473 ± 0.162
GBM	0.939 ± 0.035	0.875 ± 0.070	0.875 ± 0.064	0.019 ± 0.010	0.489 ± 0.125
LR	0.901 ± 0.033	0.848 ± 0.054	0.884 ± 0.050	0.023 ± 0.008	0.485 ± 0.119
RF	0.907 ± 0.050	0.850 ± 0.084	0.832 ± 0.083	0.024 ± 0.013	**0.563 ± 0.133**
SVM	**0.941 ± 0.038**	**0.907 ± 0.048**	0.870 ± 0.066	0.015 ± 0.007	0.488 ± 0.150

All models had very low positive predictive values at all lead times. This is because the prevalence of delirium is relatively low. This means that if a model predicts that the patient will have delirium, the probability for the patient actually to have a delirium is very low. Negative predictive values increase as the prediction is done nearer the onset of delirium. However, the values are not high enough to offer certainty that if a model predicts that the patient will not have delirium, then the patient actually will not have delirium. This is interesting as the prevalence of delirium was not very low (ca. 11%), meaning a slightly higher positive predictive value was expected. The PPV and NPV here show that the models are not particularly confident in their estimated predicted probabilities in relation to whether a patient will or will not have delirium, as will be further shown with models' calibration in Section 8.4.3.

Figure 8.7. Sensitivity and specificity for all models for each lead time when predicting delirium on an hourly basis in the ICU, using complete data.

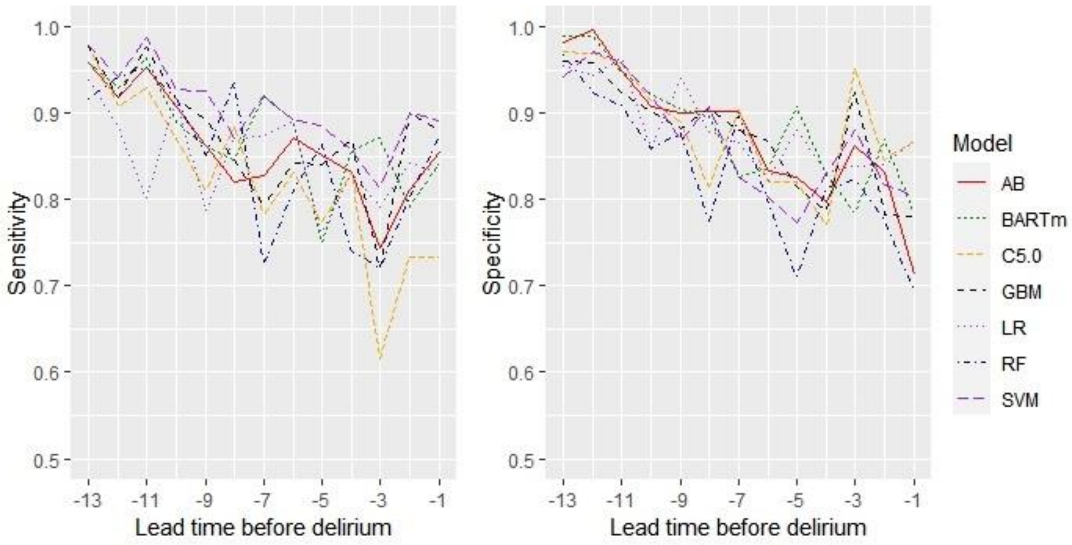
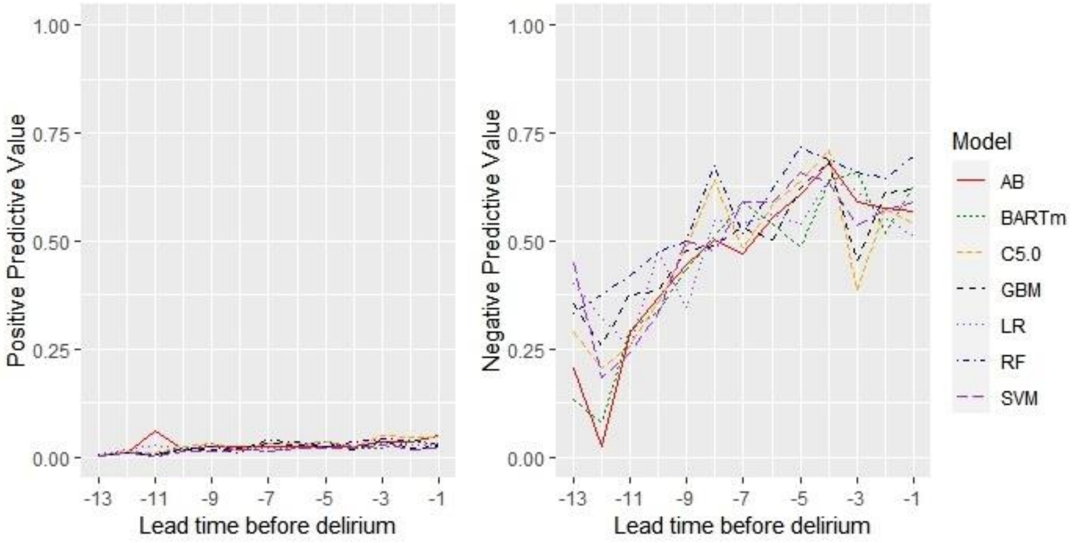


Figure 8.8 Positive and negative predictive values (PPV and NPV) for all models for each lead time when predicting delirium on an hourly basis in the ICU, using complete data.



Looking at the Table 8.3.A. from Appendix 8.3, BARTm had the highest performance when predicting delirium 13 hours in advance (AUC = 0.997). That being said, all models had very high AUC at this lead time, the lowest having been 0.959 for logistic regression. The highest sensitivity when predicting delirium 13 hours in advance belonged to gradient boosting model, support vector machine and C5.0 equally (Sens

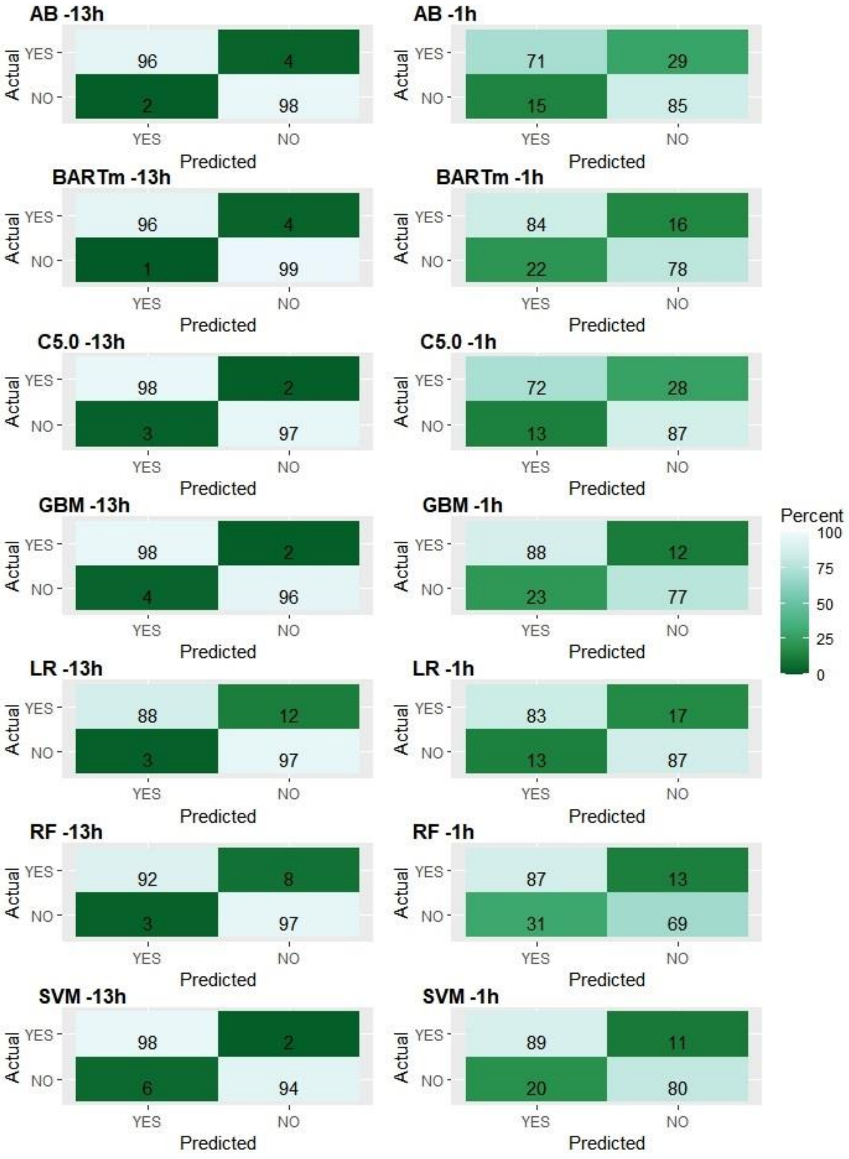
= 0.979). BARTm also had the highest specificity of 0.989 at 13 hours before the onset of delirium.

All models were still performing very highly when predicting delirium 8 hours in advance. The AUC ranged from 0.920 (random forest) to 0.942 (gradient boosting model). Random forest also had the highest sensitivity of 0.936 and support vector machine had the highest specificity of 0.907 when predicting delirium 8 hours in advance.

At 4 hours before the onset of delirium, the models had slightly lower performance overall, but was still moderately high. The performance ranged from 0.850 (random forest) to 0.905 (BARTm). Gradient boosting model had the highest sensitivity of 0.870 and support vector machine had the highest specificity of 0.832 when predicting delirium 4 hours in advance.

When looking at confusion matrices (Figure 8.9), all models were very good at predicting patients to not have delirium, regardless whether they were predicting delirium 13 hours in advance or 1 hour in advance. The proportion of predicting no delirium for 1 hour in advance was slightly lower than it was for 13 hours in advance. This was most likely because the prevalence of delirium increased as the lead times got closer to the event. Hence, due to slightly higher prevalence at 1 hour before delirium, the models were less confident about their prediction than they were at 13 hours in advance. Furthermore, due to larger number of patients at lead times closer to delirium, there was more variance within the laboratory measures (shown in Appendix 8.2), which might affect why models performed less well at lead times closer to delirium.

Figure 8.9. Confusion matrices for all models predicting delirium using complete data 13 hours vs 1 hour in advance.



8.4.3. Models' Calibration and Variable Importance

Since the support vector machine model had the highest overall mean performance (AUC = 0.941), calibration for this model was assessed. Figure 8.10 shows the calibration plots for the SVM model at lead times of -1, -8 and -13 hours before delirium. Visually, the model predicting delirium 8 hours in advance seemed to have the best calibration, where the predicted probabilities appeared to be more in accordance with the lower true probabilities. When predicting delirium 13 hours in advance, the estimated predicted probabilities had very high uncertainty, as shown by wide confidence intervals. This was because of the prevalence of delirium having been considerably low in this patient population, as shown in Table 8.3.

Figrue 8.10. Calibration plots for SVM predicting the onset of delirium 1 hour (top-left), 8 (top-right) and 13 (bottom-centre) hours in advance. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.

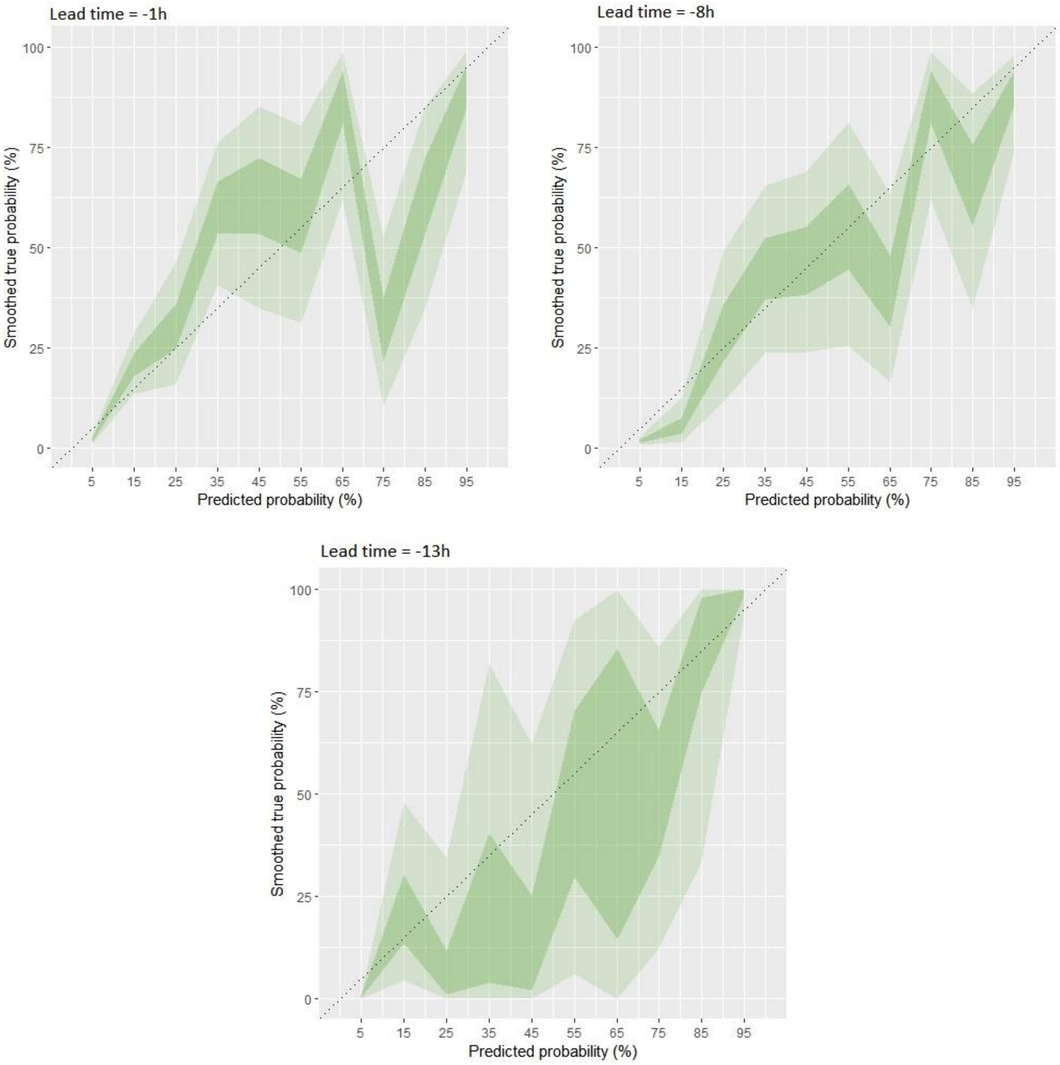


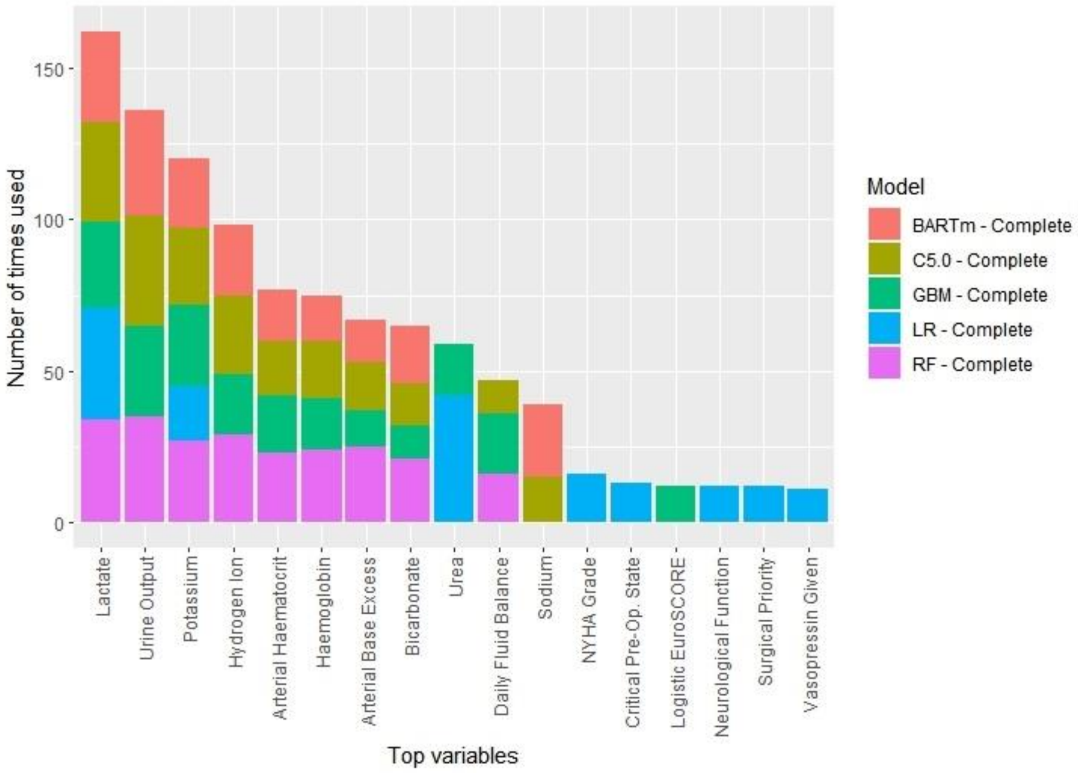
Table 8.3. Number of patients, the proportion of patients with delirium and mean predicted probability with standard deviation (SD) for the lead times of -13h, -8h and -1h for the SVM model.

Lead Time	Number of Patients	Delirium (%)	Mean Predicted Probability (%) +- SD (%)
-13	2149	6.51	7.20 +- 23.26
-8	2269	11.40	11.55 +- 22.59
-1	2307	12.60	13.01 +- 20.92

The Table 8.3 shows that the SVM model tended to overestimate risk for delirium when predicting it 13, 8 or 1 hour in advance. In addition, as also seen in calibration plots, the predicted probabilities were highly variable, as shown by standard deviation.

The Figure 8.11 shows which variables the models used the most to predict delirium across all lead times. For each lead time the top 20 variables were extracted, and then combined for all models and lead times to understand which variables were most commonly used overall. When using complete data for delirium prediction, the most common variables used by models were lactate, urine output, potassium, and hydrogen ion. All models used lactate and potassium in their prediction. Most models, apart from logistic regression, used urine output, hydrogen ion, arterial haematocrit, haemoglobin, arterial base excess and bicarbonate.

Figure 8.11. Most important variables used by models for at least 10 times.



8.5. Experiment 2 and 3 Results: Predicting Delirium, Using Complete Training Data and Incomplete Testing Data

In this section, the results of Experiment 2 and 3 are presented. These experiments are:

> **Experiment 2:** Predicting delirium on an hourly basis, using complete training data and missing values in testing data.
>
> **Experiment 3:** Predicting delirium on an hourly basis, using complete training data and imputation methods to replace missing values in testing data.

In Experiment 2, two methods that handle missing data were used to predict delirium, using complete training data and incomplete testing data. These two methods are *BARTm* and *C5.0*.

In Experiment 3, the following imputation methods were used to replace missing values in testing data: *median imputation*, *0 imputation*, and *missForest imputation*. For the model development, same methods were used as in Experiment 1.

All methods used in this section are further described in Chapter 5.

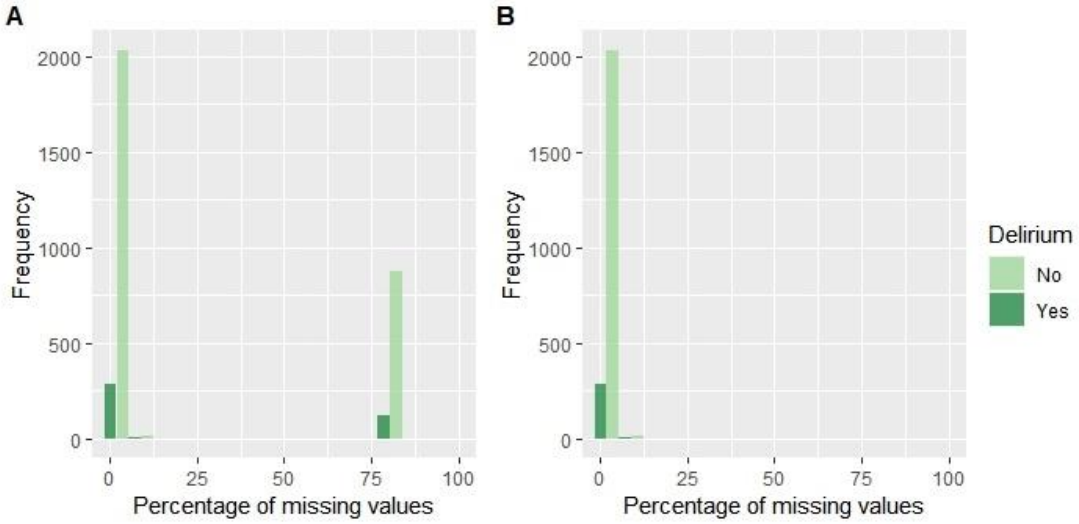
8.5.1. Data Preparation

For both experiments, to avoid excessive number of missing values, patient records with more than 40% of missing values were excluded from the analysis, as done in other studies in the literature [300], [302].

The histograms (Figure 8.12) show that the majority of patients had very little or no missing data, however a considerable number of patients also had more than 60% of missing values. A similar proportion of patients who were diagnosed with delirium, compared to the total patient population, had missing values. After removing patients with >40% of missing variables, 122 patients with delirium were excluded from the analysis.

The completeness of data was relatively high, following the removal of patient records with more than 40% of missing values.

Figure 8.12. Histogram of proportion of missing data for all patient records (A) vs for patient records when records with >40% of missing values were removed (B), where light-green and dark-green indicate patients without and with delirium, respectively.



8.5.1.1. Descriptive Statistics

In the Table 8.2.A from Appendix 8.2 descriptive statistics are shown for each laboratory variable to show how the imputation methods affected the distribution of the data. As the laboratory variables in the modelling experiments were used as minimum, maximum, first and last per time window, the mean and standard deviation values were derived for each variable. In addition, paired t-tests were used to assess whether the records with imputation methods were significantly different from the original complete dataset.

In most cases, neither of the imputation methods affected the distribution of the variables. This might be because the completeness of data was relatively high (see Table 8.4). The only statistically significant difference was found between the original maximum and last serum bicarbonate compared to these variables with 0 imputation.

8.5.1.2. Training and Testing Datasets

As shown in Table 8.4, the mean number of patients across all lead times was 2290 (SD = 37.93), completeness was 98.94% (SD = 0.14) and the proportion of patients with delirium was 11.25% (SD = 1.52). In the test set the overall completeness of data was 96.82% (SD = 0.48), which means that the BARTm and C5.0 algorithms had to handle around 4% of the patients' missing data.

The completeness of data increased and the percentage of patients with delirium increased as the lead time got closer to delirium. This was because the majority of patients have delirium between 10-13 hours since ICU admission.

Table 8.4. Number of patients in each training and testing data based on the lead time until delirium within 21 hours in ICU, completeness of data and percentage of patients with delirium.

	Total Data			**Training data (100% complete)**		**Testing data**		
Lead Time	**Patients (n)**	**Completeness (%)**	**Delirium (%)**	**Patients (n)**	**Delirium (%)**	**Patients (n)**	**Completeness (%)**	**Delirium (%)**
-13	2178	98.70	6.70	1451	7.10	727	96.00	5.91
-12	2262	98.70	10.20	1507	11.00	755	96.20	8.48
-11	2273	98.80	10.60	1514	11.20	759	96.30	9.49
-10	2278	98.80	10.80	1518	11.50	760	96.30	9.47
-9	2289	98.90	11.20	1525	11.00	764	96.60	11.60
-8	2294	98.90	11.40	1529	10.30	765	96.70	13.60
-7	2297	99.00	11.50	1531	11.20	766	96.90	12.30
-6	2304	99.00	11.80	1535	11.70	769	97.00	12.00
-5	2309	99.00	12.00	1539	11.20	770	97.10	13.50
-4	2315	99.10	12.20	1542	12.20	773	97.30	12.30
-3	2321	99.10	12.50	1547	11.80	774	97.40	13.70
-2	2327	99.10	12.70	1550	12.60	777	97.40	12.90
-1	2327	99.10	12.70	1550	12.60	777	97.40	12.90
Mean ± SD	2290 ± 37.93	98.94 ± 0.14	11.25 ± 1.52	1526 ± 25.31	11.18 ± 1.34	764 ± 12.63	96.82 ± 0.48	11.40 ± 2.27

8.5.2. Experiment 2: Models' Performance

8.5.2.1. Discriminative Performance

From Figure 8.13 it can be seen that BARTm had a higher overall performance at all lead times, compared to C5.0. Interestingly, the models had a very similar pattern of changing the performance as the lead time changed. Both models had quite high AUC, staying above 0.800 at all lead times. Similarly to the Experiment 1, here also the performance of models reduced as the lead time got closer to the onset of delirium.

The Table 8.5 shows the mean and standard deviation performance measures for both models across all lead times. As seen from Figure 8.13, BARTm had a higher mean AUC than C5.0 (AUC = 0.930 vs 0.898). BARTm also had a higher sensitivity than

C5.0 (Sens = 0.885 vs 0.822), however lower specificity than C5.0 (Spec = 0.851 vs 0.854). As seen in the Experiment 1, here also the positive and negative predictive values were generally low for both models.

Figure 8.13. AUC for both models for each lead time when predicting delirium on an hourly basis in the ICU, using complete training data and missing values in testing data.

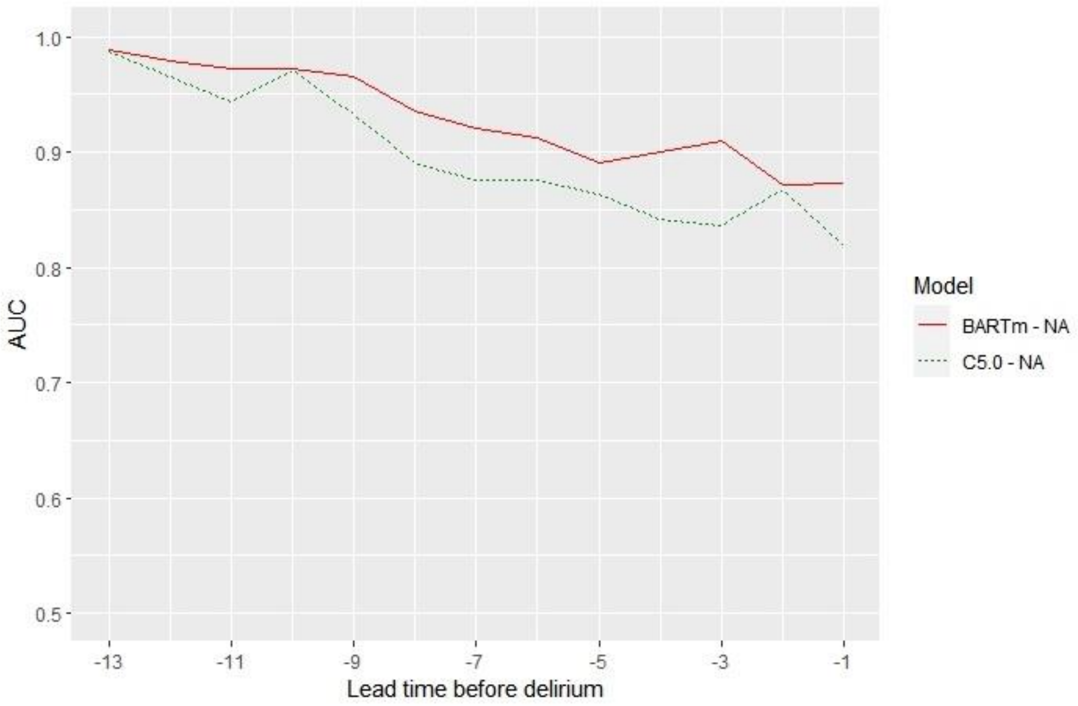
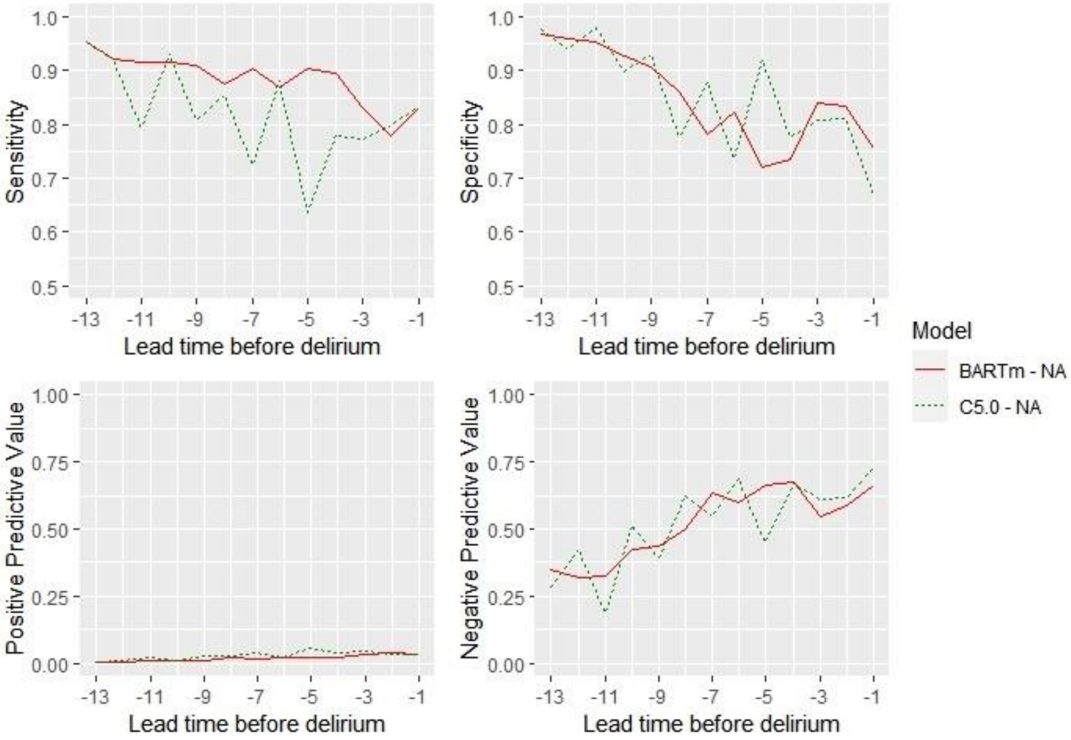


Table 8.5. Mean and standard deviation model performance measures for each model across each lead time before delirium when predicting delirium within 21h since ICU admission, using complete training data and missing values in testing data. The highest result for each performance measure is marked in bold.

	AUC	Sensitivity	Specificity	PPV	NPV
Model	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
BARTm - NA	**0.930 ± 0.041**	**0.885 ± 0.047**	0.851 ± 0.087	0.019 ± 0.010	**0.519 ± 0.133**
C5.0 - NA	0.898 ± 0.056	0.822 ± 0.089	**0.854 ± 0.098**	**0.028 ± 0.016**	0.517 ± 0.165

The Figure 8.14 shows how the sensitivity, specificity, positive and negative predictive values for both models changed with each lead time as the prediction got closer to the event of delirium. Similarly to AUC, BARTm appeared to have a higher sensitivity at most lead times, compared to C5.0. The specificity appeared to be quite similar for both models, and so do PPV and NPV.

Figure 8.14. Sensitivity and specificity, positive and negative predictive values (PPV and NPV) for both models for each lead time when predicting delirium on an hourly basis in the ICU, using complete training data and missing values in testing data.



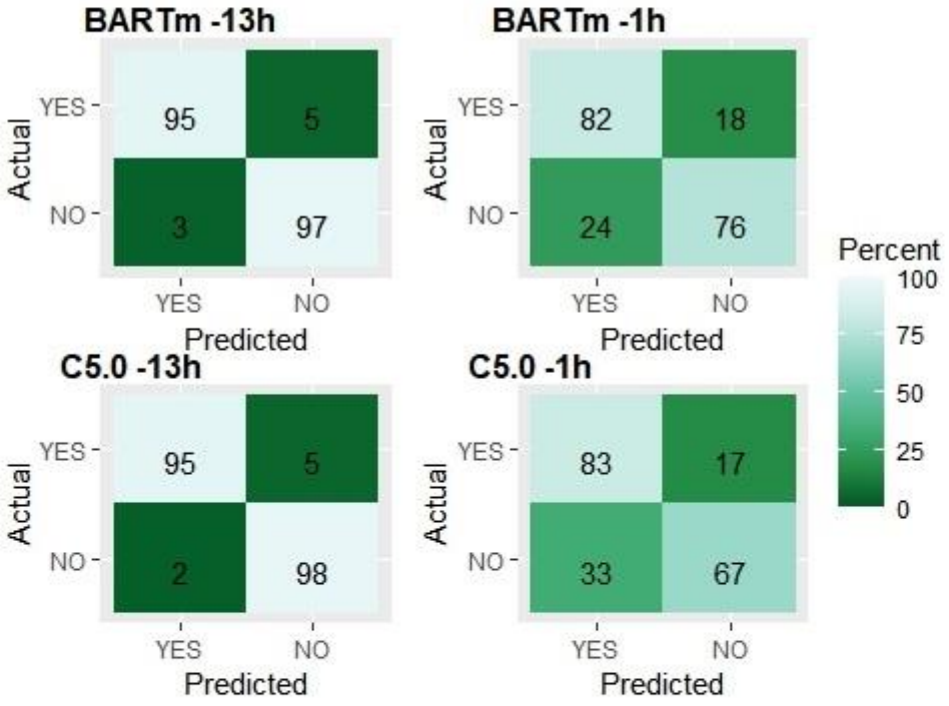
Looking at the Table 8.3.B from Appendix 8.3, when predicting delirium 13 hours in advance, the models had a very similar performance: AUC = 0.989 and 0.988 for BARTm and C5.0, respectively. Both models had an equal sensitivity of 0.953, however C5.0 had a higher specificity of 0.977 (vs 0.968).

When predicting delirium 8 hours in advance, the differences in overall model performances become clearer: BARTm had a substantially higher AUC than C5.0 (AUC = 0.935 vs 0.891), slightly higher sensitivity (0.875 vs 0.856) and remarkably higher specificity than C5.0 (0.861 vs 0.776).

At 1 hour before the occurrence of delirium, BARTm had a very high AUC of 0.900 (vs 0.842 for C5.0) and higher sensitivity than C5.0 (0.895 vs 0.779). C5.0, however, had a higher specificity of 0.777 (vs 0.736).

When looking at confusion matrices (Figure 8.15), the results were very similar to what they were when predicting delirium with complete data only within respective lead times. Similarly to Experiment 1, the models predicted a higher proportion of patients into correct classes 13 hours before delirium than they do at 1 hour before delirium. This also is in accordance with the Figure 8.13, where the overall performance of the models decreased as the lead times got closer to the event of delirium. This can be explained by a higher variability of the laboratory values at later lead times, as shown in Appendix 8.4.

Figure 8.15. Confusion matrices for both models predicting delirium 13 hours vs 1 hour in advance, using complete training data and missing values in testing data.



8.5.2.2. Calibration and Variable Importance

Since BARTm model had the highest mean AUC of 0.930, calibration for this model was assessed (Figure 8.16). The calibration when predicting delirium for this model was visibly better than it was for the best-performing model (SVM) using complete data only. The estimated values of predicted probabilities were with considerably narrow confidence intervals, indicating some certainty of the model when predicting the outcome, especially when predicting delirium 1 hour or 8 hours in advance. The model was considerably less certain about the predicted probabilities when predicting delirium 13 hours in advance due to smaller number of patients with delirium at this lead time.

According to the mean predicted probability (Table 8.6), the model considerably underestimated overall risk for delirium in the patient population when predicting delirium 13 hours in advance. The estimation of risk was considerably accurate when predicting delirium 8 hours or 1 hour in advance. The standard deviation for each mean predicted probability, however, was very high, indicating variation in predicted probabilities.

Figure 8.16. Calibration plots for BARTm evaluated on testing data that contains missing values, predicting delirium 1 hour (top-left), 8 hours (top-right) and 13 hours (bottom-centre) in advance. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.

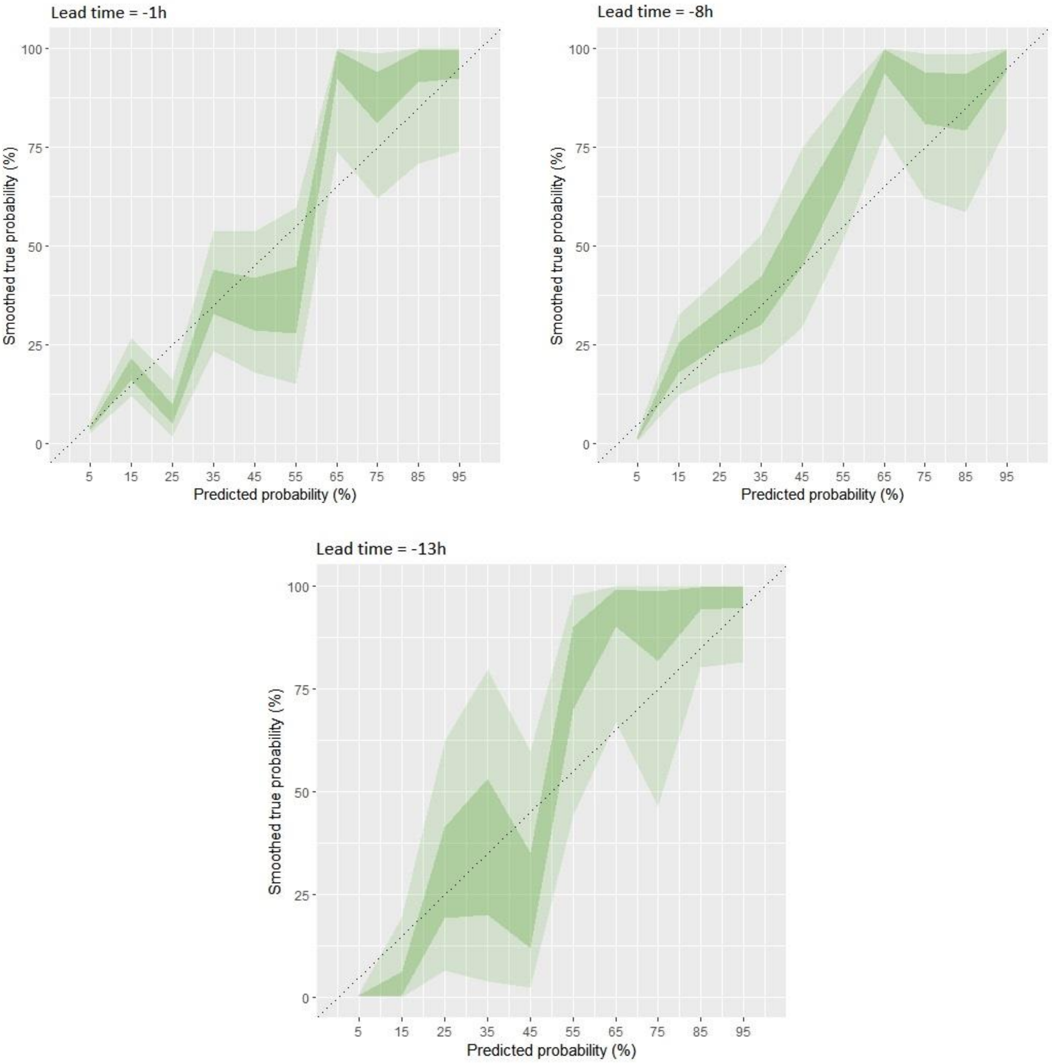
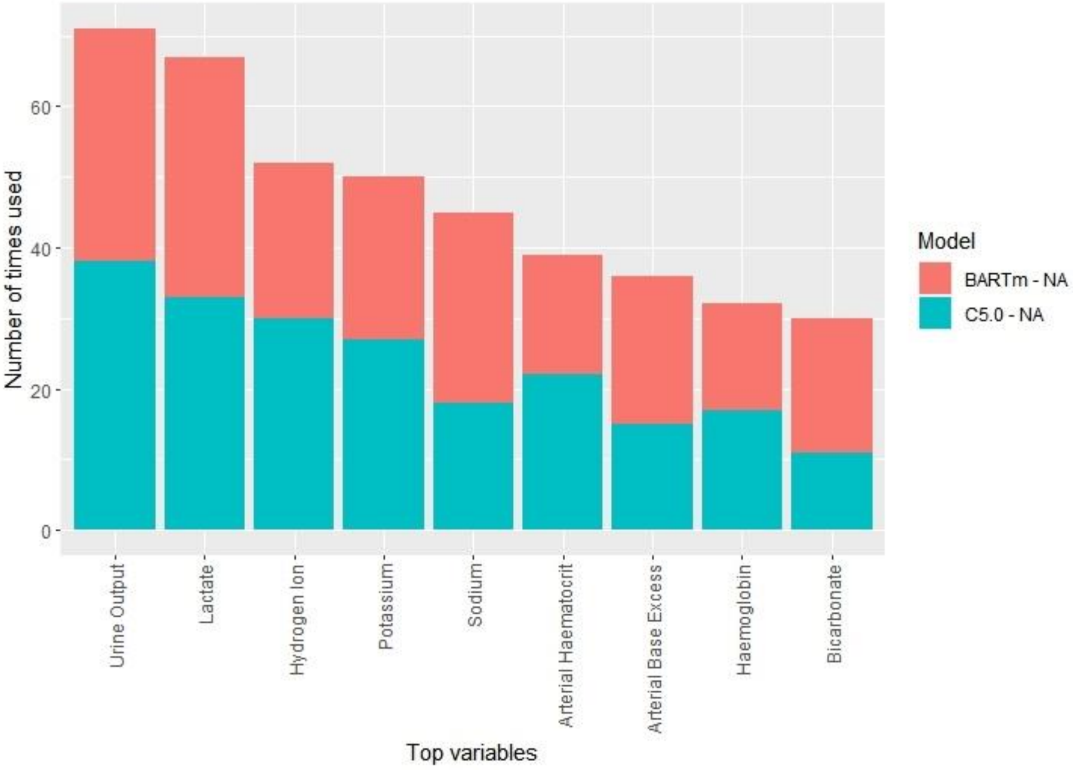


Table 8.6. Number of patients, proportion of patients with delirium and mean predicted probability with standard deviation (SD) for the BARTm model predicting delirium at lead times of -13, -8, and -1 hours.

Lead Time	**Number of patients**	**Completeness (%)**	**Delirium (%)**	**Mean Predicted Probability (%) ± SD (%)**
-13	2178	98.7	6.7	5.98 ± 18.53
-8	2294	98.9	11.4	11.82 ± 20.34
-1	2327	99.1	12.7	12.23 ± 18.96

The Figure 8.17 shows the 10 most important variables used by the two models. Both models used each of these top variables, urine output and lactate being the most popular.

Figure 8.17. Most important variables that are used in models for at least 10 times.

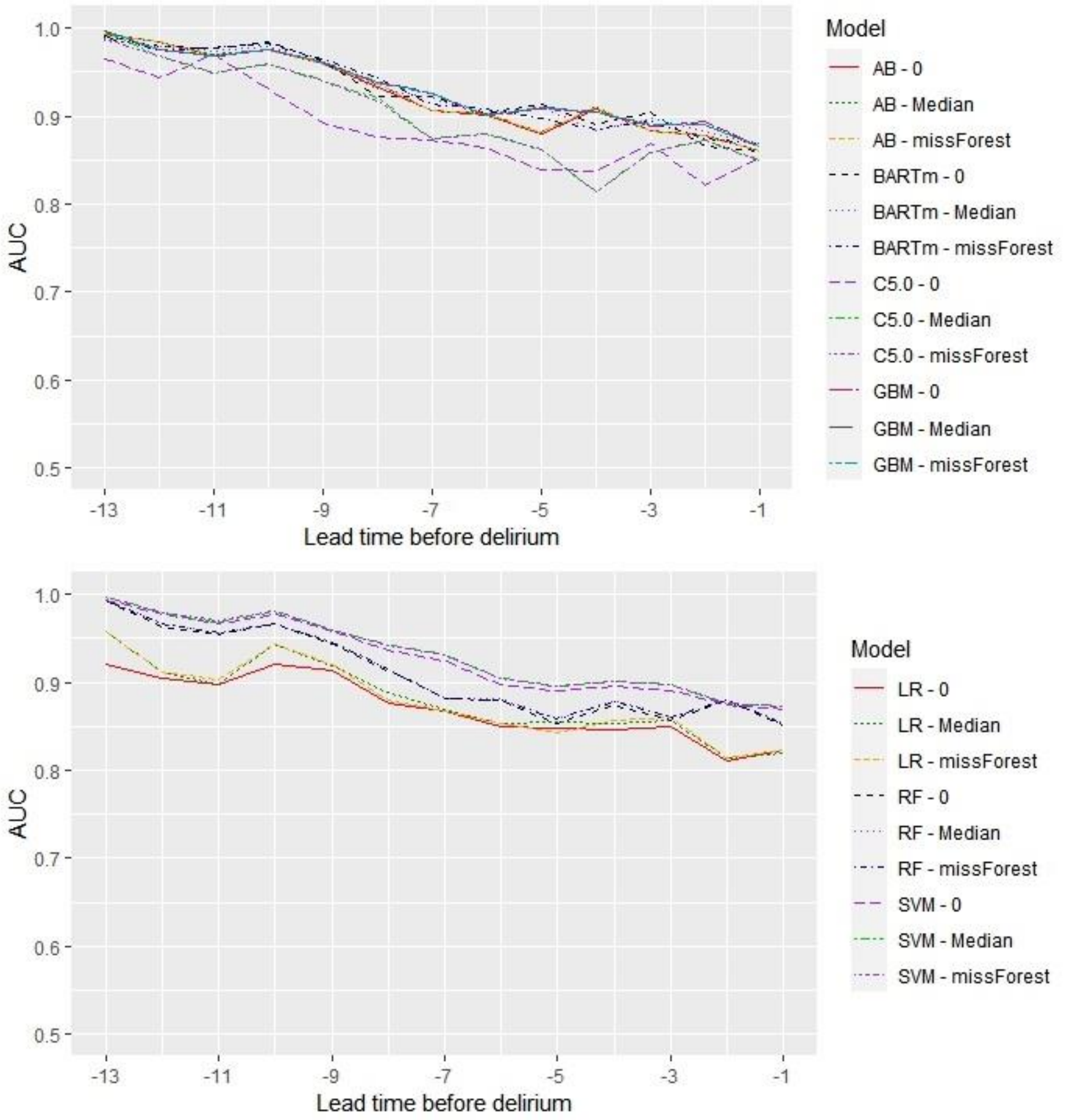


8.5.3. Experiment 3: Models' Performance

8.5.3.1. Discriminative Performance

The Figure 8.18 shows how the overall performance for each model changed with the lead time getting closer to the onset of delirium. Visually, all models, regardless of imputation method, had moderately good to very good performance. At all lead times, all models stayed above 0.800 in terms of AUC. As seen in Experiments 1 and 2, the models here also reduced in performance as the lead time got closer to delirium. As explained in earlier experiments, this is most likely due to the laboratory variables' variation increasing as the lead time got closer to delirium (shown in Appendix 8.4).

Figure 8.18. AUC for all models for each lead time when predicting delirium on an hourly basis in the ICU, using complete training data and imputation methods in testing data.



The Table 8.7 shows the mean performance measures for each model across all lead times. Most models tended to have mean AUC above 0.900, apart from C5.0 with 0 imputation, and logistic regression models. Gradient boosting model with all imputation methods and support vector machine with median and missForest imputation had equally the highest mean AUC of 0.931. In terms of sensitivity, BARTm with 0 imputation had the highest mean sensitivity of 0.876. BARTm with missForest imputation had the highest mean specificity of 0.892, directly followed by C5.0 (mean Spec = 0.890).

As seen in previous experiments, the positive and negative predictive values were very low for all models. This was most likely due to the low prevalence of delirium in the dataset, but also due to the models having been uncertain about the predicted probabilities corresponding to whether a patient had a delirium.

Table 8.7. Mean and standard deviation model performance measures for each model across each lead time before delirium when predicting delirium within 21h since ICU admission, using complete training data and imputation methods to replace missing values in testing data. The highest result for each performance measure is marked in bold.

	AUC	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Model	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
AB - 0	0.926 ± 0.046	0.860 ± 0.084	0.876 ± 0.070	0.022 ± 0.015	0.468 ± 0.178
AB - Median	0.926 ± 0.046	0.865 ± 0.083	0.871 ± 0.076	0.022 ± 0.015	0.490 ± 0.143
AB - missForest	0.926 ± 0.046	0.865 ± 0.083	0.870 ± 0.076	0.022 ± 0.015	0.491 ± 0.141
BARTm - 0	0.929 ± 0.045	**0.876 ± 0.062**	0.869 ± 0.070	0.020 ± 0.012	0.502 ± 0.122
BARTm - Median	0.929 ± 0.042	0.856 ± 0.081	0.882 ± 0.072	0.022 ± 0.014	0.463 ± 0.161
BARTm - missForest	0.929 ± 0.044	0.850 ± 0.080	**0.892 ± 0.061**	0.023 ± 0.015	0.453 ± 0.144
C5.0 - 0	0.887 ± 0.050	0.815 ± 0.075	0.830 ± 0.109	**0.030 ± 0.013**	0.543 ± 0.170
C5.0 - Median	0.904 ± 0.055	0.807 ± 0.106	0.889 ± 0.057	**0.030 ± 0.019**	0.484 ± 0.135
C5.0 - missForest	0.903 ± 0.054	0.803 ± 0.099	0.890 ± 0.059	**0.030 ± 0.018**	0.477 ± 0.144
GBM - 0	0.931 ± 0.041	0.858 ± 0.061	0.880 ± 0.068	0.022 ± 0.012	0.481 ± 0.134
GBM - Median	0.931 ± 0.041	0.860 ± 0.057	0.876 ± 0.071	0.022 ± 0.011	0.487 ± 0.144
GBM - missForest	0.931 ± 0.041	0.861 ± 0.057	0.875 ± 0.071	0.022 ± 0.011	0.487 ± 0.143
LR - 0	0.871 ± 0.037	0.807 ± 0.057	0.873 ± 0.057	0.029 ± 0.013	0.518 ± 0.115
LR - Median	0.880 ± 0.044	0.815 ± 0.068	0.875 ± 0.054	0.028 ± 0.014	0.518 ± 0.104
LR - missForest	0.879 ± 0.044	0.814 ± 0.066	0.878 ± 0.056	0.028 ± 0.014	0.504 ± 0.128
RF - 0	0.909 ± 0.049	0.840 ± 0.086	0.852 ± 0.083	0.026 ± 0.015	0.535 ± 0.139
RF - Median	0.910 ± 0.049	0.852 ± 0.076	0.845 ± 0.077	0.025 ± 0.015	0.551 ± 0.134
RF - missForest	0.910 ± 0.049	0.852 ± 0.076	0.844 ± 0.076	0.025 ± 0.015	**0.552 ± 0.131**
SVM - 0	0.927 ± 0.043	0.874 ± 0.075	0.869 ± 0.057	0.021 ± 0.014	0.519 ± 0.092
SVM - Median	**0.931 ± 0.043**	0.874 ± 0.075	0.880 ± 0.058	0.020 ± 0.014	0.491 ± 0.109
SVM - missForest	**0.931 ± 0.043**	0.874 ± 0.075	0.880 ± 0.057	0.020 ± 0.014	0.493 ± 0.105

The Figures 8.19 and 8.20 show how the sensitivity and specificity changed for the models as the lead time got closer to delirium. In general, for both sensitivity and specificity, the performance got lower as the onset of delirium got closer. Visually, C5.0 with 0 imputation and gradient boosting model with median imputation seemed to have the lowest sensitivity at most lead times. C5.0 with 0 imputation also had visually the lowest specificity at most lead times. That being said, for most lead times, models tended to stay above 0.700 for sensitivity and above 0.750 for specificity.

Looking at the Tables 8.3.C, 8.3.D and 8.3.E from Appendix 8.3, when predicting delirium 13 hours in advance, all models had very high AUC, all being above 0.950, apart from the logistic regression model with 0 imputation. The highest AUC of 0.997 belonged to four models: C5.0 with median imputation and gradient boosting model with all three imputation methods. Nine models had the highest sensitivity of 0.999: C5.0 with median imputation, gradient boosting model with median and missForest imputation, support vector machine with all three imputation methods, and random forest with all three imputation methods. All models also had very high specificity at this lead time, staying above 0.940. The highest specificity belonged to BARTm with median imputation (0.987).

When predicting delirium 8 hours in advance, BARTm had the highest AUC of 0.943. Six models had equally the highest sensitivity of 0.894: support vector machine with all three imputation methods and AdaBoost with all three imputation methods. Random forest with median and missForest imputation had the highest specificity of 0.908 when predicting delirium 8 hours in advance.

As seen in previous experiments, as the lead time got closer to the event of delirium, the performance measures got slightly lower. When predicting delirium 4 hours in advance, the AUC ranged from 0.815 (C5.0 with median and missForest) to 0.910 (AdaBoost with median and missForest). AdaBoost and support vector machine with all three imputation methods again had the highest sensitivity of 0.853. BARTm with median imputation, however, had the highest specificity of 0.898 when predicting delirium 4 hours in advance.

Figure 8.19. Sensitivity and specificity for AdaBoost, BARTm, C5.0 and gradient boosting models for each lead time when predicting delirium on an hourly basis in the ICU, using complete training data and imputation methods in testing data.

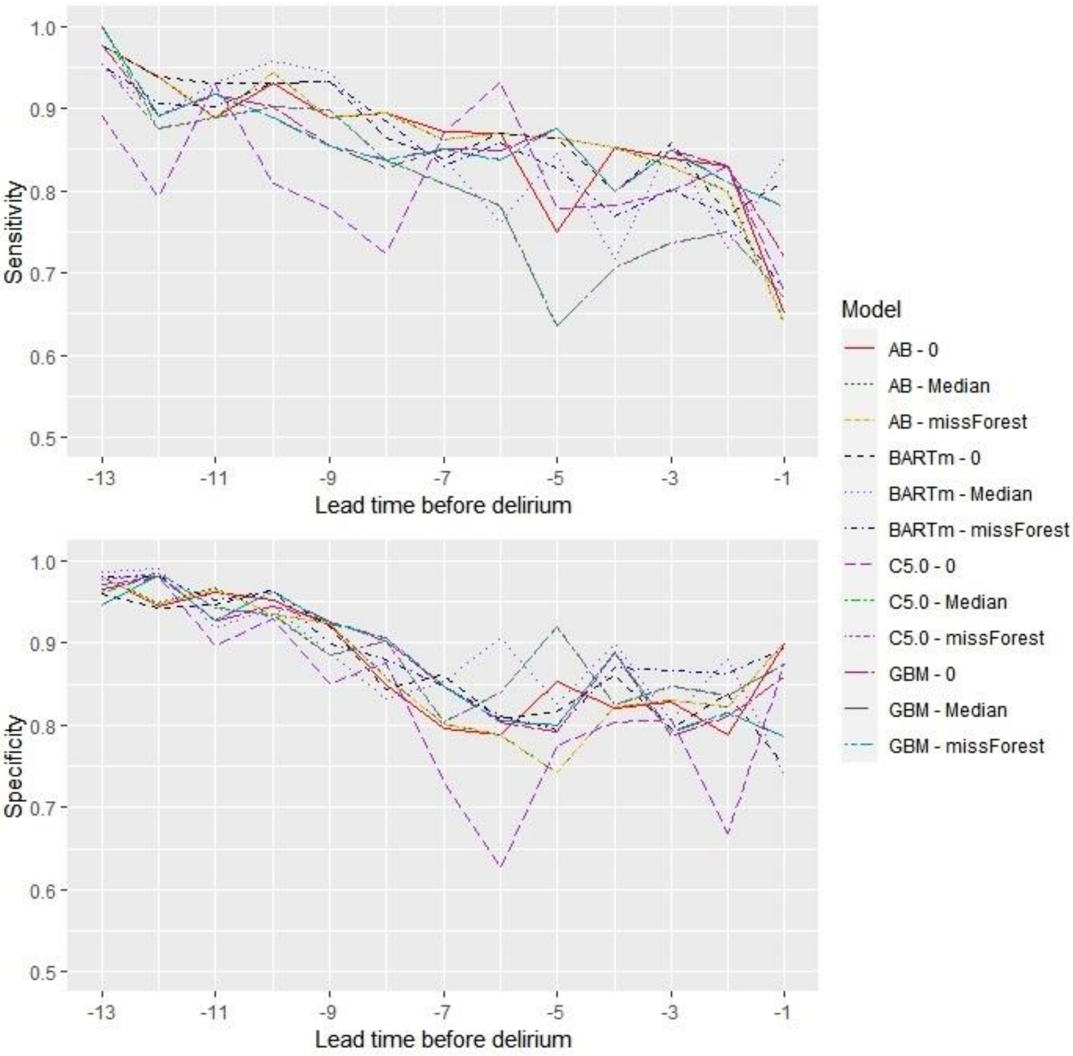


Figure 8.20. Sensitivity and specificity for logistic regression, random forest and support vector machine models for each lead time when predicting delirium on an hourly basis in the ICU, using complete training data and imputation methods in testing data.

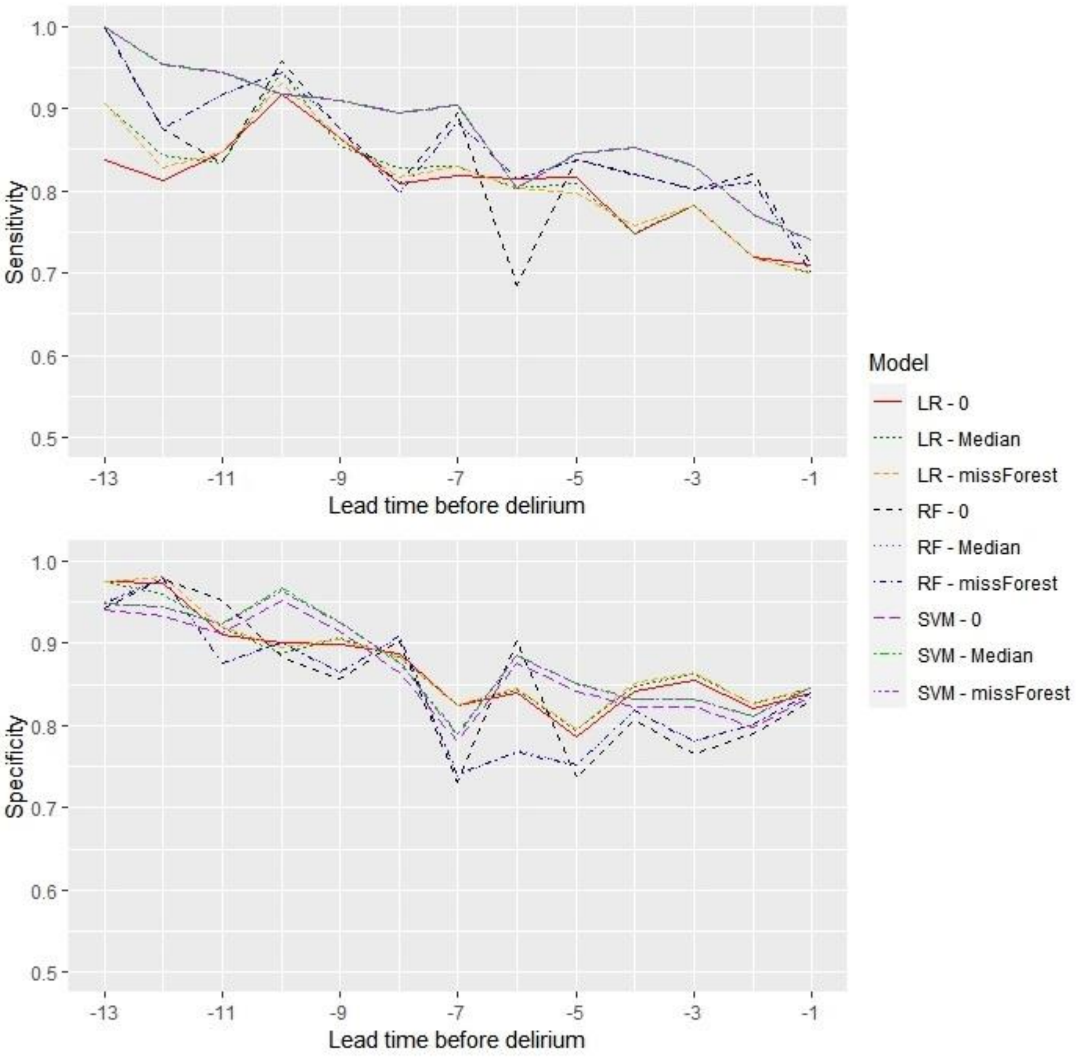


Figure 8.21. Positive predictive values (PPV) for all models for each lead time when predicting delirium on an hourly basis in the ICU, using complete training data and imputation methods in testing data.

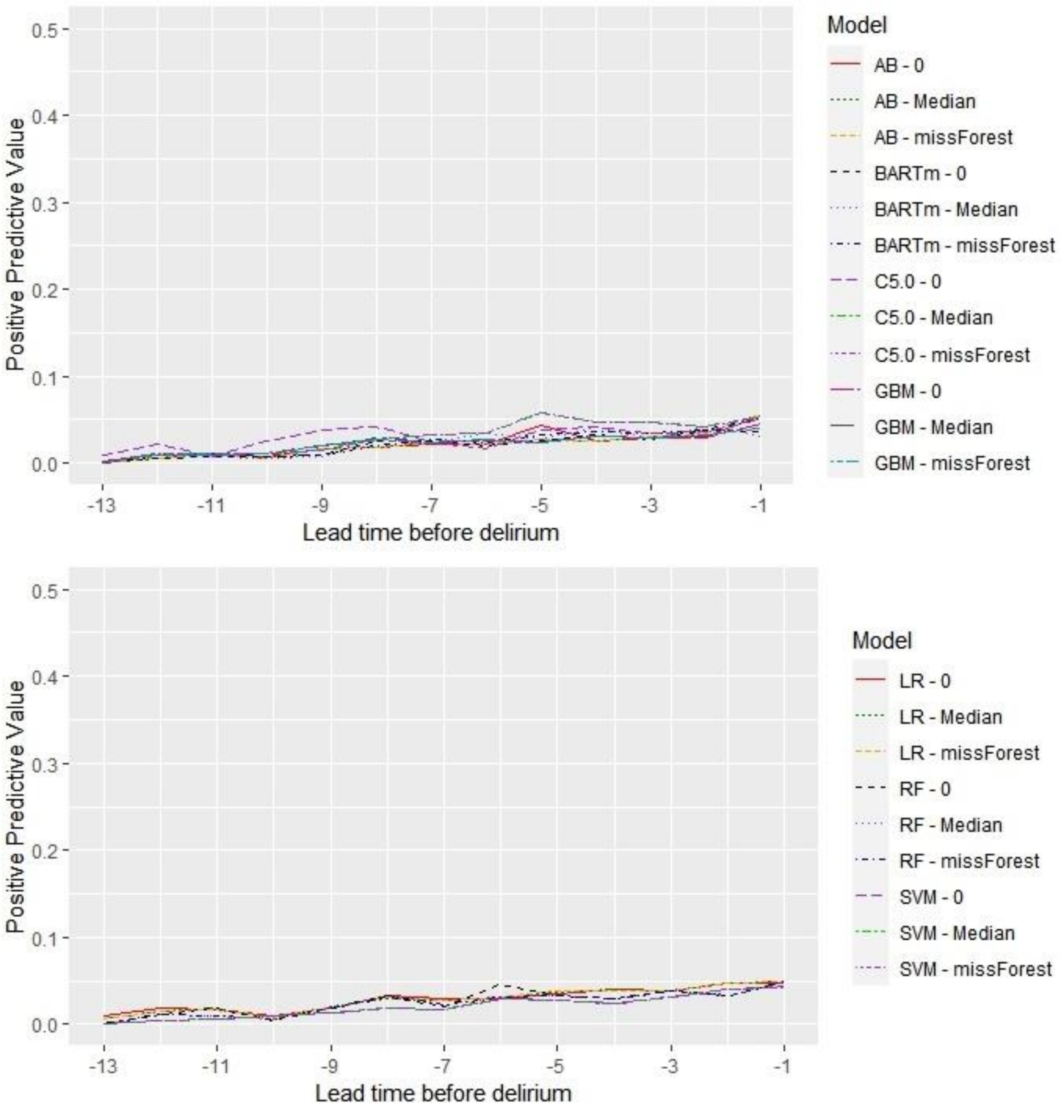
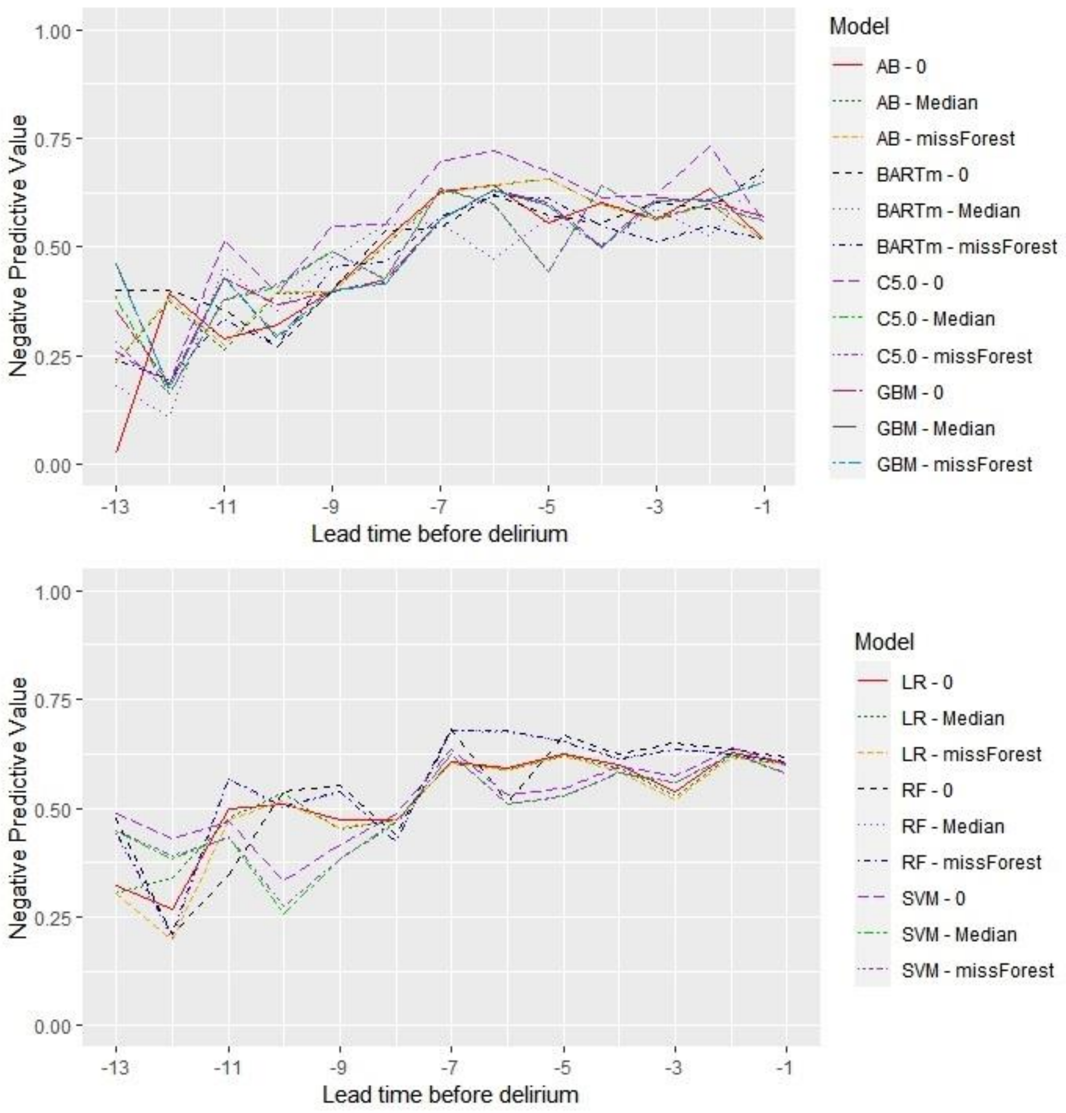


Figure 8.22. Negative predictive values (NPV) for all models for each lead time when predicting delirium on an hourly basis in the ICU, using complete training data and imputation methods in testing data.



The Figures 8.21 and 8.22 show how positive and negative predictive values changed for each model as the lead times got closer to the event of delirium. In general, for all models, as also seen in previous experiments, positive predictive values were very low. In addition to being related to relatively low prevalence of delirium, as explained before, it also indicates the models' low certainty about predicted probabilities related to patients who actually had delirium. For negative predictive value, the performance got higher as the lead time got closer to delirium, however, in general, the negative predictive values are as good as flipping a coin at their highest.

Figure 8.23. Confusion matrices for all models predicting delirium 13 hours vs 1 hour in advance, using complete training data and median imputation (A) and 0 imputation (B) in testing data.

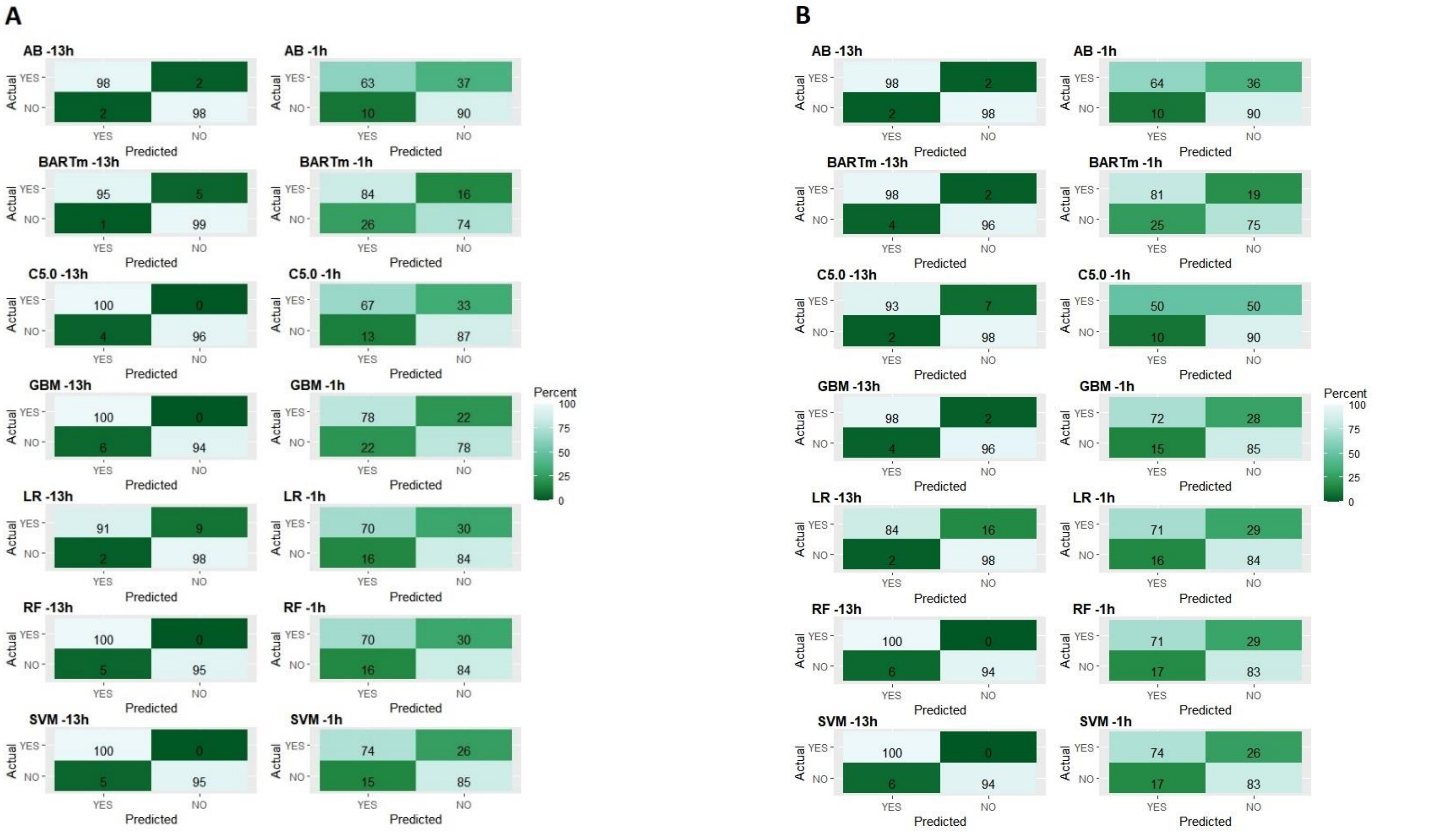
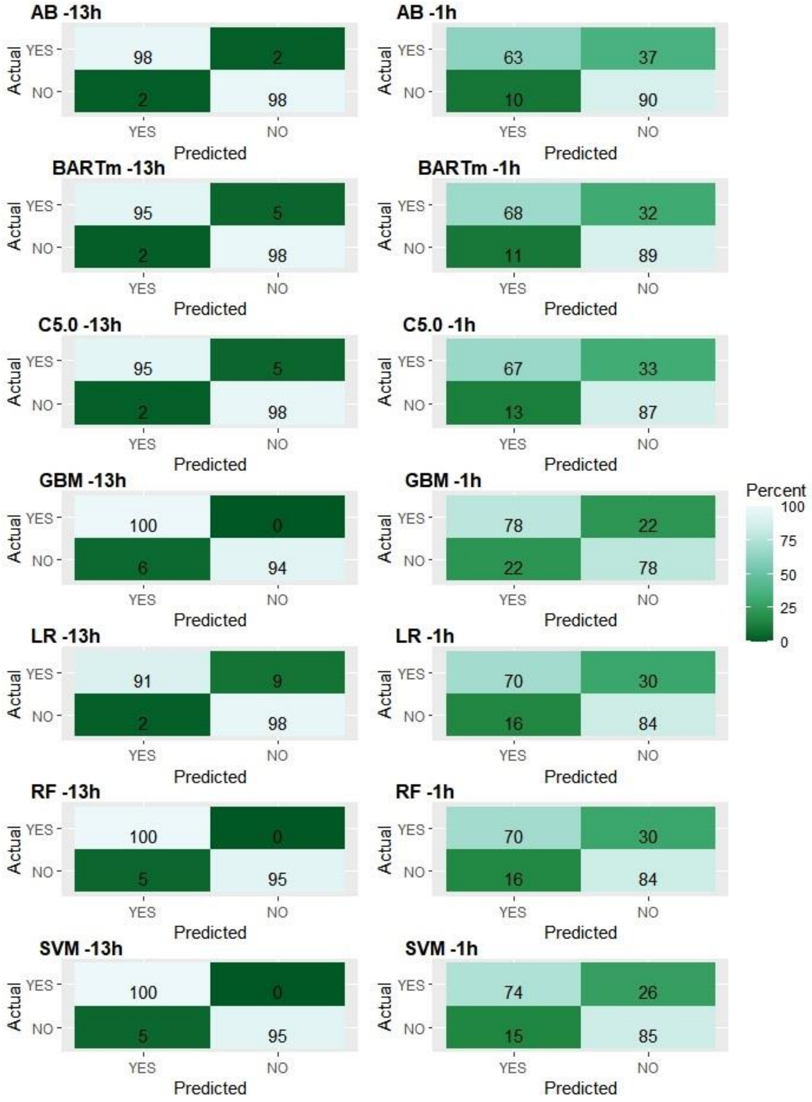


Figure 8.24. Confusion matrices for all models predicting delirium 13 hours vs 1 hour in advance, using complete training data and missForest imputation in testing data.



The Figures 8.23 and 8.24 show confusion matrices for each method predicting delirium 13 hours and 1 hour in advance. In general, all models did very well at grouping patients to have or not have delirium correctly when making the prediction 13 hours in advance. The models were substantially less capable at predicting the categories correctly when predicting delirium 1 hour in advance. This is also reflected in the reduction in overall performance of all models as the lead times got closer to delirium. Across all three imputation methods the models had very similar performance. The best performing models appeared to be BARTm with median and 0 imputation due to being able to predict patients with delirium well for both 13 and 1 hour in advance. Interestingly, with missForest imputation at 1 hour in advance, BARTm was not very good at categorising patients with delirium correctly. AdaBoost and C5.0 for all three imputation methods tended to be better at predicting patients without delirium than patients with delirium.

8.5.3.2. Calibration and Variable Importance

As the gradient boosting model with 0 imputation in testing set achieved the highest mean overall performance (AUC = 0.931), calibration for this model was assessed (Figure 8.25). For the three lead times, when the true probability of delirium was less than 50%, the model visibly underestimated risk of delirium. As the true probability went over 50%, the predicted probabilities became highly variable, and mostly showed overestimation of risk of delirium.

The variability of the predicted probabilities can be explained by considerably small number of patients with delirium in the patient population in -13h lead time. However, unlike in previous experiments, in this experiment, at lead times close to delirium, the probabilities were unexpectedly variable, which might be explained by the 0 imputation, which can affect the predicted probabilities in the patient population.

According to the mean predicted probabilities (Table 8.8), the model underestimated the risk of delirium at 13 hours before delirium (mean probability = 5.98% vs 6.70% prevalence of delirium) and at 1 hour before delirium (mean probability = 12.23% vs 11.40% prevalence of delirium), which is also in accordance with the Figure 8.25.

Figure 8.25. Calibration plots for gradient boosting model predicting delirium 1 hour (top-left), 8 hours (top-right) and 13 hours (bottom-centre) in advance, evaluated on testing set where missing values are replaced with 0. The light-green and dark-green areas indicate the 95% CI and IQR for the predicted probabilities, respectively.

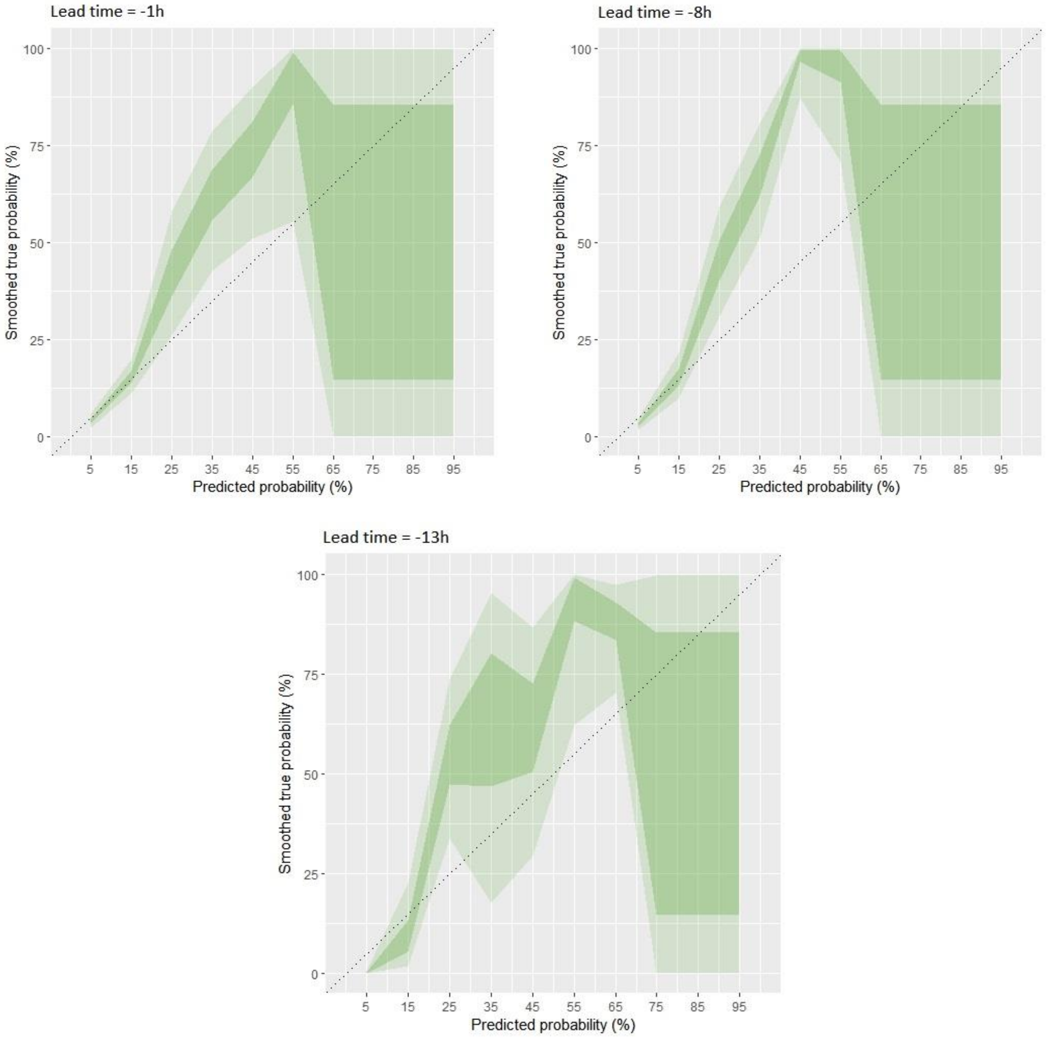
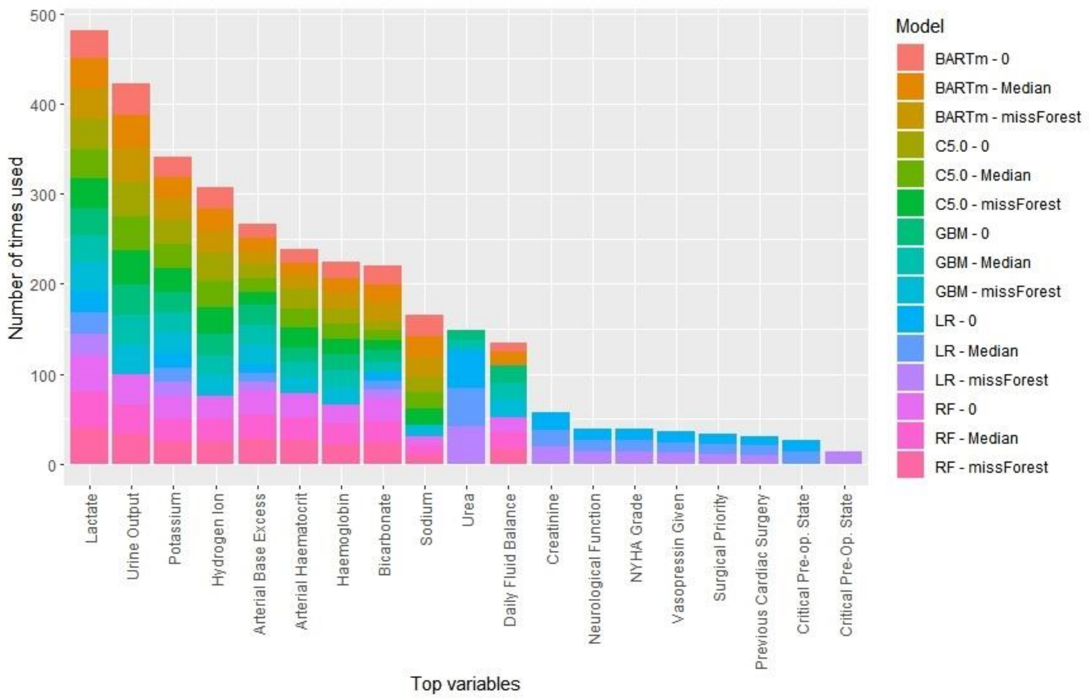


Table 8.8. Prevalence and predicted probability

Lead Time	Number of patients	Completeness (%)	Delirium (%)	Mean Predicted Probability (%) +- SD (%)
-13	2178	98.70	6.70	5.98 +- 18.53
-8	2294	98.90	11.40	11.82 +- 20.34
-1	2327	99.10	12.70	12.23 +- 18.96

The Figure 8.26 shows the most important variables in the models developed, using complete training data and imputation methods to replace missing values in testing data. All models used lactate, potassium, arterial base excess and bicarbonate in their models. Urine output, hydrogen ion, arterial haematocrit and haemoglobin were used by all models, apart from logistic regression.

Figure 8.26. Most important variables that are used by models for at least 10 times.



8.6. Discussion

8.6.1. Summary of Results

Overall, with all experiments, all models had relatively good performance in terms of AUC, sensitivity, and specificity. In terms of positive and negative predictive values, however, the performance for all models was low.

As shown in Table 8.9, based on AUC, BARTm, developed with complete data only, had the highest performance when predicting delirium 13 hours in advance (AUC = 0.997). The same model also had the highest specificity of 0.989. BARTm models in general appeared to do quite well based on AUC and sensitivity. C5.0 models tended to do well based on specificity. At 8 hours before delirium, BARTm with missForest imputation had the highest AUC (0.943). Random forest with complete data had the highest sensitivity (0.936) and random forest with median imputation had the highest specificity (0.908).

Table 8.9. Highest performance measures for each lead time if delirium occurs within 21 hours in the ICU.

Lead Time	AUC - Model	Sensitivity - Model	Specificity - Model
Highest	0.997 (-13h)	0.999 (-13h)	0.997 (-12h)
-13	0.997 - BARTm	0.999 - C5.0 – Median, GBM – Median, GBM – missForest, SVM – Median, SVM – missForest, RF – 0, RF – Median, RF – missForest, SVM - 0	0.989 - BARTm
-12	0.988 - AB	0.953 - SVM – Median, SVM – missForest, SVM - 0	0.997 - AB
-11	0.995 - SVM	0.988 - SVM	0.981 - C5.0 - NA
-10	0.984 - BARTm - 0	0.958 - BARTm - Median	0.967 - SVM - Median
-9	0.965 - BARTm – missForest, BARTm - NA	0.944 - BARTm - Median	0.940 - LR
-8	0.943 - BARTm - missForest	0.936 - RF	0.908 - RF - Median
-7	0.936 - AB	0.920 – SVM, BARTm	0.905 - C5.0
-6	0.937 - BARTm	0.933 - C5.0 - 0	0.907 - BARTm - Median
-5	0.913 - BARTm - 0	0.904 - BARTm - NA	0.920 - C5.0 – Median, C5.0 – missForest, C5.0 - NA
-4	0.910 - AB – Median, AB - missForest	0.895 - BARTm - NA	0.898 - BARTm - Median
-3	0.914 - GBM	0.872 - BARTm	0.951 - C5.0
-2	0.916 - SVM	0.901 – SVM, GBM	0.882 - BARTm - Median
-1	0.914 - SVM	0.891 - SVM	0.900 – AB – Median, AB - missForest

For all three experiments, as the lead time got closer to the event of delirium, the performance measures in general decreased. This is also noticeable in Table 8.9, where at, for example, when predicting delirium 1 hour in advance, the highest AUC is 0.914 for support vector machine with complete data. This is substantially lower than AUC for 13 hours before delirium. The same support vector machine model also had the highest sensitivity of 0.891, and AdaBoost with median and missForest imputation had equally the highest specificity of 0.900 when predicting delirium 1 hour in advance.

The decrease in models' performance as the lead time got closer to delirium was most likely due to the proportion of patients with delirium increasing, which increased the variance of the observations, and reducing the certainty of the models' predictive ability at later lead times. The changing variance of the observations as more patients are included in the prediction is evident in Appendix 8.4, where minimum, maximum, first and last laboratory values are shown as the lead time changes. The figures show a very small variation at earlier lead times, where less patients had delirium. As the lead time changed, the variance in laboratory values increased.

Table 8.10. Overall mean performance measures across all lead times for all models and experiments when predicting delirium within 21 hours in the ICU, ordered based on the highest mean AUC.

Model	**AUC** **(mean ± SD)**	**Sens** **(mean ± SD)**	**Spec** **(mean ± SD)**
Highest Overall (Model)	0.941 (SVM)	0.907 (SVM)	0.892 (BARTm – missForest)
SVM	0.941 ± 0.038	0.907 ± 0.048	0.870 ± 0.066
GBM	0.939 ± 0.035	0.875 ± 0.070	0.875 ± 0.064
BARTm	0.937 ± 0.036	0.875 ± 0.062	0.884 ± 0.069
GBM - 0	0.931 ± 0.041	0.858 ± 0.061	0.880 ± 0.068
GBM - Median	0.931 ± 0.041	0.860 ± 0.057	0.876 ± 0.071
GBM - missForest	0.931 ± 0.041	0.861 ± 0.057	0.875 ± 0.071
SVM - Median	0.931 ± 0.043	0.874 ± 0.075	0.880 ± 0.058
SVM - missForest	0.931 ± 0.043	0.874 ± 0.075	0.880 ± 0.057
BARTm - NA	0.930 ± 0.041	0.885 ± 0.047	0.851 ± 0.087
AB	0.929 ± 0.043	0.863 ± 0.060	0.877 ± 0.078
BARTm - 0	0.929 ± 0.045	0.876 ± 0.062	0.869 ± 0.070
BARTm - Median	0.929 ± 0.042	0.856 ± 0.081	0.882 ± 0.072
BARTm - missForest	0.929 ± 0.044	0.850 ± 0.080	0.892 ± 0.061
SVM - 0	0.927 ± 0.043	0.874 ± 0.075	0.869 ± 0.057
AB - 0	0.926 ± 0.046	0.860 ± 0.084	0.876 ± 0.070
AB - Median	0.926 ± 0.046	0.865 ± 0.083	0.871 ± 0.076
AB - missForest	0.926 ± 0.046	0.865 ± 0.083	0.870 ± 0.076
C5.0	0.915 ± 0.047	0.822 ± 0.097	0.885 ± 0.067
RF - Median	0.910 ± 0.049	0.852 ± 0.076	0.845 ± 0.077
RF - missForest	0.910 ± 0.049	0.852 ± 0.076	0.844 ± 0.076
RF - 0	0.909 ± 0.049	0.840 ± 0.086	0.852 ± 0.083
RF	0.907 ± 0.050	0.850 ± 0.084	0.832 ± 0.083
C5.0 - Median	0.904 ± 0.055	0.807 ± 0.106	0.889 ± 0.057
C5.0 - missForest	0.903 ± 0.054	0.803 ± 0.099	0.890 ± 0.059
LR	0.901 ± 0.033	0.848 ± 0.054	0.884 ± 0.050
C5.0 - NA	0.898 ± 0.056	0.822 ± 0.089	0.854 ± 0.098
C5.0 - 0	0.887 ± 0.050	0.815 ± 0.075	0.830 ± 0.109
LR - Median	0.880 ± 0.044	0.815 ± 0.068	0.875 ± 0.054
LR - missForest	0.879 ± 0.044	0.814 ± 0.066	0.878 ± 0.056
LR - 0	0.871 ± 0.037	0.807 ± 0.057	0.873 ± 0.057

As shown in Table 8.10, SVM with complete data had the highest mean AUC across all lead times (AUC = 0.946), compared to the other models. BARTm with 0 imputation to testing data had the highest mean sensitivity (Sens = 0.901) compared to the other models. AdaBoost with 0 imputation to test set had the highest mean NPV of 0.570.

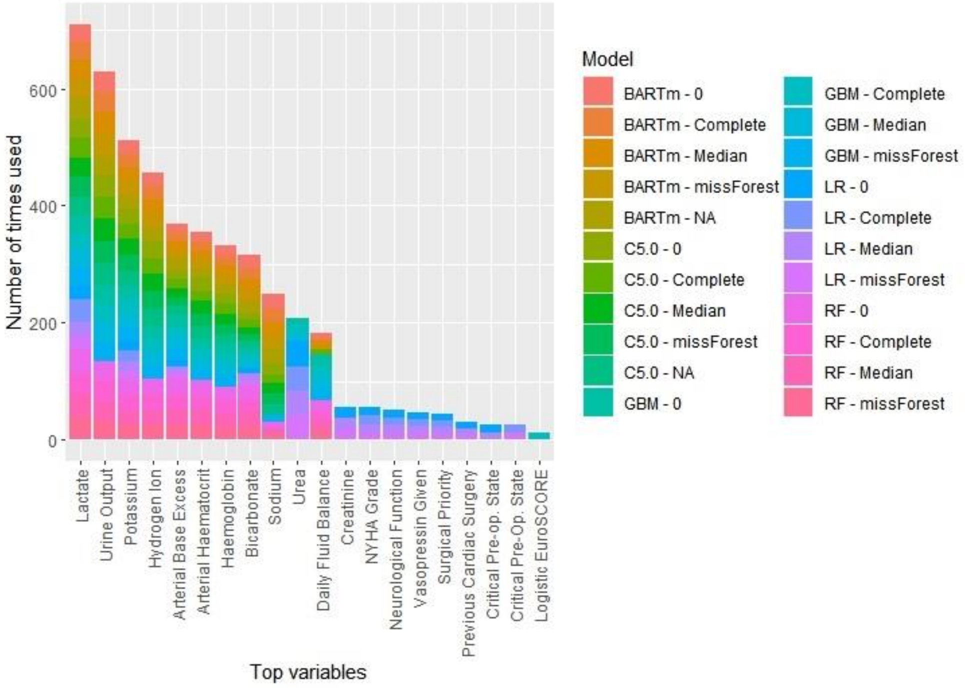
Overall, none of the models have particularly high negative predictive value at all lead times. Even though these values are important to determine the certainty of the models, as the prediction models developed in this thesis are not clinical interventions, the AUC, sensitivity and specificity have a higher importance.

As described in Section 8.2, the only existing dynamic model predicting delirium in the ICU was found to be the model by Oh et al. [388], which used heart rate variability to predict delirium and its stages. Oh et al.'s best model reached a sensitivity of 0.880 and specificity of 0.278 [388]. This is considerably lower than the support vector machine model developed in this study with mean sensitivity of 0.907 and mean specificity of 0.870. This means that the SVM model recognises patients with delirium 90.7% of the time (vs Oh et al.'s 88.0%), and patients without delirium 87.0% of the time (vs Oh et al.'s 27.8%).

Models developed with complete data performed slightly better than models with missing values or missForest imputation. However, the BARTm model with missing values in test set was the 9th best performing model based on AUC (0.930 vs SVM's 0.941) and 2nd best model based on sensitivity (0.885 vs SVM's 0.907). This is a very promising result as missing data are a vast problem in healthcare databases [110]. Ideally, more effort should be directed towards developing higher quality databases, however, in the meantime, being able to use methods that handle missing data, is a great solution. Hence, if a patient has some missing data, the clinician can still be informed whether a patient is likely to develop delirium due to the well-performing prediction model that includes missing values.

Calibration of the top-performing models in each experiment were quite similar, where the models were more certain about predicting patients without delirium, as opposed to predicting patients with delirium. That being said, based on the mean predicted probability, models in Experiment 2 and 3 tended to underestimate the risk of delirium, whereas the top-performing model in Experiment 1 tended to slightly overestimate the risk of delirium.

Figure 8.27. Most important variables for all models that are used for at least 10 times.



As shown in Figure 8.27, the most important variables in the models overall were lactate, urine output, potassium and hydrogen ion. Lactate, potassium and hydrogen ion have been confirmed to be associated with postoperative delirium elsewhere [92], [391], [392]. There is less confirmation whether urine output is associated with delirium in other studies. There is some evidence of the relationship between acute kidney injury (AKI), often defined by urine output, and delirium in the ICU, which might explain why urine output was the second-most important variable chosen by the models [393].

8.7. Conclusion

Most of the models currently predicting delirium in the ICU are static models. This chapter demonstrated that it is possible to predict delirium on an hourly basis 13 hours in advance, with the mean AUC of 0.941 (SVM), using complete data and mean AUC of 0.930 (BARTm) with missing data. The models developed in this study could help clinicians optimise treatments for patients who are at risk of developing delirium hours in advance.

Chapter 9. Overall Discussion and Conclusion

This thesis aimed to develop models predicting postoperative complications, using preoperative data and laboratory results in an intensive care unit. In order to develop these predictive models, the following research questions were answered:

> **RQ1:** What is the current landscape of dynamic prediction models in critical care in terms of prediction modelling methods?
>
> **RQ2:** What are cardiac surgery experts' challenges in cardiac surgery and priorities for a new prediction model predicting patient outcomes?
>
> **RQ3:** How can postoperative complications be classified using routinely collected medical data?

In this chapter, the research questions are answered and discussed based on the findings in this thesis. In addition, the strengths and limitations of the research presented in this thesis are discussed, and the clinical relevance of the developed prediction models are explained.

9.1. Response to Research Questions

9.1.1. What is the Current Landscape of Dynamic Prediction Models in Critical Care in Terms of Prediction Modelling Methods? (RQ1)

While there are numerous prediction models developed to predict patient outcomes in an intensive care unit, the findings from the literature review (Chapter 2) indicate that the majority of these models have been developed to predict mortality. Together with the findings from Chapter 3 and 4, this emphasises the lack of focus on postoperative complications in the literature and the need for prediction models for severe complications, such as acute kidney injury and delirium, as was done in Chapters 6, 7 and 8.

In terms of methods of how dynamic ICU prediction models have been developed, there is a lack of clarity in terms of how missing data are handled or how imbalanced classification problems are addressed.

Finally, while the use of publicly available datasets helps with accessibility of healthcare data to improve digital innovation, but also with reproducibility of studies, the majority of studies using the MIMIC databases developed and/or tested their models on datasets that include surgeries between 2001 and 2012 (MIMIC-III). This can be a problem due to the changing patient population, surgical procedures, and policies, meaning that these prediction models might not be applicable in current practice.

9.1.2. What are cardiac surgery experts' challenges in cardiac surgery and priorities for a new prediction model predicting patient outcomes? (RQ2)

This research question aimed to involve potential users of the prediction models developed in this thesis in the development process. Exploratory interviews with cardiac anaesthetists and cardiac surgeons were carried out (Chapter 3) to understand the current challenges in cardiac surgery, how clinicians in cardiac surgery currently use perioperative risk prediction models and what their current priorities are for a prediction model. It was found that the main challenges in cardiac surgery are adverse outcomes, changing population in terms of age and pre-existing conditions, and changing procedures, all of which are connected to one another. It was identified that common serious complications that occur in cardiac patients are bleeding, infections, pulmonary complications, renal complications, stroke and delirium. It was also found that adverse outcomes are mostly attempted to be avoided at the preoperative phase, where decisions about suitable treatments are made. Perioperative risk prediction models were known to the interviewed clinicians, however, surprisingly, these were not used for decision making, but rather for documentation and audit purposes. However, following from the explorative interviews, it was evident that there is a need and a place for preoperative and intensive care unit risk prediction models that could help with decision making regarding the risk of postoperative complications.

Secondly, a Delphi study (Chapter 4) was conducted to understand whether cardiac clinicians would find defining and classifying postoperative complications in cardiac surgery useful, and subsequently, how they would define and classify postoperative complications specifically following cardiac surgery. This study was undertaken due to the lack of a currently available complication classification system in cardiac surgery, which could enable more structured reporting of complications, and increased transparency of developed prediction models for combined complications. In this study consensus was reached that defining and classifying complications in cardiac surgery is useful. Consensus was also reached on the characteristics of postoperative complications in cardiac surgery, which resulted in the following definition:

> *A complication following cardiac surgery is an unplanned adverse event that occurs following cardiac surgery that can cause delay in recovery, delay in hospital discharge and affect patient's quality of life and is likely to happen due to surgical process.*

The experts agreed that the categories for complications following cardiac surgery should be "Mild", "Moderate", "Severe" and "Death". Concerning postoperative complications, consensus for characteristics of "Mild" and "Severe" complications was reached.

9.1.3. How can postoperative complications be classified using routinely collected medical data? (RQ3)

Chapter 6 showed the results of models predicting "Severe" postoperative complications, acute kidney injury and delirium, using preoperatively available data. Overall, the models produced moderate results, the highest performance belonging to random forest predicting "Severe" complications, using 24 variables (AUC = 0.713, sensitivity = 0.562, specificity = 0.748, PPV = 0.039, NPV = 0.866). When predicting delirium, logistic regression using 24 variables had the highest performance (AUC = 0.675, sensitivity = 0.672, specificity = 0.611, PPV = 0.068, NPV = 0.811). Finally, when predicting acute kidney injury, stacked model with generalised linear model, using 24 variables and upsampled training data had the best performance (AUC = 0.667, sensitivity = 0.592, specificity = 0.657, PPV = 0.131, NPV = 0.706).

It was clear that the models developed to predict “Severe” postoperative complications had better performance measures than when predicting delirium or acute kidney injury. In addition, the models’ performances benefitted from using a larger number of available variables. Also, in general, upsampling of the training data did not result in higher performance of the models.

This chapter showed that in order to predict postoperative outcomes, the models could benefit from more granular data, such as laboratory values, which were explored in Chapters 7 and 8.

When predicting acute kidney injury, all models had similarly moderately high performance. Based on AUC, BARTm had the highest mean performance across all lead times from 24 hours to 1 hour in advance (AUC = 0.850, sensitivity = 0.821, specificity = 0.741, PPV = 0.021, NPV = 0.775). Interestingly, the BARTm model that was evaluated on testing data that contained missing values achieved also a notably high mean performance (AUC = 0.830, sensitivity = 0.780, specificity = 0.741, PPV = 0.023, NPV = 0.800). As missing data are a problem for developing usable predictive models in healthcare [110], it is promising that a model, such as BARTm is robust enough to make a prediction, even when missing data are included.

When predicting delirium, all models had very high overall performance, mostly with AUC staying above 0.900. Support vector machine had the highest mean performance across lead times of 13 hours to 1 hour in advance (AUC = 0.941, sensitivity = 0.907, specificity = 0.870, PPV = 0.015, NPV = 0.488). Interestingly, the gradient boosting machine model that was evaluated on testing set which had missing values replaced with 0 had the 4th highest mean performance (AUC = 0.931, sensitivity = 0.858, specificity = 0.880, PPV = 0.022, NPV = 0.481). This is useful since it indicates that the model could also be used if a patient has missing values, where the values are simply replaced with 0.

9.2. Strengths and Limitations

9.2.1. Study Data

As seen from Chapter 2, many studies that developed dynamic prediction models for patient outcomes in ICU used MIMIC databases. Chapter 3 identified that the cardiac population and procedures have changed within the last decade, which is also confirmed by the literature [138], [155], [157]. This means the earlier versions of the MIMIC databases, such as MIMIC-II and MIMIC-III are no longer applicable for current patient population, making the currently available ICU prediction models that were developed and validated using these datasets, potentially out of date.

Using cardiac patients' data that is relevant to Scottish cardiac population can be a strength of the experiments shown in this thesis. The specific data from Golden Jubilee National Hospital was used to develop prediction models that are tailored to this institution's needs and are up to date for to the hospital's cardiac patient population.

There is an argument to be made about whether numerous prediction models for each different population are needed. On one hand, having widely used international prediction models could help with general quality improvement in cardiac centres internationally, if such a model is used for auditing. On the other hand, prediction models that are developed, using more cardiac centres' data, have to deal with too much variability that it loses its ability to make an accurate individual prediction. In addition, the nuances of a certain cardiac centre can be lost. One example of widely used preoperative risk model is logistic EuroSCORE. Since it is widely validated, it is used in many cardiac centres internationally [191]. However, EuroSCORE is criticised widely for its overestimation of risk of mortality and hence being no longer applicable for today's cardiac population [192].

This highlights the importance of updating and re-calibration of risk prediction models to capture the cardiac centre's patient population on which these models are intended to be used [321].

A limitation to the experiments undertaken in this thesis is the involvement of preoperative and postoperative laboratory data only. While both datasets included valuable information to develop predictive models for complications, inclusion of vital signs recorded in the ICU could have benefitted the predictive ability of the models. Furthermore, an inclusion of intra-operative data, such as time of surgery and anaesthetic data could have improved the performance for the hourly prediction models.

9.2.2. Predicted Outcomes

Another strength of this thesis is the objectivity of the predicted outcomes. Firstly, when looking at other studies predicting combined postoperative complications using preoperative data, there is no transparency in why these particular complications were chosen to be predicted [342]–[344], [347], [394]. Hence, the Delphi study was undertaken to take the first steps to define and classify postoperative complications in cardiac surgery (Chapter 4). Based on the characteristics found in that study, the "severe" postoperative complications were chosen as the outcome of the prediction models in Chapter 6. Knowing the exact characteristics that a "severe" complication has enables comparison between studies and makes the development of these prediction models more transparent.

A limitation to using the complication characteristics from Chapter 4 is that this was a study, including cardiac anaesthetists and intensivists only, and no cardiac surgeons took part. In addition, the experts were mainly based in the UK. While using the results of the study is a more transparent and structured approach than not using any agreed-upon definition about postoperative complications, there is also a possibility that the results might not transfer to the individual patient due to complications can have a varying severity from patient to patient. Hence, further investigation and validation of this classification needs to be undertaken.

Secondly, this study has two great advantages in terms of objectivity when predicting AKI: AKI is defined, using a widely recognised criterium (KDIGO), and with each AKI diagnosis, a timestamp is recorded with it. In the literature review (Chapter 2) only one study predicted renal complications in a dynamic manner in ICU, which was also defined based on the KDIGO criterium [74]. However, this study had various limitations, such as using a balanced dataset for both training and testing data, further explained in Chapter 2.

Different criteria for diagnosing AKI exist. Most well-known examples, other than KDIGO, are Acute Kidney Injury Network (AKIN) and Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease (RIFLE) criteria, however, KDIGO has been shown to have greater sensitivity to detect AKI than RIFLE or AKIN criteria [281], [395]. There are a number of advantages for using the KDIGO guidelines to diagnose AKI retrospectively. As described in Chapter 5, the CaTHI dataset records postoperative complications, which means that this data could have been used for understanding which patients had postoperative complications, and more specifically AKI. However, the CaTHI database does not include the timestamps when the complications occurred, which makes the hourly prediction impossible. In addition, the CaTHI database relies on the discharge information recorded by ICU staff to enter the patient outcomes into the database. Because the discharge information is entered into the ICU system retrospectively, the information can be subjective and incomplete. In addition, clinicians' notes can be difficult to understand when not knowing individual patients and not understanding the whole patient journey.

A limitation to AKI as a predicted outcome is the fact that currently available AKI diagnosis criteria involve using serum creatinine laboratory results. As seen from Chapter 5, the creatinine tests are taken in every median 23.67 hours. This means that there is a limitation to the dynamic hourly prediction, and some cases of AKI can be missed due to low frequency of creatinine tests. However, it is evident that for a small number of patients the creatinine tests are more regular, done in an ad-hoc basis, mostly because of higher previously recorded creatinine level. Hence, as the majority of patients did not have AKI, it is understandable why their creatinine tests were taken less often.

Thirdly, using the CAM-ICU diagnosis as the predicted outcome is currently the most objective way to develop prediction models for delirium due to its accuracy at diagnosing delirium patients in critical care [396]. Due to repeated assessments of delirium, this thesis showed that it is possible to predict delirium on an hourly basis.

However, there are a few limitations to defining the CAM-ICU diagnosis as the predicted outcome.

Because delirium assessment with CAM-ICU is undertaken by clinical staff, the regularity of when the assessment is done can vary from patient to patient. On average, CAM-ICU assessments were undertaken every 10.52 hours. This means that there is a possibility of under reporting of delirium due to the relatively infrequent assessments of delirium. As delirium can develop rapidly [397], the actual occurrence of it can be earlier than when delirium is assessed.

Another limitation, when analysing CAM-ICU diagnoses, is the possibility for observer bias. The score is measured by different clinical staff, which means that there can be variation in training and the results can be subjective [397]. Oh et al. minimise this problem by having the CAM-ICU scoring done both by psychiatrists and nurses [388]. However, as this thesis analyses the data retrospectively, it is unknown who exactly undertook the assessment and what their level of training was.

Even though the CAM-ICU could be recorded more often, being able to clearly state when exactly CAM-ICU was diagnosed for each patient is a strength to this study by offering transparency on how the models were developed. None of the studies discussed in Section 8.2 in Chapter 8 reported at what time delirium occurred in their patient population. Besides, if delirium was predicted in a static manner (Appendix 8.5), the hourly prediction models still achieved better performance in terms of AUC, sensitivity and specificity.

Overall, even though the outcomes predicted in this thesis were defined based on expert consensus ("severe" complications) or on internationally recognised diagnosis criteria (KDIGO, CAM-ICU), limitations to these criteria exist also for other conditions. The limitations to sepsis diagnosis criteria were discussed in Section 2.4.1 in Chapter 2. Since the focus from predicting mortality is moving more towards predicting complications, objective, and reliable definitions of diagnoses for complications are ever-changing [97]. Due to changing patient population, and postoperative complications becoming more prevalent, new definitions are developed frequently [54] and some institutions can be also slow to adopt new guidelines [398].

This further highlights the need for a prediction model to be regularly re-calibrated and updated to avoid it becoming obsolete [321].

9.2.3. Classification Methods Used

A range of methods was used to predict the outcomes, all suitable for numerical and categorical variables. These methods include logistic regression, which is a classic statistical method, tree-based methods (BCART, BARTm, C5.0), ensemble methods (random forest and stacking methods), boosting methods (AdaBoost, GBM) and others, such as naïve Bayes and SVM. Using a wide variety of methods enables the comparison of different approaches and provides an overview on not only which models have the best performance, but also on how usable they are in practice.

As shown in Chapter 6, the models chosen in this thesis are robust enough to handle the imbalanced classification problem, which was a case in this thesis, where the highest prevalence of the predicted outcome was 18.93% (AKI) and lowest prevalence was 5.91% (severe complications). In the literature review (Chapter 2) it was found that many studies used balancing methods, even though numerous classification methods have been shown to be robust when handling imbalanced classes [147]–[149].

The models being able to handle class imbalance is a strength to this thesis as models developed using balanced training data can have poor calibration when evaluated on a real-world imbalanced data due to overestimating the predicted probabilities [135]. This was also evident in Section 6.4.2.2 in Chapter 6 where the best-performing model predicting AKI, developed with upsampled training data had noticeably poor calibration.

In addition, various approaches were taken to handle missing data in the ICU dataset. Models, such as C5.0 and BARTm, that are able to make a prediction even with the presence of missing data, were used. This kind of approach was only taken by one study in the literature review (Chapter 2) [88]. Furthermore, imputation methods were experimented with, including median imputation, missForest imputation, and replacing missing values with 0 in the testing set. Even though imputation methods were shown to be a popular approach for missing data in Chapter 2 by other studies, imputation methods are suggested to be handled with caution due to the danger of introducing some bias to the results [144]. However, the results in Chapters 7 and 8 indicate similar model performance regardless of whether the models were tested on complete data, testing data with missing values, or testing data treated with imputation methods. As missing data in electronic health records are very common, and are a barrier to development of accurate and usable clinical prediction models [110], further efforts should be directed towards producing medical records that are of high quality. However, experimenting and developing classification methods that handle missing values could be a viable option in the meanwhile.While the chosen approach for hourly prediction models was simple, and therefore would be easily understandable for clinicians, there are various other ways these prediction tasks could have been approached. As seen from the literature review (Chapter 2), there are a number of methods that other papers have used to predict patient outcomes in the ICU that this thesis did not explore. For example, in the future a base model could be developed, which could be updated as the new information is added to the database, as was done by Feng et al. [60] and Deasy et al. [58], both using neural networks, and Gultepe et al. [62] and Yee et al. [86], both using Bayesian networks. Also, survival models could be explored in the future, as was done by Henry et al, using Cox proportional hazards [63]. However, the analyses in this thesis did not include vital signs, unlike these above-mentioned studies. Using less frequently measured data means that there was less data available for each patient, and therefore the number of data points available for developing a deep neural network were insufficient. Hence, the methods chosen in this thesis were less complicated, but still with adequate performance. However, exploration of methods like neural networks predicting patient outcomes should be continued, as explained in Future Work (Section 9.4).

9.2.4. Interpretability of the Models

Because the models are all performing quite well, based on the results shown in this thesis, it is difficult to say which model should be the ultimate chosen model to further develop a clinical decision support tool. On one hand, the model with the highest AUC, sensitivity and NPV should be chosen, but on the other hand, the interpretability and usability of the models should be considered.

In medicine, it is especially important to understand why a model predicts a patient to have high probability for, say, delirium. If the clinician knows what factors have contributed to the probability to be high, then the clinicians can investigate which factors to pay closer attention to and improve with medical interventions.

A limitation to most of the models developed in this thesis is the lack of interpretability due to the "black-box" nature of most machine learning methods. In addition to helping to understand how the models make their risk calculations, reporting the model coefficients helps the models to be applied for individual use as it enables calculating the probability for the predicted outcome based on patient-specific laboratory test results. Where logistic regression had the highest performance, the coefficients of the model were presented. However, for other prediction experiments, machine learning methods, such as random forest, support vector machine and BARTm often had the best performance, and for these models, where possible, variable importance was reported due to the unavailability of regression coefficients. This helps the users to understand which variables are having the biggest impact on the predicted outcome, however does not give the specifics of how the predicted probabilities are calculated, which can be a limitation in uptake of such prediction model in practice [368].

Even though the importance of the variables is known and can be obtained for most of the methods used in this thesis, logistic regression is a highly explainable classification method due to its built-in estimates that can be converted into odds ratios. Even though logistic regression's performance for hourly prediction models was rather on the lower side, compared to the other models, it is highly understandable to its intended users, which can be an important aspect for successful implementation [399]. Logistic regression has also been shown to be competitive with other machine learning algorithms in other studies [150], [301], [332].

9.3. Clinical Relevance

The prediction models developed in this thesis have a few potential usages. The preoperative prediction models could be applied to preoperative assessment and decision-making regarding the type of treatment plan the patient is receiving.

If a clinician knows, using a preoperative prediction model for AKI, that a patient is at risk of having postoperative AKI, then the renal recovery period before surgery could be applied (i.e., cardiac surgery is delayed due to using certain medicines, such as contrast agents that can have an adverse effect on kidneys) [400]. In case of a patient being at risk for AKI or delirium, certain medications, such as preoperative antipsychotics or postoperative inotropes, associated with delirium [401] and nephrotoxins, associated with AKI [402], could be avoided.

Furthermore, before surgery, the preoperative prediction models could aid the clinician at communicating risk about potential complications the patient might have. This will help managing patients' and their families' expectations regarding the outcomes of the surgery.

The hourly ICU prediction models could be developed into a clinical system that is integrated with electronic health records. These hourly prediction models could predict AKI and delirium hours in advance, so that clinicians can direct more resources towards patients at risk and make an informed decision about safe treatments for the patient.

Besides helping with clinician's decision-making and risk communication, the prediction models could be used for surgical planning and bed planning, as patients with postoperative complications are known to stay in the hospital for longer [4], [5]. Using such prediction models to manage bed spaces could help reduce cancellations of surgery [403].

9.4. Future Work

The research presented in this thesis leads to various avenues of future work needed for developing accurate and usable prediction models for postoperative complications that would enable their uptake in everyday clinical practice.

Firstly, as the study presented in Chapter 3 was not an in-depth analysis of clinicians' requirements for a clinical prediction model, further research is needed to understand the requirements for such a model, and to understand how it could fit into practical context. Furthermore, once a prototype of a system involving the developed prediction models is developed, a system evaluation study needs to be undertaken to understand the clinicians' perspectives on the performance and the utility of the models.

Secondly, the consensus about the usefulness of defining and classifying surgical complications following cardiac surgery provides a rationale for continuing the work towards developing an internationally recognised definition and classification system for such complications (Chapter 4). Hence, further study is needed that involves international participants, including cardiac surgeons, in addition to cardiac anaesthetists, to develop a definition and a classification system for postoperative complications following cardiac surgery.

Thirdly, to improve the predictive ability of the ICU models, vital signs should be included to provide more granular information that is recorded far more regularly than laboratory results. Furthermore, to improve the prediction of AKI, in addition to collecting serum creatinine more regularly in the ICU, the use of biomarkers should be further investigated and applied to practice. Because acute kidney injury is a complex, multi-factorial complication, there is currently no single biomarker that is a "kidney troponin" [404]. Currently, the most promising biomarkers for acute kidney injury diagnosis are NGAL, IL-18, kidney injury molecule-1 and cell-cycle arrest biomarkers. Using biomarkers as variables in clinical prediction models have shown to improve the accuracy of models [281].

In addition to the methods used to develop predictive models, with the addition of vital signs data, deep learning methods, such as deep neural networks should be experimented with in the future.

In addition to internal validation of the developed prediction models developed in this thesis, to ensure reproducibility, risk prediction models should also undergo external validation to support generalisability before implementation into clinical practice [405]. Following successful external validation, a continuous updating strategy needs to be developed to allow for changes in the cardiac population over time. In addition, to allow for wide-spread uptake of the developed prediction models, a system including these models needs to be developed that is integrated with electronic health records.

Finally, to help mitigate the effects of other adverse postoperative outcomes, other postoperative complications in the ICU, such as stroke, respiratory complications and various infections should be investigated. This, however, is possible once agreed-upon diagnosis criteria based on laboratory results and vital signs are developed, aiding dynamic prediction of such complications.

9.5. Final Conclusions

The results from this thesis show that there is an appetite for improving the recording and prediction of postoperative complications in cardiac surgery. Using routinely collected medical data can be used to develop both preoperative and hourly ICU predictive models for postoperative complications, such as acute kidney injury and delirium. Such prediction models could help with not only clinical decision making, but also communication about risk, research in complications and auditing.

References

[1] SCTS, “National Adult Cardiac Surgery Audit (NACSA),” 2020.

[2] NHS, “Coronary Artery Bypass Graft,” 2020. https://www.nhs.uk/conditions/coronary-artery-bypass-graft-cabg/.

[3] NHS, “Aortic Valve Replacement,” 2021. https://www.nhs.uk/conditions/aortic-valve-replacement/#:~:text=An aortic valve replacement involves,long time to recover from.

[4] N. Al-Sarraf *et al.*, “The effect of preoperative renal dysfunction with or without dialysis on early postoperative outcome following cardiac surgery,” *Int. J. Surg.*, vol. 9, no. 2, pp. 183–187, 2011, doi: 10.1016/j.ijsu.2010.11.006.

[5] M. Schubert *et al.*, “A hospital-wide evaluation of delirium prevalence and outcomes in acute care patients - a cohort study,” *BMC Health Serv. Res.*, vol. 18, no. 1, 2018.

[6] SCTS, “National Cardiac Surgery Activity and Outcomes Report 2002-2016,” 2020. doi: 10.1201/9781315164533.

[7] L. Ball, F. Constantino, and P. Pelosi, “Postoperative complications of patients undergoing cardiac surgery,” *Curr. Opin. Crit. Care*, vol. 22, no. 4, pp. 386–392, 2016.

[8] J. Maillard, N. Elia, C. S. Haller, C. Delhumeau, and B. Walder, “Preoperative and early postoperative quality of life after major surgery - A prospective observational study,” *Health Qual. Life Outcomes*, vol. 13, no. 1, pp. 1–12, 2015, doi: 10.1186/s12955-014-0194-0.

[9] A. Pinto, O. Faiz, R. Davis, A. Almoudaris, and C. Vincent, “Surgical complications and their impact on patients’ psychosocial well-being: A systematic review and meta-analysis,” *BMJ Open*, vol. 6, no. 2, 2016, doi: 10.1136/bmjopen-2014-007224.

[10] S. Eappen *et al.*, “Relationship Between Occurrence of Surgical Complications and Hospital Finances,” *JAMA*, vol. 309, no. 15, pp. 1599–1606, 2013, [Online].

Available: www.jama.com.

[11] M.-M. Bouamrane and F. S. Mair, “Implementation of an integrated preoperative care pathway and regional electronic clinical portal for preoperative assessment,” *BMC Med. Inform. Decis. Mak.*, vol. 14, no. 93, 2014.

[12] Centre for Perioperative Care, “Multidisciplinary working in perioperative care,” 2020. [Online]. Available: https://cpoc.org.uk/sites/cpoc/files/documents/2020-09/Multidisciplinary working in perioperative care - rapid review.pdf.

[13] Centre for Perioperative Care, “Shared Decision Making,” 2022. https://www.cpoc.org.uk/shared-decision-making.

[14] S. R. Moonesinghe, M. G. Mythen, P. Das, K. M. Rowan, and M. P. Grocott, “Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review,” *Anesthesiology*, vol. 119, no. 4, pp. 959–981, 2013.

[15] National Guideline Centre, “Perioperative care in adults: Evidence review for preoperative risk stratification tools,” 2020. [Online]. Available: https://www.nice.org.uk/guidance/ng180/evidence/c-preoperative-risk-stratification-tools-pdf-8833151056.

[16] S. Barnett and S. R. Moonesinghe, “Clinical risk scores to guide perioperative management,” *Postgrad. Med. J.*, vol. 87, no. 1030, pp. 535–541, 2011, doi: 10.1136/pgmj.2010.107169.

[17] L. Lapp, “Developing predictive models for postoperative complications in cardiac patients (MPhil Thesis),” University of Strathclyde, 2017.

[18] C. Xiao, E. Choi, and J. Sun, “Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review,” *J. Am. Med. Informatics Assoc.*, vol. 25, no. 10, pp. 1419–1428, 2018, doi: 10.1093/jamia/ocy068.

[19] N. Mehta and A. Pandit, “Concurrence of big data analytics and healthcare: A

systematic review," *Int. J. Med. Inform.*, vol. 114, pp. 57–65, 2018.

[20] M. G. Seneviratne, N. H. Shah, and L. Chu, "Bridging the implementation gap of machine learning in healthcare," *BMJ Innov*, vol. 6, pp. 45–47, 2019.

[21] A. Khanijahani, S. Iezadi, S. Dudley, M. Goettler, P. Kroetsch, and J. Wise, "Organizational, professional, and patient characteristics associated with artificial intelligence adoption in healthcare: A systematic review," *Heal. Policy Technol.*, vol. 11, no. 100602, 2022.

[22] M. Al-rawashdeh, P. Keikhosrokiani, B. Belaton, M. Alawida, and A. Zwiri, "IoT Adoption and Application for Smart Healthcare: A Systematic Review," *Sensors*, vol. 22, no. 5377, 2022.

[23] K. Cresswell, M. Callaghan, S. Khan, Z. Sheikh, H. Mozaffar, and A. Sheikh, "Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: A systematic review," *Health Informatics J.*, vol. 26, no. 3, pp. 2138–2147, 2020.

[24] M. Nagendran *et al.*, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, no. m689, 2020.

[25] M. van Smeden *et al.*, "Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease," *Eur. Heart J.*, vol. 43, no. 31, pp. 2921–2930, 2022, doi: https://doi.org/10.1093/eurheartj/ehac238.

[26] A. Akay and H. Hess, "Deep Learning: Current and Emerging Applications in Medicine and Technology," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 3, pp. 906–920, 2019.

[27] J. McCarthy, M. L. Minsky, N. Rochester, and C. W. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence AUgust 31, 1955," *AI Mag.*, vol. 27, no. 4, pp. 12–14, 2006.

[28] J. Mullenbach, S. Wiegreffe, J. DUke, J. Sun, and J. Einstein, "Explainable prediction of medical codes from clinical text," *arXiv*, 2018, doi: https://doi.org/10.48550/arXiv.1802.05695.

[29] R. Thirunavukarasu, G. P. Doss, R. Gnanasambandan, M. Gopikrishnan, and V. Palanisamy, "Towards computational solutions for precision medicine based big data healthcare system using deep learning models: a review," *Comput. Biol. Med.*, vol. 149, no. 106020, 2022.

[30] L. Balkenende, J. Teuwen, and R. M. Mann, "Application of Deep Learning in Breast Cancer Imaging," *Semin. Nucl. Med.*, vol. 52, no. 5, pp. 584–596, 2022.

[31] R. Aggarwal *et al.*, "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis," *npj Digit. Med.*, vol. 4, no. 65, 2021.

[32] R. R. Wildeboer, R. J. G. van Sloun, H. Wijkstra, and M. Mischi, "Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods," *Comput. Methods Programs Biomed.*, vol. 189, no. 105316, 2020.

[33] S. M. Ayyad *et al.*, "Role of AI and Histopathological Images in Detecting Prostate Cancer: A Survey," *Sensors*, vol. 21, no. 8, p. 2586, 2021.

[34] V. Romeo *et al.*, "AI-enhanced simultaneous multiparametric F-FDG PET/MRI for accurate breast cancer diagnosis," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 49, pp. 596–608, 2022.

[35] Y. Kumar, S. Gupta, R. Singla, and Y.-C. Hu, "A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis," *Arch. Comput. Methods Eng.*, vol. 29, pp. 2043–2070, 2022.

[36] K. Freeman *et al.*, "Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy," *BMJ*, vol. 374, p. n1872, 2021.

[37] S. Bedrikovetski *et al.*, "Artificial intelligence for pre-operative lymph node staging in colorectal cancer: a systematic review and meta-analysis," *BMC Cancer*, vol. 21, no. 1058, 2021.

[38] N. Sushentsev *et al.*, "Comparative performance of fully-automated and semi-automated artificial intelligence methods for the detection of clinically significant prostate cancer on MRI: a systematic review," *Insights Imaging*, vol.

13, no. 59, 2022.

[39] I. Matias *et al.*, “Prediction of Atrial Fibrillation using artificial intelligence on Electrocardiograms: A systematic review,” *Comput. Sci. Rev.*, vol. 39, p. 100334, 2021.

[40] S. Saravanan *et al.*, “A systematic review of artificial intelligence (AI) based approaches for the diagnosis of Parkinson’s disease,” *Arch. Comput. Methods Eng.*, vol. 29, pp. 3639–3653, 2022.

[41] A. E. W. Johnson, L. Bulgarelli, and T. J. Pollard, “Deidentification of free-text medical records using pre-trained bidirectional transformers,” *CHIL ’20 Proc. ACM Conf. Heal. Inference, Learn.*, pp. 214–221, 2020.

[42] K. Stone, R. Zwiggelaar, P. Jones, and N. M. Parthalain, “A systematic review of the prediction of hospital length of stay: Towards a unified framework,” *PlOS Digit. Heal.*, vol. 1, no. 4, 2022.

[43] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci. Data*, vol. 3, no. 160035, 2016.

[44] M. Syed *et al.*, “Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: systematic review,” *Informatics*, vol. 8, no. 15, 2021.

[45] T. Panch, H. Mattie, and L. A. Celi, “The ‘inconvenient truth’ about AI in healthcare,” *NPJ Digit. Med.*, vol. 2, no. 77, 2019.

[46] A. Rajkomar, J. Dean, and I. Kohane, “Machine Learning in Medicine,” *N. Engl. J. Med.*, vol. 380, pp. 1347–1358, 2019.

[47] D. Mokart *et al.*, “Delayed intensive care unit admission is associated with increased mortality in patients with cancer with acute respiratory failure,” *Leuk. Lymphoma*, vol. 54, no. 8, pp. 1724–1729, 2013, doi: 10.3109/10428194.2012.753446.

[48] L. Mardini, J. Lipes, and D. Jayaraman, “Adverse outcomes associated with delayed intensive care consultation in medical and surgical inpatients,” *J. Crit.*

Care, vol. 27, no. 6, pp. 688–693, 2012, doi: 10.1016/j.jcrc.2012.04.011.

[49] V. Huddar, B. K. Desiraju, V. Rajan, S. Bhattacharya, S. Roy, and C. K. Reddy, "Predicting Complications in Critical Care Using Heterogeneous Clinical Data," *IEEE Access*, vol. 4, pp. 7988–8001, 2016, doi: 10.1109/ACCESS.2016.2618775.

[50] QSR International, "NVivo." 2021, [Online]. Available: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home.

[51] Mendeley Ltd, "Mendeley Reference Manager." 2021, [Online]. Available: https://www.mendeley.com/reference-management/reference-manager.

[52] T. Wigmore and P. Farquhar-Smith, "Outcomes for critically ill cancer patients in the ICU: Current trends and prediction," *Int. Anesthesiol. Clin.*, vol. 54, no. 4, pp. e62–e75, 2016, doi: 0.1097/AIA.0000000000000121.

[53] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal," *BMJ*, vol. 369, 2020, doi: https://doi.org/10.1136/bmj.m1328.

[54] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016, doi: 10.1001/jama.2016.0287.

[55] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, no. n71, 2021.

[56] S. Bhattacharya, V. Huddar, V. Rajan, and C. . Reddy, "A dual boundary classifier for predicting acute hypotensive episodes in critical care," *PLoS One*, vol. 13, no. 2, 2018, doi: 10.1371/journal.pone.0193259.

[57] K. Caballero and R. Akella, "Dynamically modeling Patient's health state from electronic medical records: A time series approach," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 2015-Augus, pp. 69–78, 2015, doi: 10.1145/2783258.2783289.

[58] J. Deasy, P. Lio, and A. Ercole, “Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation,” *Nature,* vol. 10, no. 22129, 2020, doi: https://doi.org/10.1038/s41598-020-79142-z.

[59] B. Dummitt, A. Zeringue, A. Palagiri, C. Veremakis, B. Burch, and B. Yount, “Using survival analysis to predict septic shock onset in ICU patients,” *J. Crit. Care*, vol. 48, pp. 339–344, 2018, doi: 10.1016/j.jcrc.2018.08.041.

[60] X. Feng *et al.*, “Dynamic prediction of late noninvasive ventilation failure in intensive care unit using a time adaptive machine model,” *Comput. Methods Programs Biomed.*, vol. 208, no. 106290, 2021, doi: https://doi.org/10.1016/j.cmpb.2021.106290.

[61] S. Ghosh, J. Li, L. Cao, and K. Ramamohanarao, “Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns,” *J. Biomed. Inform.*, vol. 66, pp. 19–31, 2017, doi: 10.1016/j.jbi.2016.12.010.

[62] E. Gultepe, J. P. Green, H. Nguyen, J. Adams, T. Albertson, and I. Tagkopoulos, “From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system,” *J. Am. Med. Informatics Assoc.*, vol. 21, no. 2, pp. 315–325, 2014, doi: 10.1136/amiajnl-2013-001815.

[63] K. . Henry, D. . Hager, P. . Pronovost, and S. Saria, “A targeted real-time early warning score (TREWScore) for septic shock,” *Sci. Transl. Med.*, vol. 7, no. 299, p. 299ra122, 2015, doi: 10.1126/scitranslmed.aab3719.

[64] L. Hernandez *et al.*, “Multimodal tensor-based method for integrative and continuous patient monitoring during postoperative cardiac care,” *Artif. Intell. Med.*, vol. 113, no. 102032, 2021, doi: https://doi.org/10.1016/j.artmed.2021.102032.

[65] V. Huddar, B. K. Desiraju, V. Rajan, S. Bhattacharya, S. Roy, and C. K. Reddy, “Predicting Complications in Critical Care Using Heterogeneous Clinical Data,” *IEEE Access*, vol. 4, pp. 7988–8001, 2016, doi: 10.1109/ACCESS.2016.2618775.

[66] C. W. Hug and P. Szolovits, "ICU acuity: real-time models versus daily models.," *AMIA Annu. Symp. Proc.*, vol. 2009, pp. 260–264, 2009.

[67] A. E. W. Johnson and R. G. Mark, "Real-time mortality prediction in the Intensive Care Unit," *AMIA Annu. Symp. Proc.*, pp. 994–1003, 2017.

[68] R. Joshi and P. Szolovits, "Prognostic physiology: modeling patient severity in Intensive Care Units using radial domain folding," in *AMIA Annual Symposium Proceedings*, 2012, pp. 1276–83.

[69] J. Lee and R. G. Mark, "An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care," *Biomed. Eng. Online*, vol. 9, pp. 1–17, 2010, doi: 10.1186/1475-925X-9-62.

[70] L.-W. H. Lehman, S. Nemati, R. P. Adams, G. Moody, A. Malhotra, and R. G. Mark, "Tracking progression of patient state of health in critical care using inferred shared dynamics in physiological time series," in *IEEE Engineering in Medicine and Biology Society*, 2013, pp. 7072–7075, doi: 10.1109/EMBC.2013.6611187.

[71] L.-W. H. Lehman *et al.*, "A Physiological Time Series Dynamics-Based Approach to Patient Monitoring and Outcome Prediction," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 3, pp. 1068–1076, 2015, doi: 10.1109/JBHI.2014.2330827.A.

[72] J. Ma, D. K. K. Lee, M. E. Perkins, M. A. Pisani, and E. Pinker, "Using the shapes of clinical data trajectories to predict mortality in ICUs," *Crit. Care Explor.*, vol. 1, no. e0010, 2019, doi: 10.1097/CCE.0000000000000010.

[73] Y. Mao, W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, "An integrated data mining approach to real-time clinical monitoring and deterioration warning," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, no. November 2014, pp. 1140–1148, 2012, doi: 10.1145/2339530.2339709.

[74] A. Meyer *et al.*, "Machine learning for real-time prediction of complications in critical care: a retrospective study," *Lancet Respir. Med.*, vol. 6, no. 12, pp. 905–914, 2018, doi: 10.1016/S2213-2600(18)30300-X.

[75] D. Misra *et al.*, “Early detection of septic shock onset using interpretable machine learners,” *J. Clin. Med.*, vol. 10, no. 301, 2021, doi: 10.3390/jcm10020301.

[76] A. Mohammed, F. van Wyk, L. Chinthala, and A. Khojandi, “Temporal differential expression of physiomarkers predicts sepsis in critically ill adults,” *Shock*, vol. 56, no. 1, pp. 58–64, 2020, doi: 10.1097/shk.0000000000001670.

[77] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, “An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU,” *Crit. Care Med.*, vol. 46, no. 4, pp. 547–553, 2018, doi: 10.1097/CCM.0000000000002936.

[78] H. J. Park, D. Y. Jung, W. Ji, and C.-M. Choi, “Detection of bacteremia in surgical in-patients using recurrent neural network based on time series records: development and validation study,” *J. Med. Internet Res.*, vol. 22, no. 8, p. e19512, 2020, doi: 10.2196/19512.

[79] T. N. Pattalung, T. Ingviya, and S. Chaichulee, “Feature explanations in recurrent neural networks for predicting risk of mortality in intensive care patients,” *J. Pers. Med.*, vol. 11, no. 934, 2021, doi: https://doi.org/10.3390/jpm11090934.

[80] R. Raj *et al.*, “Machine learning-based dynamic mortality prediction after traumatic brain injury,” *Sci. Rep.*, vol. 9, no. 1, p. 17672, 2019, doi: https://doi.org/10.1038/s41598-019-53889-6.

[81] S. P. Shashikumar *et al.*, “Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics,” *J. Electrocardiol.*, vol. 50, no. 6, pp. 739–743, 2017, doi: 10.1016/j.jelectrocard.2017.08.013.

[82] A. Silva, P. Cortez, M. . Santos, L. Gomes, and J. Neves, “Mortality assessment in intensive care units via adverse events using artificial neural networks,” *Artif. Intell. Med.*, vol. 36, no. 3, pp. 223–34, 2006, doi: 10.1016/j.artmed.2005.07.006.

[83] P. J. Thoral *et al.*, “Explainable machine learning on AmsterdamUMCdb for

ICU discharge decision support: uniting intensivists and data scientists," *Crit. Care Explor.*, vol. 3, no. 9, 2021, doi: 10.1097/CCE.0000000000000529.

[84] F. van Wyk, A. Khojandi, A. Mohammed, E. Begoli, R. L. Davis, and R. Kamaleswaran, "A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier," *Int. J. Med. Inform.*, vol. 122, pp. 55–62, 2019, doi: 10.1016/j.ijmedinf.2018.12.002.

[85] J. Xia *et al.*, "A long short-term memory ensemble approach for improving the outcome prediction in intensive care unit," *Comput. Math. Methods Med.*, vol. 2019, 2019, doi: https://doi.org/10.1155/2019/8152713.

[86] C. R. Yee, N. R. Narain, V. R. Akmaev, and V. Vemulapalli, "A Data-Driven Approach to Predicting Septic Shock in the Intensive Care Unit," *Biomed. Inform. Insights*, vol. 4, no. 11, 2019, doi: 10.1177/1178222619885147.

[87] L. Yijing *et al.*, "Prediction of cardiac arrest in critically ill patients based on bedside vital signs monitoring," *Comput. Methods Programs Biomed.*, vol. 214, no. 106568, 2022, doi: https://doi.org/10.1016/j.cmpb.2021.106568.

[88] Q.-Y. Zhao *et al.*, "A machine-learning approach for dynamic prediction of sepsis-induced coagulopathy in critically ill patients with sepsis," *Front. Med.*, vol. 7, no. 637434, 2021, doi: 10.3389/fmed.2020.637434.

[89] E. Moniz-Cook *et al.*, "Challenge Demcare: management of challenging behaviour in dementia at home and in care homes – development, evaluation and implementation of an online individualised intervention for care homes; and a cohort study of specialist community mental health car," *Program. Grants Appl. Res.*, vol. 5, no. 15, pp. 1–290, 2017, doi: 10.3310/pgfar05150.

[90] A. Johnson, T. Pollart, and R. Mark, "MIMIC-III Clinical Database (version 1.4)," *PhysioNet*, 2016. https://doi.org/10.13026/C2XW26.

[91] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," *Sci. Data*, vol. 5, no. 180178, 2018, doi: https://doi.org/10.1038/sdata.2018.178.

[92] C. Lee *et al.*, "The association of perioperative serum lactate levels with postoperative delirium in elderly trauma patients," *Biomed Res. Int.*, 2019, doi: 10.1155/2019/3963780.

[93] M. Shankar-Hari, D. A. Harrison, and K. M. Rowan, "Differences in Impact of Definitional Elements on Mortality Precludes International Comparisons of Sepsis Epidemiology - A Cohort Study Illustrating the Need for Standardized Reporting," *Crit. Care Med.*, vol. 44, no. 12, pp. 2223–2230, 2016, doi: 10.1097/CCM.0000000000001876.

[94] M. M. Levy *et al.*, "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference," *Intensive Care Med.*, vol. 29, no. 4, pp. 530–538, 2003, doi: 10.1097/01.CCM.0000050454.01978.3B.

[95] M. G. Davies and P. O. Hagen, "Systemic inflammatory response syndrome," *Br. J. Surg.*, vol. 84, no. 7, pp. 920–935, 1997, doi: 10.1002/bjs.1800840707.

[96] J. Vermassen, J. Decruyenaere, L. de Bus, P. Depuydt, and K. Colpaert, "Characteristics of Sepsis-2 septic shock patients failing to satisfy the Sepsis-3 septic shock definition: an analysis of real-time collected data," *Ann. Intensive Care*, vol. 154, 2021, doi: https://doi.org/10.1186/s13613-021-00942-1.

[97] M. Sartelli *et al.*, "Raising concerns about the Sepsis-3 definitions," *World J. Emerg. Surg.*, vol. 13, no. 6, 2018, doi: 10.1186/s13017-018-0165-6.

[98] R. Serafim, J. A. Gomes, J. Salluh, and P. Povoa, "A Comparison of the Quick-SOFA and Systemic Inflammatory Response Syndrome Criteria for the Diagnosis of Sepsis and Prediction of Mortality: A Systematic Review and Meta-Analysis," *Chest*, vol. 153, no. 3, pp. 646–655, 2018, doi: 10.1016/j.chest.2017.12.015.

[99] H.-F. Deng *et al.*, "Evaluating machine learning models for sepsis prediction: A systematic review of methodologies," *iScience*, vol. 25, no. 1, p. 103651, 2022, doi: https://doi.org/10.1016/j.isci.2021.103651.

[100] E. Rivers *et al.*, "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *N. Engl. J. Med.*, vol. 345, no. 19, pp. 1368–1377, 2001, doi:

10.1056/NEJMoa010307.

[101] F. Sebat *et al.*, “Effect of a rapid response system for patients in shock on time of treatment and mortality during 5 years,” *Crit. Care Med.*, vol. 35, no. 11, pp. 2568–2575, 2007, doi: 10.1097/01.CCM.0000287593.54658.89.

[102] E. Antman *et al.*, “ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction,” *Circulation*, vol. 110, pp. e82e-292, 2004, doi: 10.1016/j.jacc.2004.07.002.

[103] K. Roedl *et al.*, “Epidemiology of intensive care unit cardiac arrest: Characteristics, comorbidities, and post-cardiac arrest organ failure - A prospective observational study,” *Resuscitation*, vol. 156, pp. 92–98, 2020, doi: 10.1016/j.resuscitation.2020.09.003.

[104] G. W. Smetana, V. A. Lawrence, and J. E. Cornell, “Preoperative pulmonary risk stratification for noncardiothoracic surgery: systematic review for the American College of Physicians,” *Ann. Int. Med.*, vol. 144, no. 8, pp. 581–595, 2006, doi: 10.7326/0003-4819-144-8-200604180-00009.

[105] A. M. Arozullah, J. Daley, W. G. Henderson, and S. F. Khuri, “Multifactorial risk index for predicting postoperative respiratory failure in men after major noncardiac surgery,” *Ann. Surg.*, vol. 232, no. 2, pp. 242–253, 2000, doi: 10.1097/00000658-200008000-00015.

[106] KDIGO, “KDIGO Clinical Practice Guideline for Acute Kidney Injury,” *Off. J. Int. Soc. Nephrol.*, vol. 2, no. 1, pp. 7–14, 2012, doi: 10.1038/kisup.2012.1.

[107] S. M. Dirkes, “Acute Kidney Injury vs Acute Renal Failure,” *Crit. Care Nurse*, vol. 36, no. 6, pp. 75–76, 2016, doi: https://doi.org/10.4037/ccn2016170.

[108] R. Haskell, “Acute Kidney Injury and Chronic Kidney Disease - What’s the Difference?,” *NursingCenter*, 2020. https://www.nursingcenter.com/ncblog/january-2020/acute-kidney-injury-and-chronic-kidney-disease (accessed Jun. 09, 2021).

[109] T. Bove *et al.*, “The incidence and risk of acute renal failure after cardiac surgery,” *J. Cardiothorac. Vasc. Anesth.*, vol. 18, no. 4, pp. 442–445, 2004, doi:

10.1053/j.jvca.2004.05.021.

[110] C. Mazzali and P. Duca, "Use of administrative data in healthcare research," *Intern. Emerg. Med.*, vol. 10, pp. 517–524, 2015, doi: 10.1007/s11739-015-1213-9.

[111] S. I. Goldberg, A. Niemierko, and A. Turchin, "Analysis of data errors in clinical research databases," *AMIA Annu. Symp. Proc.*, pp. 242–246, 2008.

[112] M. Jamshidian and M. Mata, "Advances in Analysis of Mean and Covariance Structure when Data are Incomplete," in *Handbook of Latent Variable and Related Models*, 2007, pp. 21–44.

[113] M. Jamshidian and P. M. Bentler, "ML estimation of mean and covariance structures with missing data using complete data routines," *J. Educ. Behav. Stat.*, vol. 24, no. 1, pp. 21–41, 1999, doi: https://doi.org/10.2307/1165260.

[114] N. Tsikriktsis, "A review of techniques for treating missing data in OM survey research," *J. Oper. Manag.*, vol. 24, no. 1, pp. 53–62, 2005, doi: https://doi.org/10.1016/j.jom.2005.03.001.

[115] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Appl. Artif. Intell.*, vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.

[116] G. F. N. Berkelmans *et al.*, "Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice," *J. Clin. Epidemiol.*, vol. 145, pp. 70–80, 2022, doi: 10.1016/j.jclinepi.2022.01.011.

[117] T. Schneider, "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values," *J. Clim.*, vol. 14, no. 5, pp. 853–571, 2001, doi: https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2.

[118] A. D'Ambrosio, M. Aria, and R. Siciliano, "Accurate Tree-based Missing Data Imputation and Data Fusion within the Statistical Learning Paradigm," *J. Classif.*, vol. 29, pp. 227–258, 2012, doi: https://doi.org/10.1007/s00357-012-

9108-1.

[119] P. Vateekul and K. Sarinnapakorn, “Tree-Based Approach to Missing Data Imputation,” in *2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 70–75, doi: 10.1109/ICDMW.2009.92.

[120] G. Rahman and Z. Islam, “A Decision Tree-based Missing Value Imputation Technique for Data Pre-processing,” in *Proceedings of the 9-th Australasian Data Mining Conference*, 2011, pp. 41–50.

[121] K. G. M. Moons *et al.*, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration,” *Ann. Intern. Med.*, vol. 162, no. 1, pp. W1-73, 2015, doi: 10.7326/M14-0698.

[122] D. J. Stekhoven and P. Bühlmann, “MissForest - non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012, doi: https://doi.org/10.1093/bioinformatics/btr597.

[123] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, “Comparison of random forest and parametric imputation models for imputing missing data using MICE: a caliber study,” *Am. J. Epidemiol.*, vol. 179, no. 6, pp. 764–774, 2014, doi: 10.1093/aje/kwt312.

[124] A. K. Waljee *et al.*, “Comparison of imputation methods for missing laboratory data in medicine,” *BMJ Open*, vol. 3, no. e002847, 2013, doi: http://dx.doi.org/10.1136/bmjopen-2013-002847.

[125] G. Huang, “Missing data filling method based on linear interpolation and lightgbm,” *J. Phys. Conf. Ser.*, vol. 1754, no. 012187, 2021.

[126] J. S. Haukoos and C. D. Newgard, “Advanced Statistics: Missing Data in Clinical Research - Part 1: An Introduction and Conceptual Framework,” *Soc. Acad. Emerg. Med.*, vol. 14, no. 7, pp. 662–668, 2007, doi: 10.1197/j.aem.2006.11.037.

[127] R. H. H. Groenwold, “Informative missingness in electronic health record systems: the curse of knowing,” *Diagnostic Progn. Res.*, vol. 4, no. 8, 2020,

doi: https://doi.org/10.1186/s41512-020-00077-0.

[128] F. J. Prevosti and M. A. Chemisquy, "The impact of missing data on real morphological phylogenies: influence of the number and distribution of missing entries," *Cladistics*, vol. 26, pp. 326–339, 2010, doi: https://doi.org/10.1111/j.1096-0031.2009.00289.x.

[129] J. R. Barr, M. Sobel, and T. Thatcher, "Upsampling, a comparative study with new ideas," in *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, 2022, pp. 318–321.

[130] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," *J. Artif. Intell. Res.*, vol. 16, 2002, doi: 10.1613jair.953.

[131] D. N. Politis, *Subsampling*. Springer Science & Business Media, 1999.

[132] K. Fujiwara *et al.*, "Over- and Under-sampling approach for extremely imbalanced and small minority data problem in health record analysis," *Front. Public Heal.*, 2020.

[133] D. B. Rubin, "The Bayesian Bootstrap," *Ann. Stat.*, vol. 9, no. 1, pp. 130–134, 1981.

[134] G. Vandewiele *et al.*, "Overly Optimistic Prediction Results on Imbalanced Data: a Case Study of Flaws and Benefits when Applying Over-sampling," *Artif. Intell. Med.*, vol. 111, no. 1, 2021.

[135] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. van Calster, "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression," 2022.

[136] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *ISICA 2009: Computational Intelligence and Intelligent Systems*, 2009, pp. 461–471.

[137] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, "Cross-validation Metrics for Evaluating Classification Performance

on Imbalanced Data," *2019 Int. Conf. Comput. Control. Informatics its Appl.*, 2019.

[138] J. C. Jentzer *et al.*, "Changes in comorbidities, diagnoses, therapies and outcomes in a contemporary cardiac intensive care unit population," *Am. Heart J.*, vol. 215, pp. 12–19, 2019.

[139] P. Knapik, D. Ciesla, D. Borowik, P. Czempik, and T. Knapik, "Prolonged ventilation post cardiac surgery - tips and pitfalls of the prediction game," *J. Cardiothorac. Surg.*, vol. 6, p. 158, 2011, doi: 10.1186/1749-8090-6-158.

[140] J.-L. Vincent, "The clinical challenge of sepsis identification and monitoring," *PLoS Med*, vol. 13, no. 5, p. e1002022, 2016.

[141] E. A. J. Hoste *et al.*, "Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study," *Intensive Care Med.*, vol. 41, pp. 1411–1423, 2015.

[142] A. Bihorac *et al.*, "Development and Validation of a Machine-Learning Risk Algorithm for Major Complications and Death After Surgery," *Ann. Surg.*, vol. 269, no. 4, pp. 652–662, 2019.

[143] S. A. Gonzalez, "Acute liver failure," *BMJ Best Pract.*, 2022.

[144] A. Tsvetanova, M. Sperrin, N. Peek, I. Buchan, S. Hyland, and G. P. Martin, "Missing data was handled inconsistently in UK prediction models: a review of method used," *J. Clin. Epidemiol.*, vol. 140, pp. 149–158, 2021.

[145] R. Pandya and J. Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *Int. J. Comput. Appl.*, vol. 117, no. 16, 2015.

[146] A. Kapelner and J. Bleich, "Prediction with Missing Data via Bayesian Additive Regression Trees," *arXiv*, 2013.

[147] H. Luo, X. Pan, Q. Wang, S. Ye, and Y. Qian, "Logistic regression and random forest for effective imbalanced classification," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 2019, pp. 916–

917.

[148] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review,” *J. Big Data*, vol. 7, no. 70, 2020.

[149] E. Lin, Q. Chen, and X. Qi, “Deep reinforcement learning for imbalanced classification,” *Appl. Intell.*, vol. 50, pp. 2488–2502, 2020.

[150] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, 2019.

[151] R. Bisaso, K, S. A. Karungi, A. Kiragga, J. K. Mukonzo, and B. Castelnuovo, “A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients,” *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 77, 2018.

[152] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson, “Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards,” *Crit. Care Med.*, vol. 44, no. 2, pp. 368–374, 2016, doi: 10.1097/CCM.0000000000001571.

[153] J. Lee and D. J. Scott, “Open-access MIMIC-II Database for Intensive Care Research,” in *IEEE Engineering in Medicine and Biology Society*, 2011, pp. 8315–8318.

[154] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, “MIMIC-IV,” *PhysioNet*, vol. 1.0, 2021.

[155] V. F. J. Newcombe and A. Chow, “The features of the typical traumatic brain injury patient in the ICU are changing: what will this mean for the intensivist?,” *Curr. Opin. Crit. Care*, vol. 27, no. 2, pp. 80–86, 2021.

[156] A. Jones, A. P. Toft-Petersen, M. Shankar-Hari, D. A. Harrison, and K. M. Rowan, “Demographic Shifts, Case Mix, Activity and Outcome for Elderly

Patients Admitted to Adult General ICUs in England, Wales, and Northern Ireland," *Crit. Care Med.*, vol. 48, no. 4, pp. 466–474, 2020, doi: https://doi.org/10.1097/CCM.0000000000004211.

[157] K. R. Sepucha, F. J. Fowler, and A. G. Mulley, "Policy support for patient-centered care: the need for measurable improvements in decision quality," *Health Aff.*, vol. 23, no. 2, 2004.

[158] D. M. Berwick, "The Science of Improvement," *JAMA*, vol. 299, no. 10, pp. 1182–1184, 2008.

[159] A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar, "Personalized Risk Scoring for Critical Care Prognosis Using Mixtures of Gaussian Processes," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 1, pp. 207–218, 2018.

[160] J.-Y. Dupuis, "Predicting outcomes in cardiac surgery: risk stratification matters?," *Curr. Opin. Cardiol.*, vol. 23, no. 6, pp. 560–567, 2008.

[161] S. Siregar, D. Niboer, M. I. M. Versteegh, E. W. Steyerberg, and J. J. M. Takkenberg, "Methods for updating a risk prediction model for cardiac surgery: a statistical primer," *Interact. Cardiovasc. Thorac. Surg.*, vol. 28, no. 3, pp. 333–338, 2019.

[162] G. L. Hickey *et al.*, "Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models," *Eur. J. cardio-thoracic Surg.*, vol. 43, no. 6, pp. 1146–1152, 2013.

[163] Scottish Intensive Care Society Audit Group, "Audit of Critical Care in Scotland 2021 Reporting on 2020," 2021.

[164] F. Roques, P. Michel, A. R. Goldstone, and S. A. M. Nashef, "The logistic EuroSCORE," *Eur. Heart J.*, vol. 24, pp. 1–2, 2003, doi: 10.1016/S0195-668X(02)00799-6.

[165] W. A. Knaus, E. A. Draper, D. P. Wagner, and Zimmerman, "APACHE II: a severity of disease classification system," *Crit. Care Med.*, vol. 13, no. 10, pp. 818–829, 1985.

[166] P. Craig *et al.*, “Developing and evaluating complex interventions: the new Medical Research Council guidance,” *BMJ*, vol. 337, no. a1655, 2008, doi: 10.1136/bmj.a1655.

[167] J. Ritchie and L. Spencer, “Qualitative Data Analysis for Applied Policy Research,” in *Analysing Qualitative Data*, London, 1994, pp. 173–194.

[168] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual. Res. Psychol.*, vol. 3, no. 2, 2008.

[169] D. M. Sinitsky, S. B. Gowda, K. Dawas, and B. S. Fernando, “Morbidity and mortality meetings to improve patient safety: a survey of 109 consultant surgeons in London, United Kingdom,” *Patient Saf. Surg.*, vol. 13, no. 27, 2019.

[170] NICE, “Shared Decision Making,” *nice.org.uk*, 2022. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-guidelines/shared-decision-making.

[171] F. Roques *et al.*, “Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients,” *Eur. J. Cardio-thoracic Surg.*, vol. 15, pp. 816–823, 1999.

[172] J. Leipsic *et al.*, “Core competencies in cardiac CT for imaging structural heart disease interventions: an expert consensus statement,” *JACC Cardiovasc. Imaging*, vol. 12, no. 12, pp. 2555–2559, 2019.

[173] S. Matthiesen *et al.*, “Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: near-live feasibility and qualitative study,” *JMIR Hum. Factors*, vol. 8, no. 4, 2021.

[174] N. M. Thalji *et al.*, “Risk of conventional cardiac surgery among patients with severe left ventricular dysfunction in the era of mechanical circulatory support,” *J. Thorac. Cardiovasc. Surg.*, vol. 156, no. 4, pp. 1530-1540.e2, 2018.

[175] S. Karthik, A. D. Grayson, E. E. McCarron, D. M. Pullan, and M. J. Desmond, “Reexploration for bleeding after coronary artery bypass surgery: risk factors, outcomes, and the effect of time delay,” *Ann. Thorac. Surg.*, vol. 78, no. 2, pp.

527–534, 2004.

[176] C. Diez, D. Koch, O. Kuss, R.-E. Silber, I. Friedrich, and J. Boergermann, “Risk factors for mediastinitis after cardiac surgery - a retrospective analysis of 1700 patients,” *J. Cardiothorac. Surg.*, vol. 2, no. 23, 2007.

[177] S. W. Chen *et al.*, “Risk factor analysis of postoperative acute respiratory distress syndrome in valvular heart surgery,” *J. Crit. Care*, vol. 31, no. 1, pp. 139–43, 2016.

[178] J. J. Olivero, J. J. Olivero, P. T. Nguyen, and A. Kagan, “Acute kidney injury after cardiovascular surgery: an overview,” *Methodist Debakey Cardiovasc. J.*, vol. 8, no. 3, pp. 31–36, 2012.

[179] K. Koftis *et al.*, “Early delirium after cardiac surgery: an analysis of incidence and risk factors in elderly (≥65 years) and very elderly (≥80 years) patients,” *Clin. Interv. Aging*, vol. 13, pp. 1061–1070, 2018.

[180] M. A. Borger, J. Ivanov, R. D. Weisel, V. Rao, and C. M. Peniston, “Stroke during coronary bypass surgery: principal role of cerebral macroemboli,” *Eur. J. cardio-thoracic Surg.*, vol. 19, no. 5, pp. 627–632, 2001.

[181] J. Bucerius *et al.*, “Stroke after cardiac surgery: a risk factor analysis of 16,184 consecutive adult patients,” *Ann. Thorac. Surg.*, vol. 75, no. 2, pp. 472–478, 2003.

[182] J. Xu *et al.*, “Preoperative hidden renal dysfunction add an age dependent risk of progressive chrinic kidney disease after cardiac surgery,” *J. Cardiothorac. Surg.*, vol. 14, no. 151, 2019.

[183] C. Bäck, M. Hornum, P. S. Olsen, and C. H. Møller, “30-day mortality in frail patients undergoing cardiac surgery: the results of the frailty in cardiac surgery (FICS) copenhagen study,” *Scand. Cardiovasc. J.*, vol. 53, no. 6, pp. 348–354, 2019.

[184] J. T. McGinn Jr *et al.*, “Prevalence of dysglycemia among coronary artery bypass surgery patients with no previous diabetic history,” *J. Cardiothorac. Surg.*, vol. 6, no. 104, 2011.

[185] M. Navaratnarajah *et al.*, "Effect of glycaemic control on complications following cardiac surgery: literature review," *J. Cardiothorac. Surg.*, vol. 13, no. 10, 2018.

[186] C. H. M. Wong, J. S. K. Chan, D. Sanli, R. Rahimli, and A. Harky, "Aortic valve repair or replacement in patients with aortic regurgitation: A systematic review and meta-analysis," *J. Card. Surg.*, vol. 34, no. 6, pp. 377–384, 2019.

[187] S. Sannakki, D. Sannakki, J. J. Echebarria, and M. Patteril, "Preoperative assessment for cardiac surgery," *Anaesth. Intensive Care Med.*, vol. 22, no. 4, pp. 216–222, 2021.

[188] M. Mihalj, T. Carrel, R. D. Urman, F. Stueber, and M. M. Luedi, "Recommendations for preoperative assessment and shared decision-making in cardiac surgery," *Curr. Anesthesiol. Rep.*, vol. 10, pp. 185–195, 2020.

[189] The Royal College of Anaesthetists, "Perioperative Medicine: The Pathway to Better Surgical Care," 2015. [Online]. Available: https://www.rcoa.ac.uk/sites/default/files/documents/2019-08/Perioperative Medicine - The Pathway to Better Care.pdf.

[190] I. Kennedy, "The Report of the Public Inquiry into Children's Heart Surgery at the Bristol Royal Infirmary 1984-1995: Learning from Bristol," Norwich, 2001.

[191] A. Gogbashian, A. Sedrakyan, and T. Treasure, "EuroSCORE: a systematic review of international performance," *Eur. J. cardio-thoracic Surg.*, vol. 25, no. 5, pp. 695–700, 2004.

[192] F. Bhatti *et al.*, "The logistic EuroSCORE in cardiac surgery: how well does it predict operative risk?," *Heart*, vol. 92, no. 12, pp. 1817–1820, 2006.

[193] S. A. Nashef *et al.*, "EuroSCORE II," *Eur. J. Cardio-Thoracic Surg.*, vol. 41, no. 4, pp. 734–744, 2012.

[194] S. L. Feder, "Data quality in electronic health records research: Quality domains and assessment methods," *West. J. Nurs. Res.*, vol. 40, no. 5, pp. 753–766, 2017.

[195] B. Ehsani-Moghaddam, K. Martin, and J. A. Queenan, "Data quality in

healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data," *Heal. Inf. Manag. J.*, vol. 50, no. 1–2, pp. 88–92, 2019.

[196] L. E. Cowley, D. M. Farewell, S. Maguire, and A. M. Kemp, "Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature," *Diagnostic Progn. Res.*, vol. 3, no. 16, 2019.

[197] I. G. Stiell and G. A. Wells, "Methodologic standards for the development of clinical decision rules in emergency medicine," *Ann. Emerg. Med.*, vol. 33, no. 4, pp. 437–447, 1999.

[198] M. Hébert *et al.*, "Standardizing Postoperative Complications - Validating the Clavien-Dindo Complications Classification in Cardiac Surgery," *Semin. Thorac. Cardiovasc. Surg.*, 2020.

[199] C. Benstoem, A. Moza, R. Autschbach, C. Stoppe, and A. Goetzenich, "Evaluating Outcomes Used in Cardiothoracic Surgery Interventional Research: A Systematic Review of Reviews to Develop a Core Outcome Set," *PLoS One*, 2015.

[200] S. Keeney, F. Hasson, and H. McKenna, *The Delphi Technique in Nursing and Health Research*. Wiley-Blackwell, 2011.

[201] P. Clavien, J. Sanabria, and S. Strasberg, "Proposed classification of complication of surgery with examples of utility in cholecystectomy.," *Surgery*, vol. 111, pp. 518–526, 1992.

[202] D. Dindo, N. Demartines, and P. A. Clavien, "Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey," *Ann. Surg.*, vol. 240, no. 2, pp. 205–213, 2004.

[203] K. Slankamenac, R. Graf, J. Barkun, M. A. Puhan, and P.-A. Clavien, "The Comprehensive Complication Index: A Novel Continuous Scale to Measure Surgical Morbidity," *Ann. Surg.*, vol. 258, no. 1, pp. 1–7, 2013.

[204] D. Dindo, N. Demartines, and P.-A. Clavien, "Classification of surgical complications," *Ann. Surg.*, vol. 240, no. 2, pp. 205–213, 2004.

[205] H. Umezawa, J. Nakao, T. Matsutani, H. Kuwahara, M. Taga, and R. Ogawa, “Usefulness of the Clavien-Dindo Classification in Understanding the Limitations and Indications of Larynx-preserving Esophageal Reconstruction,” *Int. Open Access J. Am. Soc. Plast. Surg.*, vol. 4, no. 11, p. e1113, 2016.

[206] M. Bolliger, J.-A. Kroehnert, F. Molineus, D. Kandioler, and M. Schindl, “Experiences with the standardized classification of surgical complications (Clavien-Dindo) in general surgery patients,” *Eur. Surg.*, vol. 50, pp. 256–261, 2018.

[207] R. Winter *et al.*, “Standardizing the complication rate after breast reduction using the Clavien-Dindo classification,” *Surgery*, vol. 161, no. 5, pp. 1430–1435, 2017.

[208] R. Casadei *et al.*, “The usefulness of a grading system for complications resulting from pancreatic resections: a single center experience,” *Updates Surg.*, vol. 63, 2011.

[209] E. Bosma, M. J. J. Pullens, J. de Vries, and J. A. Roukema, “The impact of complications on quality of life following colorectal surgery: a prospective cohort study to evaluate the Clavien-Dindo classification system,” *Color. Dis.*, vol. 18, no. 6, pp. 594–602, 2015.

[210] P. J. Mentula and A. K. Leppäniemi, “Applicability of the Clavien-Dindo classification to emergency surgical procedures: a retrospective cohort study on 444 consecutive patients,” *Patient Saf. Surg.*, vol. 8, 2014.

[211] E. Monteiro *et al.*, “Assessment of the Clavien-Dindo classification system for complications in head and neck surgery,” *Head Neck*, vol. 124, no. 12, pp. 2726–2731, 2014.

[212] D. Mitropoulos, W. Artibani, M. Graefen, M. Remzi, M. Rouprêt, and M. Truss, “Reporting and Grading of Complications After Urologic Surgical Procedures: An ad hoc EAU Giodelines Panel Assessment and Recommendations,” *Eur. Urol.*, vol. 61, no. 2, pp. 341–349, 2012.

[213] P. A. Clavien *et al.*, “The Clavien-Dindo Classification of Surgical

Complications Five-Year Experience," *Ann. Surg.*, vol. 250, no. 2, pp. 187–196, 2009.

[214] V. Parsonnet, D. Dean, and A. D. Bernstein, "A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease," *Circulation*, vol. 79, no. 6 Pt 2, pp. I3-12, 1989.

[215] HMSO, "Learning from Bristol: the report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995," 2002.

[216] G. M. Fried and H. Gill, "Surgery through the keyholde: a new view of an old art," *McGill J. Med.*, vol. 10, no. 2, pp. 140–143, 2007.

[217] G. Ramsay *et al.*, "Reducing surgical mortality in Scotland by use of the WHO Surgical Safety Checklist," *Br. J. Surg.*, vol. 106, no. 8, pp. 1005–1011, 2019.

[218] S. W. Grant *et al.*, "Trends and outcomes for cardiac surgery in the United Kinfdom from 2002 to 2016," *JTCVS Open*, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666273621000140.

[219] L. Wang *et al.*, "Performance of EuroSCORE II in patients who have undergone heart valve surgery: a multicentre study in a Chinese population," *Eur. J. Cardiothorac. Surg.*, vol. 45, no. 2, pp. 359–364, 2014.

[220] A. Natarajan, S. Samadian, and S. Clark, "Coronary artery bypass surgery in elderly people," *Postgrad. Med. J.*, vol. 83, no. 977, pp. 154–158, 2007.

[221] I. A. Rahman and S. Kendall, "Cardiac surgery in the very elderly: it isn't all about survival," *Br. J. Cardiol.*, vol. 27, pp. 5–7, 2020.

[222] J. G. Canto *et al.*, "Symptom presentation of women with acute coronary syndromes: myth vs reality," *Arch. Intern. Med.*, vol. 167, no. 22, pp. 2405–2413, 2007.

[223] A. Sezai *et al.*, "Long-term results (three-year) of emergency coronary artery bypass grafting for patients with unstable angina pectoris," *Am. J. Cardiol.*, vol. 106, no. 4, pp. 511–516, 2010.

[224] A. S. Adabag *et al.*, "Preoperative pulmonary function and mortality after

cardiac surgery," *Am. Heart J.*, vol. 159, no. 4, pp. 691–697, 2010.

[225] N. J. McKeon, S. N. Timmins, H. Stewart, S. T. Yerkovich, and J. L. McKeon, "Diagnosis of COPD before cardiac surgery," *Eur. Respir. J.*, vol. 46, pp. 1498–1500, 2015.

[226] R. Hasbun, H. R. Vikram, L. A. Barakat, J. Buenconsejo, and V. J. Quagliarello, "Complicated Left-Sided Native Valve Endocarditis in Adults," *JAMA*, vol. 289, no. 15, pp. 1933–1940, 2003.

[227] P. J. Conlon, F. Jefferies, H. R. Krigman, G. R. Corey, D. J. Sexton, and M. A. Abramson, "Predictors of prognosis and risk of acute renal failure in bacterial endocarditis," *Clin. Nephrol.*, vol. 49, no. 2, pp. 96–101, 1998.

[228] R. S. D'Agostino *et al.*, "The Society of Thoracic Surgeons Adult Cardiac Surgery Database: 2018 Update on Outcomes and Quality," *Ann. Thorac. Surg.*, vol. 105, pp. 15–23, 2018.

[229] S. E. Tevis and G. D. Kennedy, "Postoperative complications and implication on patient-centered outcomes," *J. Surg. Res.*, vol. 181, no. 1, pp. 106–113, 2013.

[230] M. P. Gallery, S. M. Strasberg, and N. J. Soper, "Complications of Laparoscopic General Surgery," *Gastrointest. Endosc. Clin. N. Am.*, vol. 6, no. 2, pp. 423–444, 1996.

[231] C. C. McCoy, B. R. Englum, J. E. Keenan, S. N. Vaslef, M. L. Shapiro, and J. E. Scarborough, "Impact of specific postoperative complications on the outcomes of emergency general surgery patients," *Trauma Acute Care Surg.*, vol. 78, no. 5, pp. 912–919, 2015.

[232] K. R. Wanzel, C. G. Jamieson, and J. M. A. Bohnen, "Complications on a general surgery service: incidence and reporting," *Can. J. Surg.*, vol. 43, no. 2, pp. 113–117, 2000.

[233] T. M. Goyal, E. L. Idler, T. J. Krause, and R. J. Contrada, "Quality of Life Following Cardiac Surgery: Impact of the Severity and Course of Depressive Symptoms," *Psychosom. Med.*, vol. 67, no. 5, pp. 759–765, 2005.

[234] P. B. Rahmanian, A. Kröner, G. Langebartels, O. Özel, J. Wippermann, and T. Wahlers, “Impact of major non-cardiac complications on outcome following cardiac surgery procedures: logistic regression analysis in a very recent patient cohort.,” *Interact. Cardiovasc. Thorac. Surg.*, vol. 17, no. 2, pp. 319–327, 2013.

[235] H. McKenna, “The Delphi technique: a worthwhile research approach for nursing?,” *J. Adv. Nurs.*, vol. 19, no. 6, pp. 1221–5, 1994.

[236] A. Boel, V. Navarro-Compán, R. Landewé, and D. van der Heijde, “Two different invitation approaches for consecutive rounds of a Delphi survey led to comparable final outcome,” *J. Clin. Epidemiol.*, vol. 129, pp. 31–39, 2021.

[237] RCOA, “About the College,” 2021. https://www.rcoa.ac.uk/about-college.

[238] EACTAIC, “About EACTAIC,” 2021. https://www.eactaic.org/about-us/.

[239] SCTS, “About SCTS,” 2021. https://scts.org/about_scts/.

[240] SCATA, “SCATA,” 2021. 223059 (accessed Jul. 12, 2021).

[241] Qualtrics, “Qualtrics,” *https://www.qualtrics.com/uk/core-xm/survey-software/*, 2021. .

[242] RStudio, “RStudio.” 2021, [Online]. Available: https://www.rstudio.com/.

[243] C. Vogel, S. Zwolinsky, C. Griffiths, M. Hobbs, E. Henderson, and E. Wilkins, “A Delphi study to build consensus on the definition and use of big data in obesity research,” *Int. J. Obes.*, vol. 43, pp. 2573–2586, 2019.

[244] J. E. Mahoney *et al.*, “Modified Delphi Consensus to Suggest Key Elements of Stepping On Falls Prevention Program,” *Front. Public Heal.*, vol. 5, 2017.

[245] R. Veugelers, M. I. Gaakeer, P. Patka, and R. Huijsman, “Improving design choices in Delphi studies in medicine: the case of an exemplary physician multi-round panel study with 100% response,” *BMC Med. Res. Methodol.*, vol. 20, 2020.

[246] J. Kaufman *et al.*, “Identification of preliminary core outcome domains for communication about childhood vaccination: An online Delphi survey,”

Vaccine, vol. 36, no. 44, pp. 6520–6528, 2018.

[247] L. M. Hart and T. Wade, “Identifying research priorities in eating disorders: A Delphi study building consensus across clinicians, researchers, consumers, and carers in Australia,” *Int. J. Eat. Disord.*, 2019, [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/eat.23172.

[248] F. Hasson, S. Keeney, and H. McKenna, “Research guidelines for the Delphi survey technique,” *J. Adv. Nurs.*, vol. 32, no. 4, pp. 1008–15, 2000.

[249] HIS, “Mortality and Morbidity Reviews Practice Guide,” 2018.

[250] J. Higginson, R. Walters, and N. Fulop, “Mortality and morbidity meetings: an untaped resource for improving the governance of patient safety?,” *BMJ Qual. Saf.*, vol. 21, no. 7, pp. 576–585, 2012.

[251] M. F. Al-Haddad, A. Cadamy, E. Black, and K. Slade, “Are morbidity and mortality care review practices in Scottish intensive care units aligned to national standards?,” *J. Intensive Care Soc.*, vol. 19, no. 3, pp. 264–268, 2018.

[252] L. A. Celi, L. C. Hinske, G. Alterovitz, and P. Szolovits, “An artificial intelligence tool to predict fluid requirement in the intensive care unit: a p-roof-of-concept study,” *Crit. Care*, vol. 12, no. 6, p. R151, 2008.

[253] M. Schinkel, K. Paranjape, R. S. Nannan Panday, N. Skyttberg, and P. W. B. Nanayakkara, “Clinical applications of artificial intelligence in sepsis: A narrative review,” *Comput. Biol. Med.*, vol. 115, 2019.

[254] S. B. Hu, D. J. L. Wong, A. Correa, N. Li, and J. C. Deng, “Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model,” *PLoS One*, vol. 11, no. 8, 2016, doi: 10.1371/journal.pone.0161401.

[255] C. Okoli and S. D. Pawlowski, “The Delphi method as a research tool: an example, design considerations and applications,” *Inf. Manag.*, vol. 42, pp. 15–29, 2004.

[256] D. T. Engelman *et al.*, “Guidelines for Perioperative Care in Cardiac Surgery.

Enhanced Recovery After Surgery Society Recommendations," *JAMA Surg.*, vol. 154, no. 8, pp. 755–766, 2019.

[257] L. Rodríguez-Mañas *et al.*, "Searching for an Operational Definition of Frailty: A Delphi Method Based Consensus Statement. The Frailty Operative Definition-Consensus Conference Project," *Journals Gerontol. Ser. A*, vol. 68, no. 1, pp. 62–67, 2013.

[258] G. Murphy, "Categories and concepts," in *Noba textbook series: Psychology*, R. Biswas-Diener and E. Diener, Eds. Champaign, IL: DEF, 2021.

[259] American Society of Anesthesiologists, "ASA Physical Status Classification System," 2020. https://www.asahq.org/standards-and-guidelines/asa-physical-status-classification-system.

[260] D. Mayhew, V. Mendonca, and B. V. S. Murthy, "A review of ASA physical status - historical perspectives and modern developments," *Anesthesia*, vol. 74, no. 3, pp. 373–379, 2019.

[261] K. G. M. Moons *et al.*, "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration," *Ann. Intern. Med.*, vol. 162, no. 1, pp. W1-73, 2015, doi: 10.7326/M14-0698.

[262] Public Health Scotland, "Inpatient and Day Case Activity," *Data and Intelligence: Previously ISD Scotland*, 2022. https://www.isdscotland.org/Health-Topics/Hospital-Care/Inpatient-and-Day-Case-Activity/.

[263] GE Healthcare, "About Us," *https://www.gehealthcare.co.uk/about/about-ge-healthcare-systems#what-we-do*, 2021. https://www.gehealthcare.co.uk/about/about-ge-healthcare-systems#what-we-do (accessed Jan. 07, 2022).

[264] G. P. Burns, "Arterial blood gases made easy," *Clin. Med. (Northfield. Il).*, vol. 14, no. 1, pp. 66–68, 2014.

[265] H. H. Billett, "Hemoglobin and Hematocrit," in *Clinical Methods: The History,*

Physical, and Laboratory Examinations, H. K. Walker, W. D. Hall, and J. W. Hurst, Eds. 1990.

[266] C. Haldeman-Englert, M. Foley, and R. Turley Jr, “Bicarbonate,” *University of Rochester Medical Center Health Encyclopedia*, 2021. https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=bicarbonate (accessed Dec. 30, 2021).

[267] S. M. Nehring, A. Goyal, and B. C. Patel, “C Reactive Protein,” in *StatPearls*, 2021.

[268] S. S. Waikar and J. V Bonventre, “Creatinine Kinetics and the Definition of Acute Kidney Injury,” *J. Am. Soc. Nephrol.*, vol. 20, no. 3, pp. 672–679, 2009.

[269] K. Welch, “Fluid Balance,” *Learn. Disabil. Pract.*, vol. 13, no. 6, pp. 33–38, 2010.

[270] C. Bannerman, “Fluid Balance Monitoring,” *NHS Brighton and Sussex University Hospitals*, 2018. https://www.bsuh.nhs.uk/library/wp-content/uploads/sites/8/2019/01/Fluid-Balance-Monitoring-Poster.pdf (accessed Jan. 07, 2022).

[271] P. Swietach *et al.*, “Hydrogen ion dynamics in human red blood cells,” *J. Physiol.*, vol. 588, no. Pt24, pp. 4995–5014, 2010.

[272] K. Brandis, “The Hydrogen Ion,” in *Acid-base pHysiology*, 2015.

[273] L. W. Andersen, J. Mackenhauer, J. C. Roberts, K. M. Berg, M. N. Cocchi, and M. W. Donnino, “Etiology and therapeutic approach to elevated lactate,” *Mayo Clin. Proc.*, vol. 88, no. 10, pp. 1127–1140, 2013.

[274] G. N. Nakhoul *et al.*, “Serum potassium, end-stage renal disease and mortality in chronic kidney disease,” *Am. J. Nephrol.*, vol. 41, pp. 456–463, 2015.

[275] R. C. Morris Jr., O. Schmidlin, L. A. Frassetto, and A. Sebastian, “Relationship and interaction between sodium and potassium,” *J. Am. Coll. Nutr.*, vol. 25, no. sup3, pp. 262S-270S, 2006.

[276] R. Vanholder, T. Gryp, and G. Glorieux, “Urea and chronic kidney disease: the

comeback of the century? (in uraemia research)," *Nephrol. Dial. Transplant.*, vol. 33, no. 1, pp. 4–12, 2018.

[277] M. Legrand and D. Payen, "Understanding urine output in critically ill patients," *Ann. Intensive Care*, vol. 1, no. 13, 2011.

[278] L. Lapp, M. M. Bouamrane, K. Kavanagh, M. Roper, D. Young, and S. Schraag, "Evaluation of random forest and ensemble methods at predicting complications following cardiac surgery," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11526 LNAI, pp. 376–385, 2019, doi: 10.1007/978-3-030-21642-9_48.

[279] K. Birnie *et al.*, "Predictive models for kidney disease: improving global outcomes (KDIGO) defined acute kidney injury in UK cardiac surgery," *Crit. Care*, vol. 18, no. 606, 2014.

[280] J. R. Brown, R. S. Kramer, S. G. Coca, and C. R. Parikh, "Duration of Acute Kidney Injury Impacts Long-Term Survival After Cardiac Surgery," *Ann. Thorac. Surg.*, vol. 90, no. 4, pp. 1142–1148, 2010.

[281] S. T. H. Chew and N. C. Hwang, "Acute Kidney Injury After Cardiac Surgery: A Narrative Review of the Literature," *J. Cardiothorac. Vasc. Anesth.*, vol. 33, no. 4, pp. 1122–1138, 2019.

[282] M. H. Rosner and M. D. Okusa, "Acute kidney injury associated with cardiac surgery," *Clin. J. Am. Soc. Nephrol.*, vol. 1, no. 1, pp. 19–32, 2006.

[283] Oxford English Dictionary, "Delirium," *Lexico*. https://www.lexico.com/definition/delirium (accessed Mar. 24, 2022).

[284] E. W. Ely *et al.*, "Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU)," *Crit. Care Med.*, vol. 29, no. 7, pp. 1370–1379, 2001.

[285] S. K. Inouye, R. G. J. Westendorp, and J. S. Saczynski, "Delirium in elderly people," *Lancet*, vol. 383, no. 9920, pp. 911–922, 2014.

[286] C. H. Brown, "Delirium in the Cardiac Surgical Intensive Care Unit," *Curr.*

Opin. Anaesthesiol., vol. 27, no. 2, pp. 117–122, 2014.

[287] A. S. Evans *et al.*, “Current approach to diagnosis and treatment of delirium after cardiac surgery,” *Ann Card Anaesth*, vol. 19, no. 2, pp. 328–337, 2016.

[288] R. Collier, “Hospital-induced delirium hits hard,” *CMAJ*, vol. 184, no. 1, pp. 23–24, 2012.

[289] L. J. Gleason *et al.*, “Effect of delirium and other major complications on outcomes after elective surgery in older adults,” *JAMA Surg.*, vol. 150, no. 12, pp. 1134–1140, 2015.

[290] R. L. Plackett, “Karl Pearson and the Chi-Squared Test,” *Int. Stat. Rev.*, vol. 51, no. 1, pp. 59–72, 1983.

[291] B. Derrick, D. Toher, and P. White, “Why Welch’s test is Type I error robust,” *Quant. Methods Psychol.*, vol. 12, no. 1, pp. 30–38, 2016.

[292] L. I. Kuncheva, *Combining Patterns Classifiers: Methods and Algorithms.* New York: Wiley-Interscience, 2004.

[293] M. Kuhn *et al.*, “Package ‘caret.’” 2018.

[294] D. Meyer, “Package ‘e1071.’” 2021, [Online]. Available: https://cran.r-project.org/web/packages/e1071/e1071.pdf.

[295] S. Chatterjee, “Package ‘fastAdaboost.’” 2016.

[296] B. Greenwell, B. Boehmke, and J. Cunningham, “Package ‘gbm.’” 2019.

[297] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[298] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A Review,” *Int. J. Adv. soft Comput. its Appl.*, vol. 7, no. 3, pp. 176–204, 2015.

[299] M. Kuhn, “Package ‘caret.’” 2018, [Online]. Available: https://cran.r-project.org/web/packages/caret/caret.pdf.

[300] C. W. Hug, “Computer Science and Artificial Intelligence Laboratory Technical

Report Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models by," *Development*, 2009.

[301] A. E. . Johnson and R. . Mark, "Real-time mortality prediction in the Intensive Care Unit," in *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2017, pp. 994–1003.

[302] J. C. Ho, C. H. Lee, and J. Ghosh, "Septic Shock Prediction for Patients with Missing Data," *ACM Trans. Manag. Inf. Syst.*, vol. 5, no. 1, pp. 1–15, 2014.

[303] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Mak.*, vol. 51, 2011.

[304] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *The Thirteenth International Conference on Machine Learning*, 1996, pp. 148–156.

[305] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[306] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[307] T. van der Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints," *BMC Med. Res. Methodol.*, vol. 14, no. 137, 2014.

[308] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," in *Ensemble Machine Learning*, 2012, pp. 157–175.

[309] D. A. Pisner and D. M. Schnyer, "Chapter 6 - Support vector machine," in *Machine Learning: Methods and Applications to Brain Disorders*, 2020, pp. 101–121.

[310] S. Suthaharan, "Support Vector Machine," in *Machine Learning Models and Algorithms for Big Data Classification*, 2016, pp. 207–235.

[311] K. Mehrabani-Zeinabad, M. Doostfatemeh, and S. M. T. Ayatollahi, "An

Efficient and Effective Model to Handle Missing Data in Classification," *Biomed Res. Int.*, 2020.

[312] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[313] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–740, 1984.

[314] A. Kapelner and J. Bleich, "Package 'bartMachine,'" 2020, [Online]. Available: https://cran.r-project.org/web/packages/bartMachine/bartMachine.pdf.

[315] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., 1993.

[316] M. Kuhn, S. Weston, M. Culp, N. Coulter, and R. Quinlan, "Package 'C50.'" 2021.

[317] Z. Zhang, "Missing data imputation: focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, p. 9, 2016.

[318] X.-H. Zhou, G. J. Eckert, and W. M. Tierney, "Multiple imputation in public health research," *Stat. Med.*, vol. 20, pp. 1541–1549, 2001.

[319] S. Dávila and H. Rosado, "Performance of missing value imputation schemes in health-related data," in *IIE Annual Conference*, 2017, pp. 2105–2110.

[320] D. J. Stekhoven, "Package 'missForest.'" 2016.

[321] B. van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Med.*, vol. 17, no. 230, 2019.

[322] G. P. Findlay, A. P. L. Goodwin, K. Protopapa, N. C. E. Smith, and M. Mason, "Knowing the risk: A review of the peri-operative care of surgical patients," London, 2011.

[323] M.-M. Bouamrane and F. S. Mair, "A study of clinical and information management processes in the surgical pre-assessment clinic," *BMC Med.*

Inform. Decis. Mak., vol. 14, no. 22, 2014.

[324] S. W. Grant and D. P. Jenkins, “National Cardiac Surgery Activity and Outcomes Report 2002-2016,” 2020.

[325] M. Ruel *et al.*, “How detrimental is reexploration for bleeding after cardiac surgery?,” *J. Thorac. Cardiovasc. Surg.*, vol. 154, no. 3, pp. 927–935, Mar. 2018, doi: 10.1016/j.jtcvs.2016.04.097.

[326] F. D. Wang and C. H. Chang, “Risk factors of deep sternal wound infections in coronary artery bypass graft surgery,” *J. Cardiovasc. Surg. (Torino).*, vol. 41, no. 5, pp. 709–13, 2000.

[327] A. Salehi Omran *et al.*, “Superficial and deep sternal wound infection after more than 9000 coronary artery bypass graft (CABG): incidence, risk factors and mortality.,” *BMC Infect. Dis.*, vol. 7, p. 112, 2007, doi: 10.1186/1471-2334-7-112.

[328] T. L. Higgins, F. G. Estafanous, F. D. Loop, G. J. Beck, J. M. Blum, and L. Paranandi, “Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients,” *JAMA*, vol. 267, no. 17, pp. 2344–2348, 1992.

[329] D. M. Shahian *et al.*, “The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1-- coronary artery bypass grafting surgery,” *Ann. Thorac. Surg.*, vol. 88, no. 1 Suppl, pp. S2-22, 2009.

[330] J. Sanders, N. Makarious, A. Tocock, R. Magboo, A. Thomas, and L. M. Aitken, “Preoperative risk assessment tools for morbidity after cardiac surgery: a systematic review,” *Eur. J. Cardiovasc. Nurs.*, no. zvaz003, 2022.

[331] J. L. Rudolph *et al.*, “Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery,” *Circulation*, vol. 119, pp. 229–236, 2009.

[332] J. J. Levy and A. J. O’Malley, “Don’t dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning,” *BMC Med. Res. Methodol.*, vol. 20, no. 171, 2020.

[333] Z. Yang, W. H. Tang, A. Shintemirov, and Q. H. Wu, "Association Rule Mining-Based Dissolved Gas Analysis for Fault Diagnosis of Power Transformers," *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.*, vol. 39, no. 6, pp. 597–610, 2009.

[334] T. D. Henry *et al.*, "Long-term survival in patients with refractory angina," *Eur. Heart J.*, vol. 34, no. 34, pp. 2683–2688, 2013.

[335] S. A. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, and R. Salamon, "European system for cardiac operative risk evaluation (EuroSCORE)," *Eur. J. Cardiothorac. Surg.*, vol. 16, no. 1, pp. 9–13, 1999.

[336] D. Kristovic *et al.*, "Cardiac surgery-associated acute kidney injury: risk factors analysis and comparison of prediction models," *Interact. Cardiovasc. Thorac. Surg.*, vol. 21, no. 3, pp. 366–373, 2015.

[337] H. Gombotz and H. Knotzer, "Preoperative identification of patients with increased risk for perioperative bleeding," *Curr. Opin. Anesthesiol.*, vol. 26, no. 1, pp. 82–90, 2013.

[338] M. K. Ford, S. Beattie, and D. N. Wijeysundera, "Systematic Review: Prediction of Perioperative Cardiac Complications and Mortality by the Revised Cardiac Risk Index," *Ann. Intern. Med.*, vol. 152, pp. 26–35, 2010.

[339] A. L. Shroyer *et al.*, "The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models," *Ann. Thorac. Surg.*, vol. 75, no. 6, pp. 1856–1864, 2003.

[340] T. K. Wang, A. Y. Li, T. Ramanathan, R. A. Stewart, G. Gamble, and H. D. White, "Comparison of four risk scores for comtemporary isolated coronary artery bypass grafting," *Hear. lung Circ.*, vol. 23, no. 5, pp. 469–474, 2014.

[341] J.-Y. Dupuis, F. Wang, H. Nathan, M. Lam, S. Grimes, and M. Bourke, "The Cardiac Anaesthesia Risk Evaluation Score: A clinically useful predictor of mortality and morbidity after cardiac surgery," *Anesthesiology*, vol. 94, pp. 194–204, 2001.

[342] H. J. Geissler *et al.*, "Risk stratification in heart surgery: comparison of six score

systems," *Eur. J. cardio-thoracic Surg.*, vol. 17, no. 4, pp. 400–406, 2000.

[343] H. Hirose *et al.*, "EuroSCORE predicts postoperative mortality, certain morbidities, and recovery time," *Interact. Cardiovasc. Thorac. Surg.*, vol. 9, no. 4, pp. 613–617, 2009.

[344] O. Pitkanen, M. Niskanen, S. Rehnberg, M. Hippelainen, and M. Hynynen, "Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE.," *Eur. J. cardio-thoracic Surg.*, vol. 18, no. 6, pp. 703–710, Dec. 2000.

[345] S. Scolletta, P. Giomarelli, G. Cevenini, and B. Biagioli, "Estimation of morbidity risk factors in intensive care unit: a Bayesian discriminant approach: 028," *Eur. J. Anaesthesiol.*, vol. 21, no. 14, 2004.

[346] A. U. Syed, H. Fawzy, A. Farag, and A. Nemlander, "Predictive value of EuroSCORE and Parsonnet scoring in Saudi population," *Hear. lung Circ.*, vol. 13, no. 4, pp. 384–388, 2004.

[347] T. K. Wang, S. Harmos, G. D. Gamble, T. Ramanathan, and P. N. Ruygrok, "Performance of contemporary surgical risk scores for mitral valve surgery," *J. Card. Surg.*, vol. 32, no. 3, pp. 172–176, 2017.

[348] M. D. Coleman, S. Shaefi, and R. N. Sladen, "Preventing acute kidney injury after cardiac surgery," *Curr. Opin. Anesthesiol.*, vol. 24, no. 1, pp. 70–76, 2011.

[349] A. Candela-Toha *et al.*, "Predicting acute renal failure after cardiac surgery: external validation of two new clinical scores," *Clin. J. Am. Soc. Nephrol.*, vol. 3, no. 5, pp. 1260–1265, 2008.

[350] G. M. Chertow, E. M. Levy, K. E. Hammermeister, F. Grover, and J. Daley, "Independent Association between Acute Renal Failure and Mortality following Cardiac Surgery," *Am. J. Med.*, vol. 104, no. 4, pp. 343–348, 1998.

[351] C. V Thakar, S. Arrigain, S. Worley, J.-P. Yared, and E. P. Paganini, "A Clinical Score to Predict Acute Renal Failure after Cardiac Surgery," *J. Am. Soc. Nephrol.*, vol. 16, no. 1, pp. 162–168, 2005.

[352] R. H. Mehta *et al.*, “Bedside Tool for Predicting the Risk of Postoperative Dialysis in Patients Undergoing Cardiac Surgery,” *Circulation*, vol. 114, pp. 2208–2216, 2006.

[353] D. N. Wijeysundera *et al.*, “Derivation and Validation of a Simplified Predictive Index for Renal Replacement Therapy after Cardiac Surgery,” *JAMA*, vol. 297, no. 16, pp. 1801–1809, 2007.

[354] C. E. Hobson *et al.*, “Acute kidney injury is associated with increased lon-term mortality after cardiothoracic surgery,” *Circulation*, vol. 119, pp. 2444–2453, 2009.

[355] J. Penny-Dimri, C. Bergmeir, C. M. Reid, J. Williams-Spence, A. D. Cochrane, and J. A. Smith, “Machine Learning Algorithms for Predicting and Risk Profiling for Cardiac Surgery-Associated Acute Kidney Injury,” *Semin. Thorac. Cardiovasc. Surg.*, vol. 33, no. 3, pp. 735–745, 2021.

[356] Y. Li *et al.*, “A novel machine learning algorithm, Bayesian networks model, to predict the high-risk patients with cardiac surgery-associated acute kidney injury,” *Clin. Cardiol.*, vol. 43, no. 7, pp. 752–761, 2020.

[357] P. Thottakkara *et al.*, “Application of Machine Learning Techniques to High-Dimensional Clinical Data to Forecast Postoperative Complications,” *PLoS One*, vol. 11, no. 5, p. e0155705, 2016.

[358] M. A. Rozner, “Preoperative Prediction of Postoperative Delirium,” *JAMA*, vol. 271, no. 20, pp. 1573–1574, 1994.

[359] S. T. Dillon *et al.*, “Higher C-reactive protein levels predict postoperative delirium in older patients undergoing major elective surgery: a longitudinal nested case-control study,” *Biol. Psychiatry*, vol. 81, no. 2, pp. 145–153, 2017.

[360] A. Kapoor and J. E. Fixley, “Postoperative Delirium,” in *Perioperative Medicine*, 2011, pp. 531–542.

[361] L. Ansaloni *et al.*, “Risk factors and incidence of postoperative delirium in elderly patients after elective and emergency surgery,” *Br. J. Surg.*, vol. 97, no. 2, pp. 273–280, 2010.

[362] X. Chen, Y. Lao, Y. Zhang, L. Qiao, and Y. Zhuang, "Risk predictive models for delirium in the intensive care unit: a systematic review and meta-analysis," *Ann. Palliat. Med.*, vol. 10, no. 2, pp. 1467–1479, 2021.

[363] J. Chen, J. Yu, and A. Zhang, "Delirium risk prediction models for intensive care unit patients: A systematic review," *Intensive Crit. Care Nurs.*, vol. 60, 2020.

[364] M. M. Ruppert *et al.*, "ICU Delirium-Prediction Models: A Systematic Review," *Crit. Care Explor.*, vol. 2, no. e0296, 2020.

[365] H. Chen, L. Mo, H. Hu, Y. Ou, and J. Luo, "Risk factors of postoperative delirium after cardiac surgery: a meta-analysis," *J. Cardiothorac. Surg.*, vol. 16, no. 113, 2021.

[366] R. Katznelson *et al.*, "Delirium following vascular surgery: increased incidence with preoperative β-blocker administration," *Can. J. Anesth.*, vol. 56, no. 793, 2009.

[367] Y. Kawatani *et al.*, "Development of delirium in the intensive care unit in patients after endovascular aortic repair: a retrospective evaluation of the prevalence and risk factors," *Crit. Care Res. Pract.*, p. 405817, 2015.

[368] M. van Smeden, R. H. H. Groenwold, and K. G. M. Moons, "A cautionary note on the use of the missing indicator method for handling missing data in prediction research," *J. Clin. Epidemiol.*, vol. 125, pp. 188–190, 2020.

[369] G. C. Galyfos, G. E. Geropapas, A. Sianou, F. Sigala, and K. Filis, "Risk factors for postoperative delirium in patients undergoing vascular surgery," *J. Vasc. Surg.*, vol. 66, no. 3, pp. 937–946, 2017.

[370] L. Visser *et al.*, "Predicting postoperative delirium after vascular surgical procedures," *J. Vasc. Surg.*, vol. 62, no. 1, pp. 183–189, 2015.

[371] B. Koebrugge, R. J. A. van Wensen, K. Bosscha, P. L. J. Dautzenberg, and O. H. J. Koning, "Delirium after emergency/elective open and endovascular aoroiliac surgery at a surgical ward with a high-standard delirium care protocol," *Vascular*, vol. 18, no. 5, pp. 279–287, 2010.

[372] L. J. Krzych *et al.*, "Complex assessment of the incidence and risk factors of delirium in a large cohort of cardiac surgery patients: a single-center 6-year experience," *Biomed Res. Int.*, p. 835850, 2013.

[373] S. Cai *et al.*, "Preoperative cardiac function parameters as valuable predictors for nurses to recognise delirium after cardiac surgery: A prospective cohort study," *Eur. J. Cardiovasc. Nurs.*, vol. 19, no. 4, pp. 310–319, 2020.

[374] K. Koftis, A. Szylinska, M. Listewnik, M. Brykczynski, E. W. Ely, and I. Rotter, "Diabetes and elevated preoperative HbA1c level as risk factors for postoperative delirium after cardiac surgery: an observational cohort study," *Neuropsychiatr. Dis. Treat.*, vol. 15, pp. 511–521, 2019.

[375] P. J. Tully, R. A. Baker, H. R. Winefield, and D. A. Turnbull, "Depression, anxiety disorders and type D personality as risk factors for delirium after cardiac surgery," *Aust. N. Z. J. Psychiatry*, vol. 44, no. 11, pp. 1005–1011, 2010.

[376] J. Kazmierski *et al.*, "Incidence and predictors of delirium after cardiac surgery: Results from the IPDACS Study," *J. Psychosom. Res.*, vol. 69, no. 2, pp. 179–185, 2010.

[377] A. G. Benoit *et al.*, "Risk factors and prevalence of perioperative cognitive dysfunction in abdominal aneurysm patients," *J. Vasc. Surg.*, vol. 42, no. 5, pp. 884–890, 2005.

[378] K. Karkouti, D. N. Wijeysundera, T. M. Yau, S. A. McCluskey, A. van Rensburg, and W. S. Beattie, "The influence of baseline hemoglobin concentration on tolerance of anemia in cardiac surgery," *Transfusion*, vol. 48, no. 4, pp. 666–672, 2008.

[379] F. Husain-Syed *et al.*, "Preoperative Renal Functional Reserve Predicts Risk of Acute Kidney Injury After Cardiac Operation," *Ann. Thorac. Surg.*, vol. 105, no. 4, pp. 1094–1101, 2018.

[380] J. O. Friedrich, N. Adhikari, M. S. Herridge, and J. Beyene, "Meta-Analysis: Low-Dose Dopamine Increases Urine Output but Does Not Prevent Renal Dysfunction or Death," *Ann. Intern. Med.*, vol. 142, pp. 510–524, 2005.

[381] M. Park, S. G. Coca, S. U. Nigwekar, A. X. Garg, S. Garwood, and C. R. Parikh, "Prevention and Treatment of Acute Kidney Injury in Patients Undergoing Cardiac Surgery: A Systematic Review," *Am. J. Nephrol.*, vol. 31, pp. 408–418, 2010.

[382] B. Hiemstra *et al.*, "Dopamine in critically ill patients with cardiac dysfunction: A systematic review with meta-analysis and trial sequential analysis," *Acta Anaesthesiol. Scand.*, vol. 63, no. 4, pp. 424–437, 2018.

[383] K. Grynberg *et al.*, "Early serum creatinine accurately predicts acute kidney injury post cardiac surgery," *BMC Nephrol.*, vol. 18, no. 93, 2017.

[384] P. J. Connelly, M. Lonergan, E. Soto-Pedre, L. Donnelly, K. Zhou, and E. R. Pearson, "Acute kidney injury, plasma lactate concentrations and lactic acidosis in metformin users: A GoDarts study," *Diabetes, Obes. Metab.*, vol. 19, no. 11, pp. 1579–1586, 2017.

[385] H. Fan *et al.*, "Development and validation of a dynamic delirium prediction rule in patients admitted to the Intensive Care Units (DYNAMIC-ICU): A prospective cohort study," *Int. J. Nurs. Stud.*, vol. 93, pp. 64–73, 2019.

[386] K. J. Moon, Y. Jin, T. Jin, and S.-M. Lee, "Development and validation of an automated delirium risk assessment system (Auto-DelRAS) implemented in the electronic health record system," *Int. J. Nurs. Stud.*, vol. 77, pp. 46–53, 2018.

[387] A. Marra *et al.*, "Acute Brain Dysfunction. Development and Validation of a Daily Prediction Model," *Chest*, vol. 154, no. 2, pp. 293–301, 2018.

[388] J. Oh *et al.*, "Prediction and early detection of delirium in the intensive care unit by using heart rate variability and machine learning," *Physiol. Meas.*, vol. 39, no. 3, 2018.

[389] S. P. Ellner and J. Guckenheimer, "What are dynamic models?," in *Dynamic Models in Biology*, Princeton University Press, 2011, pp. 1–30.

[390] M. G. Cole and J. Mccusker, "Delirium in older adults: a chronic cognitive disorder?," *Int. Psychogeriatrics*, vol. 28, no. 8, pp. 1229–1233, 2016.

[391] M. A. Pisani, T. E. Murphy, P. H. van Ness, K. L. B. Araujo, and S. K. Inouye, "Characteristics associated with delirium in older patients in a medical intensive care unit," *Arch. Intern. Med. Intern. Med.*, vol. 167, no. 15, pp. 1629–1634, 2007.

[392] R. Horacek, B. Krnacova, J. Prasko, and K. Latalova, "Delirium as a complication of the surgical intensive care," *Neuropsychiatr. Dis. Treat.*, vol. 12, pp. 2425–2434, 2016.

[393] R. Y. Y. Wan and M. Ostermann, "Acute Kidney Injury and Delirium: Kidney-Brain Crosstalk," in *Annual Update in Intensive Care and Emergency Medicine*, 2019, pp. 397–404.

[394] T. K. M. Wang, D. H. M. Choi, T. Ramanathan, and P. N. Ruygrok, "Comparing performance of risk scores for combined aortic valve replacement and coronary bypass grafting surgery," *Hear. lung Circ.*, vol. 25, no. 11, pp. 1118–1123, 2016.

[395] J. A. Lopes and S. Jorge, "The RIFLE and AKIN classifications for acute kidney injury: a critical and comprehensive review," *Clin. Kidney J.*, vol. 6, no. 1, pp. 8–14, 2013.

[396] D. Gusmao-Flores, J. I. F. Salluh, R. A. Chalhub, and L. C. Quarantini, "The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: a systematic review and meta-analysis of clinical studies," *Crit. Care*, vol. 16, no. 4, p. R115, 2012.

[397] G. Mistraletti, P. Pelosi, E. S. Mantovani, M. Bernardino, and C. Gregoretti, "Delirium: Clinical approach and prevention," *Best Pract. Res. Clin. Anaesthesiol.*, vol. 26, pp. 311–326, 2012.

[398] Institute of Medicine, "Promoting Adoption of Clinical Practice Guidelines," in *Clinical Practice Guidelines We Can Trust*, 2011.

[399] H. White, "Theory-based impact evaluation: principles and practice," *J. Dev. Eff.*, vol. 1, no. 3, pp. 271–284, 2009.

[400] M. Medalion *et al.*, “The effect of cardiac angiography timing, contrast media dose, and preoperative renal function on acute renal failure after coronary artery bypass grafting,” *J. Thorac. Cardiovasc. Surg.*, vol. 139, no. 6, pp. 1539–1544, 2010.

[401] L. Tse *et al.*, “Pharmacological risk factors for delirium after cardiac surgery: a review,” *Curr. Neuropharmacol.*, vol. 10, no. 3, pp. 181–196, 2012.

[402] R. H. Thiele, J. M. Isbell, and M. H. Rosner, “AKI Associated with Cardiac Surgery,” *Clin. J. Am. Soc. Nephrol.*, vol. 10, no. 3, pp. 500–514, 2015.

[403] R. Kumar and R. Gandhi, “Reasons for cancellation of operation on the day of intended surgery in a multidisciplinary 500 bedded hospital,” *J. Anaesthesiol. Clin. Pharmacol.*, vol. 28, no. 1, pp. 66–69, 2012.

[404] C. R. Parikh and G. Han, “Variation in performance of kidney injury biomarkers due to cause of acute kidney injury,” *Am. J. Kidney Dis.*, vol. 62, no. 6, pp. 1023–1026, 2013.

[405] J. Labarère, R. Bertrand, and M. J. Fine, “How to derive and validate clinical prediction models for use in intensive care medicine,” *Intensive Care Med.*, vol. 40, pp. 513–527, 2014.

[406] Oxford English Dictionary, “mortality, n.,” *OED Online*, 2021. https://www.oed.com/view/Entry/122442?redirectedFrom=mortality (accessed Sep. 21, 2021).

[407] Oxford English Dictionary, “morbidity, n.,” *OED Online*, 2021. https://www.oed.com/view/Entry/122124?redirectedFrom=morbidity (accessed Sep. 21, 2021).

[408] M. H. Weil, “Defining Hemodynamic Instability,” in *Functional Hemodynamic Monitoring*, 2005, pp. 9–17.

[409] J. B. Patel and A. Sapra, “Nephrotoxic Medications,” *StatPearls*, 2021.

[410] C. B. Overgaard and V. Džavík, “Inotropes and Vasopressors,” *Circulation*, vol. 118, no. 10, pp. 1047–1056, 2008.

Appendix 3.1: Interview Schedule

Introduction

- What is your job title and expertise?
- Could you please describe your role?
- What are the main challenges in your day-to-day practice?
- What are the main challenges in cardio-thoracic surgery?

Perioperative medicine

- What could be done to improve cardiac perioperative medicine?
- What should be in priority for improvement?
- What could improve the care of your patients?
- What in your opinion is perioperative medicine? *What perioperative medicine entails?*

Risk scoring

- What is your opinion on risk scoring tools?
- Why do you find the risk stratification systems to be important / not important?
- Do you use risk prediction tools? Which ones?
- At which stage are you using the risk prediction system (before surgery, during, after)?
- What information are you using from these risk prediction systems?
- How do you use that information? *Do you have an example?*
- What kind of information do you find the most helpful from risk prediction models (e.g. probability of the outcome, a certain threshold or score, alert)?
- What are the advantages of the risk prediction systems you are using?
- What are the disadvantages of the risk prediction systems you are using?
- Are there any problems that the currently available predictions systems don’t address? *Which problems?*
- For what problem would you want a new risk prediction system?
- What kind of decision support do you need?

Surgical complications

- What is your take on surgical complications? *Are they a problem? Do they happen often? Are they recorded well enough? Do you look at complications as a measure of outcome after surgery?*
- Do you perceive complications to be an issue for patients?
- If yes, which complications do you think are the most common for patients?
- What could be done to prevent complications? *Using a risk tool? Decision support? Better patient management? How?*
- Do you think complications can be prevented or managed by using a risk prediction tool?
- If yes, what kind of complications should that tool predict?

Is there anything else you would like to say?

Appendix 4.1: Delphi Study Round 1 Questionnaire

Demographics and Expertise

1. Are you in any way involved with cardiac surgery patients? (Can be preoperatively, intraoperatively, and/or postoperatively.)
    - Yes
    - No
2. What is your country of residence?
3. What is your speciality?
    - Cardiac Anaesthetist
    - Cardiac Surgeon
    - Cardiac Critical Care
    - Other (Please specify)
4. How long have you worked in this field (in years)?
5. What stages of cardiac surgery are you involved in?
    - Preoperative assessment
    - Decision making (e.g., if patient is fit for surgery)
    - The surgery itself
    - ICU
    - Long-term follow-up of the patient
    - Other (Please explain)

Defining Postoperative Complications

6. How would you define the term “postoperative complication” in cardiac surgery? Please bring examples to explain.

7. How useful do you think it is to classify postoperative complications for cardiac surgery? Please explain your answer.
    - Extremely useful
    - Very useful
    - Moderately useful
    - Slightly useful
    - Not at all useful

Classification of Postoperative Complications

8. In order to classify postoperative complication in cardiac surgery, how many grades should there be?
9. Based on the number of grades suggested, how would you define each of these grades?
10. Please provide an example of a complication for each of the suggested levels.
11. Is there anything else you would like to comment on the topic of postoperative complications?

Appendix 4.2: Delphi Study Round 2 Questionnaire

1. Are you in any way involved with cardiac surgery patients? (Can be preoperatively, intraoperatively and/or postoperatively.)
2. From the selection below, please choose how postoperative complications after cardiac surgery should be defined. Multiple options are possible if combination of definitions is deemed important.
    - An unplanned adverse event occurring after cardiac surgery that may be caused or compounded by the surgical process.
    - An unplanned adverse event arising as a result of cardiac surgery, which was otherwise unlikely to have occurred in the same period.
    - Any adverse event that impairs a patient's physical, cognitive, psychological, or emotional function and quality of life.
    - Any deviation from the ideal recovery pattern after cardiac surgery.
    - Unexpected, or expected but unwanted, outcome of cardiac surgery which notably delays recovery from the procedure compared to the desired outcome or leads to the patient failing to derive the intended benefits of surgery.
    - Any event resulting from surgery which lengthens the patient's stay in hospital or reduces their quality of life beyond normal.
    - Any unplanned clinical event that leads to a delay in hospital discharge or requires additional treatment or intervention to mitigate or reverse the event.
    - Any deviation of any physiological system which adversely affects rapid recovery to good health.

- An event which may have an impact on patient's survival or quality and longevity.

3. Based on your opinion, should death be included in the grading of postoperative complications? Please explain your answer.
    - Yes
    - No
4. For Mild Complication the definition should be (multiple options are possible):
    - No notable effect on overall length of stay
    - No notable effect on final outcome
    - Lasting 1 week – 1 month
    - No or only short-term clinical relevance
    - Mildly debilitating
    - Common
    - No or small amount of intervention required
    - Minimal impact on patient
    - Minimal impact on institution
5. For Moderate Complication the definition should be (multiple options are possible):
    - Some effect on overall length of stay
    - Some effect on final outcome
    - Lasting 1-3 months
    - Acutely important, but less clinical consequence long term
    - Moderately debilitating
    - Less common

- Some intervention required
- Limited impact on patient
- Limited impact on institution

6. For Severe Complication the definition should be (multiple options are possible):
    - Extended length of stay
    - Potentially life-threatening
    - Lasting 3 months – 1 year
    - With sustained relevance and life-limiting
    - Severely debilitating
    - Uncommon
    - Notable amount of intervention required
    - Notable or long-standing impact on the patient
    - Notable or long-standing impact on institution
7. If you have any comments regarding the topic of postoperative complications following cardiac surgery, please write them below.

Appendix 4.3: Examples for Each Proposed Complication Level

Complication level	**Example of Complication**
Mild	Atrial Fibrillation
	Leg wound breakdown
	Surgical bleeding tamponade and discharge from ICU and hospital at usual time point
	Minor bleeding following nasal temperature probe insertion
	Residual pneumothorax
	Hypoxia
	High oxygen concentration via face mask to maintain adequate oxygenation
	Mild acute kidney injury
	Blood transfusion
	Uncontrolled pain
	More than one inotrope
	Coagulopathy requiring products
	Bleeding resulting in blood transfusion but no other measurable pathology
	Lower respiratory tract infection delaying but not preventing full recovery
	Constipation
	Minor adverse drug reactions
	Pain that is resolved with analgesia
	Simple wound infections
	Bleeding
	Acute kidney injury not needing renal replacement therapy
	Chest infection requiring antibiotics and two days extra in hospital
	Minimal bleeding from a wound requiring redressing
	Torn skin from tape, etc
	Bilateral basal collapse
	Bild inotropic support
	Minor chest infection
	Postoperative fever
	Reduced mobility
	Dental damage
	Mild postoperative pain
	Mild respiratory tract infection
	Mild urinary tract infection
	Bruising
	Drug reaction causing a short-lived skin rash
	Simple chest infection

Complication level	Example of Complication
Moderate	Pneumonia needing prolonged ventilation and tracheostomy
	Cognitive impairments
	Wound infection leading to increased length of stay
	Seizures
	Prolonged sedation
	Bleeding
	Acute kidney injury not needing renal replacement therapy
	Atrial Fibrillation
	Acute renal failure requiring renal replacement therapy for several days
	Moderate respiratory impairment
	Severe cardiovascular compromise
	Postoperative chest infection
	Myocardial infarction
	More than tyhree drugs for cardiovascular support
	Postoperative delirium
	Persisting atrial fibrillation
	Chest infection
	Prolonged ventilation
	Sternal wound infection
	Pneumonia
	Bleeding
	Renal failure needing temporary dialysis
Severe	Reopening
	Stroke
	Catastrophic cardiac dysfunction
	Death or severe disability
	Cardiac arrest
	Renal failure requiring ongoing support
	Respiratory arrest
	Severe acute kidney injury
	Renal replacement therapy in previous normal renal function
	Deep wound infection
	Delirium
	Sternal dehiscence
	Unable to extubate
	Bleeding resulting in death
	Neurological injury leading to long-term hospitalisation
	Renal failure requiring dialysis
	Significant injury to patient which does not allow independent living at discharge
	Limb loss
	Permanent cognitive deficit
	Permanent frailty

Complication level	**Example of Complication**
	Aortic dissection
	Sepsis
	Multiorgan failure
	Bleeding requiring products and re-opening
	Major haemorrhage
	Cardiogenic shock
	Pneumonia

Appendix 5.1: TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Checklist Item	Section No
Background and objectives	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	6.1 & 7.2 & 8.2
	Specify the objectives, including whether the study describes the development or validation of the model or both.	6.1 & 7.1 & 8.1
Source of data	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5.5
	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5.4
Participants	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5.3
	Describe eligibility criteria for participants.	5.4
	Give details of treatments received, if relevant.	5.4
Outcome	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	5.6 & 5.9.1 & 6.2.1
	Report any actions to blind assessment of the outcome to be predicted.	N/A
Predictors	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	5.5.1.2 & 5.5.2.1
	Report any actions to blind assessment of predictors for the outcome and other predictors.	N/A
Sample size	Explain how the study size was arrived at.	5.4 & 6.2.1.3, 6.2.2
Missing data	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	5.5.1.1. & 5.5.2.2.
Statistical analysis methods	Describe how predictors were handled in the analyses.	5.7 & 5.8 & 5.9.2 & 6.2.3
	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	5.8.1.
	For validation, describe how the predictions were calculated.	5.8 & 5.9.3
	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	5.8.2.
	Describe any model updating (e.g., recalibration) arising from the validation, if done.	N/A
Risk groups	Provide details on how risk groups were created, if done.	N/A
Development vs. validation	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	N/A
Participants	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Appendix 6.2
	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	5.5.2.2 & Appendix 6.2 & 7.1 & 8.1
	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Appendix 6.3
Model development	Specify the number of participants and outcome events in each analysis.	6.3.1 & 7.4.1 & 8.4.1
	If done, report the unadjusted association between each candidate predictor and outcome.	6.3.1 & 6.4.1 & 6.5.1
Model specification	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Done where possible:6.5.2
	Explain how to the use the prediction model.	9.3
Model performance	Report performance measures (with CIs) for the prediction model.	6.3.2 & 6.4.2 & 6.5.2
Model-updating	If done, report the results from any model updating (i.e., model specification, model performance).	7.5 & 7.6 & 8.5 & 8.6

Section/Topic	Checklist Item	Section No
Limitations	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	9.2
Interpretation	For validation, discuss the results with reference to performance in the development data, and any other validation data.	N/A
	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	6.6 & 7.7 & 8.7
Implications	Discuss the potential clinical use of the model and implications for future research.	9.3 & 9.4

Appendix 6.1: Prevalence of Each Severe Postoperative Complication Recorded in the CaTHI Database

Complication	Prevalence
Acute renal failure	1.71 (1.43 - 2.05)
Acute renal dysfunction	0.01 (0.00 - 0.08)
Acute kidney injury	0.04 (0.01 - 0.13)
Cardiac arrest	0.82 (0.63 - 1.06)
Left ventricular wall dissection	0.03 (0.01 - 0.11)
Reopening requiring cardiopulmonary bypass	0.04 (0.01 - 0.13)
Severe heart failure	0.35 (0.24 - 0.52)
Biventricular failure	0.01 (0.00 - 0.08)
Cardiogenic shock	0.03 (0.01 - 0.11)
Deterioration in LV function	0.03 (0.01 - 0.11)
Pericardial effusion	0.12 (0.06 - 0.23)
Paraparesis	0.01 (0.00 - 0.08)
Stroke	0.80 (0.62 - 1.05)
Acute delirium	0.01 (0.00 - 0.08)
Haemorrhagic stroke	0.01 (0.00 - 0.08)
Biventricular failure	0.01 (0.00 - 0.08)
Adult respiratory distress syndrome	0.32 (0.21 - 0.49)
Percutaneous tracheostomy	0.98 (0.77 - 1.24)
Respiratory arrest	0.03 (0.01 - 0.11)
Respiratory failure	0.69 (0.52 - 0.91)
Hepatic failure	0.10 (0.05 - 0.21)
Laparotomy	0.28 (0.18 - 0.43)
Amputation	0.01 (0.00 - 0.08)
Multiorgan failure	0.35 (0.24 - 0.52)
Deep sternal wound infection	0.79 (0.61 - 1.03)
Septicaemia	2.02 (1.71 - 2.38)
Sepsis	0.12 (0.06 - 0.23)

Appendix 6.2: Descriptive Statistics for Preoperative Prediction Data

		Severe Complication and AKI Prediction Population (2012-2018)					Delirium Population (2016-2018)		
		Total Population	Severe Complication = Yes		AKI = Yes		Total Population	Delirium = Yes	
		N = 6839	N = 404 (5.91%)		N=1295 (18.93%)		N=3344	N=417 (12.47%)	
Variable	**Level**	**n (%)**	**n (%)**	**P-value**	**n (%)**	**P-value**	**n (%)**	**n (%)**	**P-value**
Age Group	16 to 60	1894 (27.7)	87 (21.5)	<0.0001	297 (22.9)	<0.0001	967 (28.9)	76 (18.2)	<0.0001
	61 to 67	1506 (22.0)	80 (19.8)		251 (19.4)		730 (21.8)	68 (16.3)	
	68 to 74	1769 (25.9)	95 (23.5)		332 (25.6)		861 (25.7)	110 (26.4)	
	75 to 99	1670 (24.4)	142 (35.1)		415 (32.0)		786 (23.5)	163 (39.1)	
Sex	Male	4965 (72.6)	261 (64.6)	0.0003	909 (70.2)	0.0339	2389 (71.4)	268 (64.3)	0.0007
	Female	1874 (27.4)	143 (35.4)		386 (29.8)		955 (28.6)	149 (35.7)	
BMI	18.5-25.0	1255 (18.4)	86 (21.3)	0.0740	215 (16.6)	<0.0001	701 (21.0)	91 (21.8)	0.7988
	25.1-30.0	2782 (40.7)	173 (42.8)		620 (47.9)		1288 (38.5)	163 (39.1)	
	Over 30.0	2802 (41.0)	145 (35.9)		460 (35.5)		1355 (40.5)	163 (39.1)	
Type II Diabetes	No	5080 (74.3)	269 (66.6)	0.0003	894 (69.0)	<0.0001	2503 (74.9)	298 (71.5)	0.1002
	Yes	1759 (25.7)	135 (33.4)		401 (31.0)		841 (25.1)	119 (28.5)	
Smoking Status	Never smoked	1932 (28.2)	119 (29.5)	0.8991	377 (29.1)	0.0926	1137 (34.0)	142 (34.1)	0.8382
	Ex-smoker	2191 (32.0)	129 (31.9)		436 (33.7)		1146 (34.3)	139 (33.3)	
	Current smoker	943 (13.8)	57 (14.1)		153 (11.8)		532 (15.9)	64 (15.3)	
	Unknown	1773 (25.9)	99 (24.5)		329 (25.4)		529 (15.8)	72 (17.3)	
Procedure	CABG	3849 (56.3)	157 (38.9)	<0.0001	559 (43.2)	<0.0001	1738 (52.0)	140 (33.6)	<0.0001
	Valve	2010 (29.4)	161 (39.9)		450 (34.7)		1108 (33.1)	153 (36.7)	
	CABG and Valve	980 (14.3)	86 (21.3)		286 (22.1)		498 (14.9)	124 (29.7)	
Priority	Elective	4372 (63.9)	229 (56.7)	<0.0001	791 (61.1)	<0.0001	1623 (48.5)	193 (46.3)	0.0001
	Emergency	73 (1.1)	23 (5.7)		31 (2.4)		52 (1.6)	17 (4.1)	
	Priority	1132 (16.6)	53 (13.1)		194 (15.0)		895 (26.8)	106 (25.4)	
	Urgent	1262 (18.5)	99 (24.5)		279 (21.5)		774 (23.1)	101 (24.2)	
Critical Pre-op. State	No	6706 (98.1)	379 (93.8)	<0.0001	1246 (96.2)	<0.0001	3255 (97.3)	392 (94.0)	<0.0001

		Severe Complication and AKI Prediction Population (2012-2018)					**Delirium Population (2016-2018)**		
		Total Population **N = 6839**	**Severe Complication = Yes** **N = 404 (5.91%)**		**AKI = Yes** **N=1295 (18.93%)**		**Total Population** **N=3344**	**Delirium = Yes** **N=417 (12.47%)**	
Variable	**Level**	**n (%)**	**n (%)**	**P-value**	**n (%)**	**P-value**	**n (%)**	**n (%)**	**P-value**
	Yes	133 (1.9)	25 (6.2)		49 (3.8)		89 (2.7)	25 (6.0)	
Previous Cardiac Surgery	No	6640 (97.1)	363 (89.9)	<0.0001	1219 (94.1)	<0.0001	3228 (96.5)	398 (95.4)	0.2485
	Yes	199 (2.9)	41 (10.1)		76 (5.9)		116 (3.5)	19 (4.6)	
Previous Percutaneous Coronary Intervention	No	5920 (86.6)	330 (81.7)	0.0039	1122 (86.6)	0.9627	2883 (86.2)	368 (88.2)	0.2253
	Yes	919 (13.4)	74 (18.3)		173 (13.4)		461 (13.8)	49 (11.8)	
Extracardiac Arteriopathy	No	6024 (88.1)	338 (83.7)	0.0060	1093 (84.4)	<0.0001	3002 (89.8)	372 (89.2)	0.7490
	Yes	815 (11.9)	66 (16.3)		202 (15.6)		342 (10.2)	45 (10.8)	
Left Ventricular Function	Good	5438 (79.5)	291 (72.0)	<0.0001	986 (76.1)	0.0007	2562 (76.6)	316 (75.8)	0.2112
	Moderate	1203 (17.6)	88 (21.8)		256 (19.8)		670 (20.0)	81 (19.4)	
	Poor	198 (2.9)	25 (6.2)		53 (4.1)		112 (3.3)	20 (4.8)	
NYHA Grade	I	1358 (19.9)	57 (14.1)	<0.0001	224 (17.3)	<0.0001	717 (21.4)	60 (14.4)	<0.0001
	II	3416 (49.9)	151 (37.4)		573 (44.2)		1599 (47.8)	178 (42.7)	
	III	1821 (26.6)	150 (37.1)		411 (31.7)		870 (26.0)	140 (33.6)	
	IV	244 (3.6)	46 (11.4)		87 (6.7)		158 (4.7)	39 (9.4)	
Angina Status	0	2200 (32.2)	167 (41.3)	<0.0001	497 (38.4)	<0.0001	1271 (38.0)	179 (42.9)	0.1452
	I	937 (13.7)	32 (7.9)		176 (13.6)		442 (13.2)	46 (11.0)	
	II	2299 (33.6)	110 (27.2)		384 (29.7)		970 (29.0)	108 (25.9)	
	III	1044 (15.3)	57 (14.1)		164 (12.7)		469 (14.0)	57 (13.7)	
	IV	359 (5.2)	38 (9.4)		74 (5.7)		192 (5.7)	27 (6.5)	
Rhythm	Normal	5651 (82.6)	290 (71.8)	<0.0001	1016 (78.5)	<0.0001	2707 (81.0)	314 (75.3)	0.0029
	Abnormal	789 (11.5)	85 (21.0)		208 (16.1)		490 (14.7)	84 (20.1)	
	Unknown	399 (5.8)	29 (7.2)		71 (5.5)		147 (14.4)	19 (4.6)	
Renal Function Before Surgery	Normal	2825 (41.3)	125 (30.9)	<0.0001	475 (36.7)	<0.0001	1776 (53.1)	156 (37.4)	<0.0001
	Moderately Impaired	1916 (28.0)	135 (33.4)		378 (29.2)		1252 (37.4)	177 (42.4)	
	Severely Impaired	480 (7.0)	62 (15.3)		159 (12.3)		316 (9.4)	84 (20.1)	

		Severe Complication and AKI Prediction Population (2012-2018)					**Delirium Population (2016-2018)**		
		Total Population N = 6839	**Severe Complication = Yes N = 404 (5.91%)**		**AKI = Yes N=1295 (18.93%)**		**Total Population N=3344**	**Delirium = Yes N=417 (12.47%)**	
Variable	**Level**	**n (%)**	**n (%)**	**P-value**	**n (%)**	**P-value**	**n (%)**	**n (%)**	**P-value**
	Unknown	1618 (23.7)	82 (20.3)		283 (21.9)		0 (0.0)	0 (0.0)	
Preoperative Creatinine	<100	5252 (76.8)	243 (60.1)	<0.0001	864 (66.7)	<0.0001	2572 (76.9)	278 (66.7)	<0.0001
	100 or higher	1587 (23.2)	161 (39.9)		431 (33.3)		772 (23.1)	139 (33.3)	
Neurological Dysfunction	No	6746 (98.6)	398 (98.5)	0.9978	1276 (98.5)	0.8125	3330 (99.6)	416 (99.8)	0.8421
	Yes	93 (1.4)	6 (1.5)		19 (1.5)		14 (0.4)	1 (0.2)	
Previous Myocardial Infarction	No	4317 (63.1)	246 (60.9)	0.3652	834 (64.4)	0.3044	2147 (64.2)	279 (66.9)	0.2398
	Yes	2522 (36.9)	158 (39.1)		461 (35.6)		1197 (35.8)	138 (33.1)	
Left Main Stem Disease	No	3210 (46.9)	202 (50.0)	0.4265	643 (49.7)	0.0911	1797 (53.7)	248 (59.5)	0.0348
	Yes	955 (14.0)	55 (13.6)		169 (13.1)		484 (14.5)	49 (11.8)	
	Unknown	2674 (39.1)	147 (36.4)		483 (37.3)		1063 (31.8)	120 (28.8)	
Pulmonary Disease	No	5693 (83.2)	329 (81.4)	0.3502	1046 (80.8)	0.0092	2843 (85.0)	349 (83.7)	0.4612
	Yes	1146 (16.8)	75 (18.6)		249 (19.2)		501 (15.0)	68 (16.3)	
Hypertension History	No	1864 (27.3)	95 (23.5)	0.0924	288 (22.2)	<0.0001	923 (27.6)	111 (26.6)	0.6735
	Yes	4975 (72.7)	309 (76.5)		1007 (77.8)		2421 (72.4)	306 (73.4)	
Congestive Cardiac Failure	No	6135 (89.7)	305 (75.5)	<0.0001	1075 (83.0)	<0.0001	2907 (86.9)	333 (79.9)	<0.0001
	Yes	704 (10.3)	99 (24.5)		220 (17.0)		437 (13.1)	84 (20.1)	
Active Endocarditis	No	6761 (98.9)	384 (95.0)	<0.0001	1267 (97.8)	0.0002	3288 (98.3)	409 (98.1)	0.8331
	Yes	78 (1.1)	20 (5.0)		28 (2.2)		56 (1.7)	8 (1.9)	

Appendix 6.3: Training and Testing Datasets for Preoperative Models

	Severe complication and AKI experiments (2012-2018 data)			Delirium experiments (2016 to 2018 data)		
Demographic	**Train (n = 4583)**	**Test (n = 2256)**	**P-value**	**Train (n = 2241)**	**Test (n = 1103)**	**P-value**
Severe Complication	258 (5.63)	146 (6.47)	0.1821	-	-	-
Acute Kidney Injury	846 (18.46)	449 (19.90)	0.1618	-	-	-
Delirium	-	-	-	286 (12.76)	131 (11.88)	0.5009
Age			0.6737			0.6990
16 to 60	1261 (27.51)	633 (28.06)		652 (29.09)	315 (28.56)	
61 to 67	1019 (22.23)	487 (21.59)		476 (21.24)	254 (23.03)	
68 to 74	1199 (26.16)	570 (25.27)		580 (25.88)	281 (25.48)	
75 to 99	1104 (24.09)	566 (25.09)		533 (23.78)	253 (22.94)	
Sex			0.6730			0.1541
Female	1248 (27.23)	626 (27.75)		658 (70.64)	297 (26.93)	
Male	3335 (72.77)	1630 (72.25)		658 (29.36)	806 (73.07)	
BMI			0.9412			0.4821
18.5-25.0	836 (18.24)	419 (18.57)		466 (20.79)	235 (21.31)	
25.1-30.0	1865 (40.69)	917 (40.65)		879 (39.22)	409 (37.08)	
Over 30.1	1882 (41.06)	920 (40.78)		896 (39.98)	459 (41.61)	
Type II Diabetes			0.2573			0.3555
No	3424 (74.71)	1656 (73.40)		1666 (74.34)	837 (75.88)	
Yes	1159 (25.29)	600 (26.60)		575 (25.66)	266 (24.12)	
Smoking Status			0.1148			0.2220
Never smoked	1337 (29.17)	595 (26.37)		762 (34.00)	375 (34.00)	
Ex-smoker	1445 (31.53)	746 (33.07)		774 (34.54)	372 (33.73)	
Current smoker	627 (13.68)	316 (14.01)		338 (15.08)	194 (17.59)	
Unknown	1174 (25.62)	599 (26.55)		367 (16.38)	162 (14.69)	
Procedure			0.6513			0.0486
CABG	2564 (55.95)	1285 (56.96)		1140 (50.87)	598 (54.22)	
Valve	1363 (29.74)	647 (28.68)		774 (34.54)	334 (30.28)	
CABG and Valve	656 (14.31)	324 (14.36)		327 (14.59)	171 (15.50)	
Surgical Priority			0.4438			0.6969
Elective	2922 (63.76)	1450 (64.27)		1080 (48.19)	543 (49.23)	

	Severe complication and AKI experiments (2012-2018 data)			**Delirium experiments (2016 to 2018 data)**		
Demographic	**Train (n = 4583)**	**Test (n = 2256)**	**P-value**	**Train (n = 2241)**	**Test (n = 1103)**	**P-value**
Emergency	52 (1.13)	21 (0.93)		34 (1.52)	18 (1.63)	
Priority	777 (16.95)	355 (15.74)		614 (27.40)	281 (25.48)	
Urgent	832 (18.15)	430 (19.06)		513 (22.89)	261 (23.66)	
Critical Pre-op. State			0.9446			0.9738
No	4493 (98.04)	2213 (98.09)		2182 (97.37)	1073 (97.28)	
Yes	90 (1.96)	43 (1.91)		59 (2.63)	30 (2.72)	
Previous Cardiac Surgery			0.8610			0.9618
No	4448 (97.05)	2192 (97.16)		2164 (96.56)	1064 (96.46)	
Yes	135 (2.95)	64 (2.84)		77 (3.44)	39 (3.54)	
Previous PCI			0.8595			0.2175
No	3970 (86.62)	1950 (86.44)		1920 (85.68)	963 87.31)	
Yes	613 (13.38)	306 (13.56)		321 (14.32)	140 (12.69)	
Extracardiac Arteriopathy			0.5974			0.6881
No	4044 (88.24)	1980 (87.77)		2008 (89.60)	994 (90.12)	
Yes	539 (11.76)	276 (12.23)		233 (10.40)	109 (9.88)	
LV Function			0.5553			0.7011
Good	3636 (79.34)	1802 (79.88)		1721 (76.80)	841 (76.25)	
Moderate	819 (17.87)	384 (17.02)		442 (19.72)	228 (20.67)	
Poor	128 (2.79)	70 (3.10)		78 (3.48)	34 (3.08)	
NYHA Grade			0.2661			0.1152
I	940 (20.51)	418 (18.53%)		490 (21.87)	227 (20.58)	
II	2266 (49.44)	1150 (50.98)		1039 (46.36)	560 (50.77)	
III	1217 (26.55)	604 (26.77)		602 (26.86)	268 (24.30)	
IV	160 (3.49)	84 (3.72)		110 (4.91)	48 (4.35)	
Angina Status			0.1273			0.7680
0	1479 (32.27)	721 (31.96)		869 (38.78)	402 (36.45)	
I	650 (14.18)	287 (12.72)		291 (12.99)	151 (13.69)	
II	1520 (33.17)	779 (34.53)		641 (28.60)	329 (29.83)	
III	710 (15.49)	334 (14.80)		311 (13.88)	158 (14.32)	
IV	224 (4.89)	135 (5.98)		129 (5.76)	63 (5.71)	
Rhythm			0.6357			0.8676
Normal	3778 (82.44)	1873 (83.02)		1811 (80.81)	896 (81.23)	

	Severe complication and AKI experiments (2012-2018 data)			Delirium experiments (2016 to 2018 data)		
Demographic	**Train (n = 4583)**	**Test (n = 2256)**	**P-value**	**Train (n = 2241)**	**Test (n = 1103)**	**P-value**
Abnormal	529 (11.54)	260 (11.52)		333 (14.86)	157 (14.23)	
Unknown	276 (6.02)	123 (5.45)		97 (4.33)	50 (4.53)	
Renal Function Before Surgery			0.0081			0.5138
Normal	1942 (42.37)	883 (39.14)		1179 (52.61)	597 (54.13)	
Moderately Impaired	1247 (27.21)	669 (29.65)		854 (38.11)	398 (36.08)	
Severely Impaired	338 (7.38)	142 (6.29)		208 (9.28)	108 (9.79)	
Unknown	1056 (23.04)	562 (24.91)		0 (0)	0 (0)	
Preoperative Creatinine			0.3294			0.5493
<100 µmol/l	3503 (76.43)	1749 (77.53)		1731 (77.24)	841 (76.25)	
>= 100 µmol/l	1080 (23.57)	507 (22.47)		510 (22.76)	262 (23.75)	
Neurological Dysfunction			0.0824			0.5243
No	4529 (98.82)	2217 (98.27)		2230 (99.51)	1100 (99.73)	
Yes	54 (1.18)	39 (1.73)		11 (0.49)	3 (0.27)	
Previous Myocardial Infarction			0.8914			0.1750
No	2896 (63.19)	1421 (62.99)		1457 (65.02)	690 (62.56)	
Yes	1687 (36.81)	835 (37.01)		784 (34.98)	413 (37.44)	
Left Main Stem Disease			0.9778			0.8988
No	2147 (46.85)	1063 (47.12)		1201 (53.59)	596 (54.03)	
Yes	641 (13.99)	314 (13.92)		322 (14.37)	162 (14.69)	
Unknown	1795 (39.17)	879 (38.96)		718 (32.04)	345 (31.28)	
Pulmonary Disease			0.6561			0.8567
No	3822 (83.40)	1871 (82.93)		1903 (84.92)	940 (85.22)	
Yes	761 (16.60)	385 (17.07)		338 (15.08)	163 (14.78)	
Hypertension History			0.5022			0.6184
No	1237 (26.99)	627 (27.79)		612 (27.31)	311 (28.20)	
Yes	3346 (73.01)	1629 (72.21)		1629 (72.69)	792 (71.80)	
Congestive Cardiac Failure			0.9144			0.4686
No	4113 (89.74)	2022 (89.63)		1941 (86.61)	966 (87.58)	
Yes	470 (10.26)	234 (10.37)		300 (13.39)	137 (12.42)	
Active Endocarditis			0.7658			0.5720
No	4529 (98.82)	2232 (98.94)		2201 (98.22)	1087 (98.55)	
Yes	54 (1.18)	24 (1.06)		40 (1.78)	16 (1.45)	

Appendix 7.1: Descriptive Statistics when Predicting Acute Kidney Injury using Preoperative and ICU Data

Table 7.1.A Patient demographics and comparison between patients with and without AKI, based on chi-square test of independence for categorical variables and t-tests for numerical variables.

	Total Population	**AKI=No**	**AKI=Yes**	**AKI=Yes vs No**
	N=6294	**N=5544 (88.08%)**	**N=750 (11.92%)**	
Demographic	**Mean ± SD or %**	**Mean ± SD or %**	**Mean ± SD or %**	**P-value**
Age	66.09 ± 10.96	65.74 ± 10.92	69.52 ± 10.68	<0.0001
Sex				<0.0001
Female	27.66%	26.84%	33.73%	
Male	72.34%	73.16%	66.27%	
BMI				
18.5-25.0	18.94%	20.27%	18.76%	0.1507
25.1-30.0	39.26%	41.20%	39.00%	
Over 30.1	41.80%	38.53%	42.24%	
Type II Diabetes				0.0281
No	75.06%	71.73%	75.51%	
Yes	24.94%	28.27%	24.49%	
Smoking Status				
Never smoked	28.55%	32.27%	28.05%	0.0195
Ex-smoker	31.63%	31.47%	31.66%	
Current smoker	13.84%	10.80%	14.25%	
Unknown	25.98%	25.47%	26.05%	
Procedure				<0.0001
CABG	56.93%	59.34%	39.07%	
Valve	29.57%	28.14%	40.13%	
CABG and Valve	13.5%	12.52%	20.80%	
Logistic EuroSCORE	5.16 ± 5.91	4.76 ± 5.24	8.10 ± 8.96	<0.0001

Demographic	Total Population
	N=6294
	Mean ± SD or %
		AKI=No
		N=5544 (88.08%)
		Mean ± SD or %
			AKI=Yes
			N=750 (11.92%)
			Mean ± SD or %
				AKI=Yes vs No
				
				P-value

Renal Function Before Surgery				<0.0001
Normal	40.94%	42.39%	30.27%	
Moderately Impaired	28.28%	27.74%	32.27%	
Severely Impaired	7.10%	5.79%	16.80%	
Unknown	23.68%	24.08%	20.66%	
Preoperative Creatinine	91.18 ± 50.24	89.77 ± 50.88	101.58 ± 43.85	<0.0001
ICU Hours	44.05 ± 91.46	34.82 ± 57.24	112.26 ± 201.83	<0.0001
Total Days in Hospital	11.20 ± 8.78	10.54 ± 7.78	16.12 ± 13.09	<0.0001
Outcome				<0.0001
Alive	99.09%	99.57%	95.60%	
Dead	0.91%	0.43%	4.40%	

Table 7.1.B. Descriptive statistics for the total study population and for patients with and without AKI, where the two groups were compared with chi-square tests for categorical and t-tests for numerical variables.

Preoperative Variable	**Total Population** **N=6294** **Percentage**	**AKI=No** **N=5544 (88.08%)** **Percentage**	**AKI=Yes** **N=750 (11.92%)** **Percentage**	**AKI=Yes vs No** **P-value**
Priority				<0.0001
Elective	64.08%	60.27%	64.59%	
Emergency	1.02%	2.93%	0.76%	
Priority	16.59%	14.13%	16.92%	
Urgent	18.32%	22.67%	17.73%	
Critical Pre-op. State				0.0002
No	98.25%	96.53%	98.48%	
Yes	1.75%	3.47%	1.52%	
Previous Cardiac Surgery				
No	97.44%	94.93%	97.78%	<0.0001
Yes	2.56%	5.07	2.22%	
Previous Percutaneous Coronary Intervention				0.5221
No	86.65%	87.47%	86.54%	
Yes	13.35%	12.53%	13.46%	
Extracardiac Arteriopathy				0.0044
No	88.51%	54.33%	88.94%	
Yes	11.49%	14.67%	11.06	
Left Ventricular Function				0.0019
Good	79.74%	75.60%	80.30%	
Moderate	17.43%	20.00%	17.08%	
Poor	2.83%	4.40%	2.62%	
NYHA Grade				<0.0001
I	20.18%	18.13%	20.45%	
II	50.14%	41.73%	51.28%	
III	26.29%	32.67%	25.43%	
IV	3.38%	7.47%	2.83%	
Angina Status				<0.0001
0	32.14%	42.67%	30.72%	

Preoperative Variable	Total Population
	N=6294
	Percentage
		AKI=No
		N=5544 (88.08%)
		Percentage
			AKI=Yes
			N=750 (11.92%)
			Percentage
				AKI=Yes vs No
				
				P-value

I	13.68%	13.33%	13.73%	
II	33.81%	28.40%	34.54%	
III	15.25%	10.67%	15.87%	
IV	5.12%	4.93%	5.14%	
Rhythm				<0.0001
Normal	82.84%	77.20%	83.60%	
Abnormal	11.38%	18.00%	10.48%	
Unknown	5.78%	4.80%	5.92%	
Neurological Dysfunction				0.4772
No	98.62%	98.27%	98.67%	
Yes	1.38%	1.73%	1.33%	
Previous Myocardial Infarction				0.6071
No	62.95%	63.87%	62.82%	
Yes	37.05%	36.13%	37.18%	
Left Main Stem Disease				0.0023
No	47.11%	53.07%	46.30%	
Yes	13.98%	12.53%	14.18%	
Unknown	38.91%	34.40%	39.52%	
Pulmonary Disease				0.2258
No	83.60%	82.00%	83.82%	
Yes	16.40%	18.00%	16.18%	
Hypertension History				0.0040
No	27.82%	23.33%	28.43%	
Yes	72.18%	76.67%	71.57%	
Congestive Cardiac Failure				<0.0001
No	90.28%	82.93%	91.27%	
Yes	9.72%	17.07%	8.73%	
Active Endocarditis				0.0001
No	98.90%	97.47%	99.10%	
Yes	1.10%	2.53%	0.90%	

Table 7.1.C. Descriptive statistics for laboratory variables for the total population and for patients with and without AKI, where the two groups are compared, using chi-squared tests for categorical and t-tests for numerical variables.

Laboratory Variable (Unit)	**Total Population** **N=6294** **Mean ± SD**	**AKI=No** **N=5544 (88.08%)** **Mean ± SD**	**AKI=Yes** **N=750 (11.92%)** **Mean ± SD**	**P-value**
Arterial Base Excess (mmol/l)	-0.34 ± 3.29	-0.20 ± 2.85	-0.43 ± 3.79	<0.0001
Arterial Haematocrit (%)	28.82 ± 4.51	29.34 ± 4.65	27.71 ± 4.12	<0.0001
Bicarbonate (mEq/l)	24.20 ± 3.35	24.34 ± 2.98	24.08 ± 3.82	<0.0001
C-Reactive Protein (µmol/L)	137.60 ± 94.28	135.00 ± 88.34	140.40 ± 100.51	0.0009
Creatinine (µmol/L)	112.10 ± 76.73	89.54 ± 49.17	143.60 ± 92.58	<0.0001
Daily Fluid Balance	300.90 ± 265.00	322.60 ± 886.34	255.50 ± 965.38	<0.0001
Haemoglobin (g/l)	97.97 ± 15.31	99.64 ± 15.82	94.19 ± 13.65	<0.0001
Hydrogen Ion (mmol/l)	39.32 ± 6.68	39.12 ± 5.98	39.28 ± 6.71	<0.0001
Lactate (mmol/l)	1.60 ± 1.03	1.55 ± 0.86	1.60 ± 1.07	<0.0001
Potassium (mmol/l)	4.52 ± 0.68	4.52 ± 0.56	4.54 ± 0.66	<0.0001
Sodium (mmol/l)	135.40 ± 4.71	135.30 ± 4.11	135.20 ± 5.83	0.0062
Urea (mmol/l)	9.28 ± 6.10	6.78 ± 3.12	13.37 ± 7.35	<0.0001
Urine Output (l per day)	92.99 ± 87.90	97.50 ± 92.03	85.91 ± 80.48	<0.0001
Medicines				
Dobutamine (dose)	3.52 ± 5.76	3.04 ± 5.65	4.08 ± 5.91	<0.0001
N	1860 (29.55%)	1509 (27.22%)	351 (46.80%)	<0.0001
Dopamine (dose)	3.84 ± 6.80	3.58 ± 7.82	3.77 ± 6.35	0.1522
N	362 (5.75%)	274 (4.94%)	88 (11.73%)	<0.0001
Noradrenaline (dose)	4.05 ± 6.88	3.33 ± 5.06	4.73 ± 6.31	<0.0001
N	2491 (39.58%)	2087 (37.64%)	404 (53.87%)	<0.0001
Vasopressin (dose)	4.89 ± 2.63	4.62 ± 2.61	4.93 ± 2.69	<0.0001
N	146 (2.32%)	69 (1.24%)	77 (10.27%)	<0.0001

Appendix 7.2: Comparison of Laboratory Variables in Test Data Based on Imputation Methods

Table 7.2.A Mean and standard deviation (SD) for each laboratory variable in the test data when predicting AKI within 25 hours since ICU admission. The p-values are calculated based on t-test where the comparison is made in variable means based on the original data (without imputation) and the corresponding imputation method.

	Without imputation	**With median imputation**		**With 0 imputation**		**With missForest imputation**	
Variable	**mean ± SD**	**mean ± SD**	**P-value**	**mean ± SD**	**P-value**	**mean ± SD**	**P-value**
Arterial Base Excess							
Min	-3.78 ± 2.17	-3.78 ± 2.17	1.0000	-3.78 ± 2.17	1.0000	-3.78 ± 2.17	1.0000
Max	1.45 ± 2.07	1.45 ± 2.07	1.0000	1.45 ± 2.07	1.0000	1.45 ± 2.07	1.0000
First	0.23 ± 2.36	0.23 ± 2.36	1.0000	0.23 ± 2.36	1.0000	0.23 ± 2.36	1.0000
Last	-1.04 ± 2.00	-1.04 ± 2.00	1.0000	-1.04 ± 2.00	1.0000	-1.04 ± 2.00	1.0000
Arterial Haematocrit							
Min	24.76 ± 4.67	24.76 ± 4.67	0.9975	24.75 ± 4.72	0.9211	24.77 ± 4.67	0.9934
Max	37.51 ± 5.22	37.51 ± 5.22	0.9986	37.48 ± 5.32	0.8936	37.51 ± 5.22	0.9897
First	35.92 ± 6.70	35.92 ± 6.70	0.9976	35.90 ± 6.77	0.9202	35.92 ± 6.70	0.9895
Last	30.39 ± 4.47	30.39 ± 4.46	0.9987	30.37 ± 4.54	0.8990	30.39 ± 4.46	0.9978
Creatinine							
Min	89.80 ± 45.14	89.44 ± 43.75	0.8338	84.28 ± 48.76	0.0021	90.03 ± 43.90	0.8928
Max	103.33 ± 59.17	102.64 ± 57.39	0.7545	96.98 ± 62.47	0.0063	103.38 ± 57.57	0.9847
First	96.28 ± 48.60	95.90 ± 47.10	0.8328	90.36 ± 52.46	0.0022	96.46 ± 47.26	0.9228
Last	96.69 ± 55.90	96.03 ± 54.21	0.7549	90.74 ± 58.93	0.0067	96.78 ± 54.47	0.9642
C-Reactive Protein							
Min	62.85 ± 38.75	62.75 ± 38.47	0.9444	61.92 ± 39.20	0.5260	62.89 ± 38.57	0.9814
Max	157.08 ± 77.30	157.08 ± 76.73	0.9997	154.75 ± 79.05	0.4294	157.15 ± 76.87	0.9814
First	62.98 ± 39.34	62.88 ± 39.06	0.9441	62.05 ± 39.78	0.5313	63.02 ± 39.16	0.9797
Last	156.98 ± 77.13	156.98 ± 76.56	0.9999	154.65 ± 78.88	0.4287	157.05 ± 76.70	0.9801
Daily Fluid Balance							
Min	346.48 ± 785.89	337.21 ± 667.64	0.7599	250.00 ± 685.32	0.0016	400.08 ± 685.53	0.0801

	Without imputation	**With median imputation**		**With 0 imputation**		**With missForest imputation**	
Variable	**mean ± SD**	**mean ± SD**	**P-value**	**mean ± SD**	**P-value**	**mean ± SD**	**P-value**
Max	1266.41 ± 649.52	1273.02 ± 551.76	0.7920	913.79 ± 791.69	<0.0001	1213.84 ± 573.25	0.0386
First	1108.41 ± 749.59	1124.73 ± 637.19	0.5730	799.78 ± 807.67	<0.0001	1065.58 ± 656.12	0.1430
Last	513.58 ± 839.76	521.04 ± 713.33	0.8180	370.58 ± 749.48	<0.0001	553.24 ± 727.87	0.2244
Bicarbonate							
Min	20.77 ± 2.03	20.78 ± 1.98	0.9331	19.74 ± 4.92	<0.0001	20.79 ± 2.01	0.8446
Max	26.18 ± 2.11	26.17 ± 2.05	0.9626	24.88 ± 6.04	<0.0001	26.14 ± 2.08	0.6752
First	24.70 ± 2.43	24.70 ± 2.37	0.9563	23.48 ± 5.86	<0.0001	24.64 ± 2.41	0.5483
Last	23.66 ± 2.20	23.66 ± 2.14	0.9818	22.49 ± 5.56	<0.0001	23.65 ± 2.17	0.9203
Haemoglobin							
Min	83.82 ± 15.95	83.82 ± 15.95	1.0000	83.82 ± 15.95	1.0000	83.82 ± 15.95	1.0000
Max	128.18 ± 17.03	128.18 ± 17.03	1.0000	128.18 ± 17.03	1.0000	128.18 ± 17.03	1.0000
First	123.24 ± 22.41	123.24 ± 22.41	1.0000	123.24 ± 22.41	1.0000	123.24 ± 22.41	1.0000
Last	103.08 ± 14.04	103.08 ± 14.04	1.0000	103.08 ± 14.04	1.0000	103.08 ± 14.04	1.0000
Hydrogen Ion							
Min	34.10 ± 3.67	34.12 ± 3.58	0.8574	32.41 ± 8.22	<0.0001	34.11 ± 3.59	0.9431
Max	47.89 ± 16.34	47.82 ± 15.93	0.9109	45.52 ± 19.01	0.0004	47.85 ± 15.94	0.9482
First	38.55 ± 4.41	38.53 ± 4.30	0.8937	36.64 ± 9.40	<0.0001	38.53 ± 4.32	0.9161
Last	40.11 ± 4.13	40.09 ± 4.03	0.8956	38.13 ± 9.59	<0.0001	40.11 ± 4.04	0.9710
Lactate							
Min	1.08 ± 0.43	1.08 ± 0.42	0.8643	1.03 ± 0.47	0.0030	1.09 ± 0.42	0.6550
Max	2.68 ± 1.28	2.67 ± 1.26	0.7907	2.55 ± 1.38	0.0123	2.69 ± 1.26	0.8689
First	1.65 ± 0.90	1.64 ± 0.88	0.7585	1.57 ± 0.94	0.0258	1.66 ± 0.88	0.8447
Last	1.51 ± 0.64	1.50 ± 0.62	0.8617	1.44 ± 0.70	0.0050	1.51 ± 0.62	0.7951
Potassium							
Min	3.90 ± 0.35	3.90 ± 0.35	1.0000	3.90 ± 0.35	1.0000	3.90 ± 0.35	1.0000
Max	5.49 ± 0.64	5.49 ± 0.64	1.0000	5.49 ± 0.64	1.0000	5.49 ± 0.64	1.0000
First	4.18 ± 0.54	4.18 ± 0.54	1.0000	4.18 ± 0.54	1.0000	4.18 ± 0.54	1.0000
Last	4.62 ± 0.38	4.62 ± 0.38	1.0000	4.62 ± 0.38	1.0000	4.62 ± 0.38	1.0000
Sodium							
Min	133.31 ± 4.22	133.31 ± 4.22	1.0000	133.31 ± 4.22	1.0000	133.31 ± 4.22	1.0000
Max	139.86 ± 2.95	139.86 ± 2.95	1.0000	139.86 ± 2.95	1.0000	139.86 ± 2.95	1.0000
First	138.46 ± 4.57	138.46 ± 4.57	1.0000	138.46 ± 4.57	1.0000	138.46 ± 4.57	1.0000

Variable	Without imputation mean ± SD	With median imputation mean ± SD	P-value	With 0 imputation mean ± SD	P-value	With missForest imputation mean ± SD	P-value
Last	134.61 ± 2.91	134.61 ± 2.91	1.0000	134.61 ± 2.91	1.0000	134.61 ± 2.91	1.0000
Urea							
Min	5.79 ± 2.01	5.79 ± 2.01	0.9682	5.74 ± 2.07	0.5586	5.79 ± 2.01	0.9650
Max	6.90 ± 2.78	6.89 ± 2.77	0.9646	6.84 ± 2.83	0.6121	6.90 ± 2.77	0.9902
First	6.06 ± 2.02	6.06 ± 2.01	0.9706	6.01 ± 2.08	0.5416	6.06 ± 2.01	0.9661
Last	6.62 ± 2.84	6.61 ± 2.83	0.9642	6.57 ± 2.89	0.6333	6.62 ± 2.83	0.9920
Urine							
Min	24.28 ± 15.05	24.28 ± 15.05	1.0000	24.28 ± 15.05	1.0000	24.28 ± 15.05	1.0000
Max	306.25 ± 211.29	306.25 ± 211.29	1.0000	306.25 ± 211.29	1.0000	306.25 ± 211.29	1.0000
First	187.54 ± 131.77	187.54 ± 131.77	1.0000	187.54 ± 131.77	1.0000	187.54 ± 131.77	1.0000
Last	66.90 ± 61.28	66.90 ± 61.28	1.0000	66.90 ± 61.28	1.0000	66.90 ± 61.28	1.0000

Appendix 7.3: Performance Measures for Each Model Predicting AKI at Each Lead Time

Table 7.3.A. Performance measures for each model at each lead time when predicting AKI within 25h since ICU admission, using complete data.

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**AB**					
-24	0.770 (0.678 - 0.862)	0.802 (0.715 - 0.889)	0.652 (0.548 - 0.756)	0.022 (0.000 - 0.054)	0.856 (0.780 - 0.932)	0.146
-23	0.785 (0.694 - 0.876)	0.709 (0.609 - 0.809)	0.750 (0.655 - 0.845)	0.027 (0.000 - 0.063)	0.833 (0.751 - 0.915)	0.181
-22	0.812 (0.727 - 0.897)	0.691 (0.590 - 0.792)	0.782 (0.692 - 0.872)	0.028 (0.000 - 0.064)	0.813 (0.728 - 0.898)	0.197
-21	0.778 (0.691 - 0.865)	0.828 (0.749 - 0.907)	0.623 (0.521 - 0.725)	0.021 (0.000 - 0.051)	0.853 (0.779 - 0.927)	0.143
-20	0.770 (0.680 - 0.860)	0.643 (0.541 - 0.745)	0.756 (0.664 - 0.848)	0.034 (0.000 - 0.073)	0.834 (0.754 - 0.914)	0.197
-19	0.792 (0.709 - 0.875)	0.582 (0.481 - 0.683)	0.837 (0.761 - 0.913)	0.039 (0.000 - 0.079)	0.774 (0.688 - 0.860)	0.231
-18	0.793 (0.708 - 0.878)	0.761 (0.672 - 0.850)	0.686 (0.589 - 0.783)	0.027 (0.000 - 0.061)	0.839 (0.762 - 0.916)	0.171
-17	0.811 (0.730 - 0.892)	0.689 (0.593 - 0.785)	0.808 (0.727 - 0.889)	0.030 (0.000 - 0.065)	0.775 (0.689 - 0.861)	0.224
-16	0.773 (0.683 - 0.863)	0.759 (0.667 - 0.851)	0.703 (0.605 - 0.801)	0.025 (0.000 - 0.059)	0.841 (0.762 - 0.920)	0.177
-15	0.743 (0.655 - 0.831)	0.628 (0.530 - 0.726)	0.788 (0.705 - 0.871)	0.038 (0.000 - 0.077)	0.800 (0.719 - 0.881)	0.222
-14	0.743 (0.655 - 0.831)	0.840 (0.766 - 0.914)	0.576 (0.476 - 0.676)	0.023 (0.000 - 0.053)	0.857 (0.786 - 0.928)	0.143
-13	0.815 (0.739 - 0.891)	0.772 (0.690 - 0.854)	0.738 (0.652 - 0.824)	0.027 (0.000 - 0.059)	0.789 (0.709 - 0.869)	0.199
-12	0.792 (0.713 - 0.871)	0.743 (0.658 - 0.828)	0.714 (0.626 - 0.802)	0.032 (0.000 - 0.066)	0.810 (0.733 - 0.887)	0.177
-11	0.813 (0.738 - 0.888)	0.740 (0.656 - 0.824)	0.745 (0.661 - 0.829)	0.032 (0.000 - 0.066)	0.787 (0.708 - 0.866)	0.200
-10	0.832 (0.760 - 0.904)	0.670 (0.579 - 0.761)	0.860 (0.793 - 0.927)	0.034 (0.000 - 0.069)	0.693 (0.604 - 0.782)	0.275
-9	0.817 (0.741 - 0.893)	0.810 (0.733 - 0.887)	0.700 (0.610 - 0.790)	0.024 (0.000 - 0.054)	0.805 (0.727 - 0.883)	0.178
-8	0.834 (0.765 - 0.903)	0.804 (0.730 - 0.878)	0.766 (0.688 - 0.844)	0.025 (0.000 - 0.054)	0.743 (0.662 - 0.824)	0.220
-7	0.830 (0.760 - 0.900)	0.829 (0.759 - 0.899)	0.701 (0.616 - 0.786)	0.024 (0.000 - 0.052)	0.784 (0.707 - 0.861)	0.177
-6	0.869 (0.806 - 0.932)	0.775 (0.697 - 0.853)	0.831 (0.761 - 0.901)	0.026 (0.000 - 0.056)	0.686 (0.600 - 0.772)	0.250
-5	0.852 (0.787 - 0.917)	0.816 (0.745 - 0.887)	0.732 (0.651 - 0.813)	0.025 (0.000 - 0.054)	0.762 (0.684 - 0.840)	0.195
-4	0.844 (0.777 - 0.911)	0.842 (0.775 - 0.909)	0.738 (0.657 - 0.819)	0.021 (0.000 - 0.047)	0.753 (0.674 - 0.832)	0.182
-3	0.847 (0.781 - 0.913)	0.798 (0.724 - 0.872)	0.748 (0.668 - 0.828)	0.027 (0.000 - 0.057)	0.755 (0.676 - 0.834)	0.199
-2	0.838 (0.771 - 0.905)	0.713 (0.630 - 0.796)	0.835 (0.767 - 0.903)	0.034 (0.001 - 0.067)	0.691 (0.607 - 0.775)	0.230

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-1	0.888 (0.827 - 0.949)	0.806 (0.730 - 0.882)	0.802 (0.725 - 0.879)	0.022 (0.000 - 0.050)	0.729 (0.643 - 0.815)	0.229
Mean ± SD	**0.810 ± 0.037**	**0.752 ± 0.072**	**0.745 ± 0.069**	**0.028 ± 0.005**	**0.786 ± 0.052**	
Model	**BARTm**					
-24	0.832 (0.751 - 0.913)	0.914 (0.853 - 0.975)	0.601 (0.494 - 0.708)	0.010 (0.000 - 0.032)	0.857 (0.781 - 0.933)	0.036
-23	0.843 (0.763 - 0.923)	0.886 (0.816 - 0.956)	0.651 (0.546 - 0.756)	0.012 (0.000 - 0.036)	0.847 (0.768 - 0.926)	0.043
-22	0.841 (0.761 - 0.921)	0.840 (0.760 - 0.920)	0.698 (0.598 - 0.798)	0.016 (0.000 - 0.043)	0.832 (0.751 - 0.913)	0.063
-21	0.817 (0.736 - 0.898)	0.793 (0.708 - 0.878)	0.719 (0.625 - 0.813)	0.022 (0.000 - 0.053)	0.819 (0.738 - 0.900)	0.054
-20	0.831 (0.751 - 0.911)	0.857 (0.782 - 0.932)	0.693 (0.594 - 0.792)	0.015 (0.000 - 0.041)	0.826 (0.745 - 0.907)	0.059
-19	0.837 (0.761 - 0.913)	0.813 (0.733 - 0.893)	0.729 (0.638 - 0.820)	0.021 (0.000 - 0.050)	0.803 (0.721 - 0.885)	0.066
-18	0.843 (0.767 - 0.919)	0.795 (0.711 - 0.879)	0.739 (0.647 - 0.831)	0.021 (0.000 - 0.051)	0.806 (0.723 - 0.889)	0.069
-17	0.856 (0.783 - 0.929)	0.789 (0.705 - 0.873)	0.749 (0.659 - 0.839)	0.022 (0.000 - 0.052)	0.797 (0.714 - 0.880)	0.077
-16	0.837 (0.758 - 0.916)	0.855 (0.779 - 0.931)	0.693 (0.594 - 0.792)	0.015 (0.000 - 0.041)	0.829 (0.748 - 0.910)	0.061
-15	0.814 (0.735 - 0.893)	0.723 (0.633 - 0.813)	0.753 (0.666 - 0.840)	0.030 (0.000 - 0.064)	0.802 (0.721 - 0.883)	0.084
-14	0.830 (0.754 - 0.906)	0.840 (0.766 - 0.914)	0.700 (0.607 - 0.793)	0.019 (0.000 - 0.047)	0.809 (0.730 - 0.888)	0.063
-13	0.832 (0.759 - 0.905)	0.743 (0.658 - 0.828)	0.769 (0.687 - 0.851)	0.030 (0.000 - 0.063)	0.774 (0.692 - 0.856)	0.092
-12	0.823 (0.749 - 0.897)	0.832 (0.759 - 0.905)	0.686 (0.595 - 0.777)	0.022 (0.000 - 0.051)	0.806 (0.729 - 0.883)	0.053
-11	0.837 (0.766 - 0.908)	0.837 (0.766 - 0.908)	0.723 (0.637 - 0.809)	0.021 (0.000 - 0.049)	0.780 (0.700 - 0.860)	0.069
-10	0.848 (0.779 - 0.917)	0.767 (0.685 - 0.849)	0.788 (0.709 - 0.867)	0.027 (0.000 - 0.058)	0.750 (0.666 - 0.834)	0.105
-9	0.838 (0.766 - 0.910)	0.730 (0.643 - 0.817)	0.801 (0.723 - 0.879)	0.029 (0.000 - 0.062)	0.754 (0.670 - 0.838)	0.110
-8	0.851 (0.785 - 0.917)	0.812 (0.740 - 0.884)	0.756 (0.676 - 0.836)	0.024 (0.000 - 0.052)	0.749 (0.669 - 0.829)	0.084
-7	0.876 (0.815 - 0.937)	0.811 (0.738 - 0.884)	0.789 (0.713 - 0.865)	0.023 (0.000 - 0.051)	0.723 (0.640 - 0.806)	0.101
-6	0.871 (0.809 - 0.933)	0.865 (0.801 - 0.929)	0.726 (0.643 - 0.809)	0.018 (0.000 - 0.043)	0.761 (0.682 - 0.840)	0.081
-5	0.887 (0.829 - 0.945)	0.798 (0.724 - 0.872)	0.837 (0.769 - 0.905)	0.024 (0.000 - 0.052)	0.665 (0.578 - 0.752)	0.121
-4	0.887 (0.829 - 0.945)	0.868 (0.806 - 0.930)	0.765 (0.687 - 0.843)	0.017 (0.000 - 0.041)	0.726 (0.644 - 0.808)	0.088
-3	0.881 (0.822 - 0.940)	0.816 (0.745 - 0.887)	0.814 (0.743 - 0.885)	0.023 (0.000 - 0.051)	0.690 (0.605 - 0.775)	0.104
-2	0.876 (0.816 - 0.936)	0.791 (0.717 - 0.865)	0.833 (0.765 - 0.901)	0.025 (0.000 - 0.054)	0.671 (0.585 - 0.757)	0.115
-1	0.918 (0.865 - 0.971)	0.932 (0.883 - 0.981)	0.764 (0.682 - 0.846)	0.008 (0.000 - 0.025)	0.734 (0.649 - 0.819)	0.082
Mean ± SD	**0.850 ± 0.026**	**0.821 ± 0.053**	**0.741 ± 0.057**	**0.021 ± 0.006**	**0.775 ± 0.054**	
Model	**C5.0**					
-23	0.763 (0.669 - 0.857)	0.734 (0.637 - 0.831)	0.747 (0.651 - 0.843)	0.025 (0.000 - 0.059)	0.829 (0.746 - 0.912)	0.087
-22	0.763 (0.670 - 0.856)	0.716 (0.618 - 0.814)	0.702 (0.602 - 0.802)	0.029 (0.000 - 0.066)	0.852 (0.775 - 0.929)	0.083
-21	0.748 (0.657 - 0.839)	0.713 (0.618 - 0.808)	0.692 (0.595 - 0.789)	0.032 (0.000 - 0.069)	0.847 (0.771 - 0.923)	0.077
-20	0.806 (0.721 - 0.891)	0.774 (0.685 - 0.863)	0.753 (0.661 - 0.845)	0.022 (0.000 - 0.053)	0.809 (0.725 - 0.893)	0.133

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-19	0.754 (0.666 - 0.842)	0.703 (0.609 - 0.797)	0.762 (0.675 - 0.849)	0.031 (0.000 - 0.067)	0.805 (0.724 - 0.886)	0.136
-18	0.787 (0.701 - 0.873)	0.625 (0.524 - 0.726)	0.856 (0.783 - 0.929)	0.033 (0.000 - 0.070)	0.744 (0.653 - 0.835)	0.169
-17	0.780 (0.694 - 0.866)	0.789 (0.705 - 0.873)	0.706 (0.612 - 0.800)	0.024 (0.000 - 0.056)	0.822 (0.743 - 0.901)	0.078
-16	0.750 (0.657 - 0.843)	0.699 (0.600 - 0.798)	0.720 (0.623 - 0.817)	0.030 (0.000 - 0.067)	0.844 (0.766 - 0.922)	0.086
-15	0.729 (0.639 - 0.819)	0.745 (0.657 - 0.833)	0.646 (0.549 - 0.743)	0.032 (0.000 - 0.068)	0.849 (0.777 - 0.921)	0.078
-14	0.773 (0.688 - 0.858)	0.777 (0.693 - 0.861)	0.668 (0.573 - 0.763)	0.027 (0.000 - 0.060)	0.835 (0.760 - 0.910)	0.079
-13	0.798 (0.720 - 0.876)	0.822 (0.747 - 0.897)	0.685 (0.594 - 0.776)	0.023 (0.000 - 0.052)	0.808 (0.731 - 0.885)	0.083
-12	0.767 (0.685 - 0.849)	0.683 (0.592 - 0.774)	0.801 (0.723 - 0.879)	0.035 (0.000 - 0.071)	0.763 (0.680 - 0.846)	0.150
-11	0.769 (0.688 - 0.850)	0.798 (0.721 - 0.875)	0.634 (0.541 - 0.727)	0.029 (0.000 - 0.061)	0.831 (0.759 - 0.903)	0.083
-10	0.763 (0.681 - 0.845)	0.728 (0.642 - 0.814)	0.691 (0.602 - 0.780)	0.035 (0.000 - 0.070)	0.822 (0.748 - 0.896)	0.086
-9	0.786 (0.706 - 0.866)	0.740 (0.654 - 0.826)	0.739 (0.653 - 0.825)	0.030 (0.000 - 0.063)	0.798 (0.719 - 0.877)	0.088
-8	0.806 (0.733 - 0.879)	0.768 (0.690 - 0.846)	0.725 (0.642 - 0.808)	0.031 (0.000 - 0.063)	0.781 (0.704 - 0.858)	0.142
-7	0.800 (0.726 - 0.874)	0.766 (0.687 - 0.845)	0.719 (0.635 - 0.803)	0.031 (0.000 - 0.063)	0.786 (0.710 - 0.862)	0.097
-6	0.853 (0.787 - 0.919)	0.856 (0.791 - 0.921)	0.732 (0.650 - 0.814)	0.019 (0.000 - 0.044)	0.759 (0.679 - 0.839)	0.148
-5	0.792 (0.717 - 0.867)	0.667 (0.580 - 0.754)	0.792 (0.717 - 0.867)	0.041 (0.005 - 0.077)	0.752 (0.673 - 0.831)	0.159
-4	0.817 (0.746 - 0.888)	0.719 (0.636 - 0.802)	0.827 (0.758 - 0.896)	0.034 (0.001 - 0.067)	0.702 (0.618 - 0.786)	0.172
-3	0.809 (0.737 - 0.881)	0.746 (0.666 - 0.826)	0.765 (0.687 - 0.843)	0.033 (0.000 - 0.066)	0.754 (0.675 - 0.833)	0.161
-2	0.807 (0.735 - 0.879)	0.722 (0.640 - 0.804)	0.776 (0.700 - 0.852)	0.036 (0.002 - 0.070)	0.750 (0.671 - 0.829)	0.158
-1	0.874 (0.810 - 0.938)	0.786 (0.707 - 0.865)	0.829 (0.756 - 0.902)	0.023 (0.000 - 0.052)	0.703 (0.615 - 0.791)	0.162
Mean ± SD	**0.787 ± 0.034**	**0.742 ± 0.052**	**0.738 ± 0.058**	**0.030 ± 0.005**	**0.793 ± 0.045**	
Model	**GBM**					
-24	0.816 (0.732 - 0.900)	0.877 (0.805 - 0.949)	0.602 (0.495 - 0.709)	0.015 (0.000 - 0.041)	0.861 (0.786 - 0.936)	0.072
-23	0.813 (0.727 - 0.899)	0.722 (0.623 - 0.821)	0.789 (0.699 - 0.879)	0.024 (0.000 - 0.058)	0.805 (0.718 - 0.892)	0.097
-22	0.841 (0.761 - 0.921)	0.765 (0.673 - 0.857)	0.777 (0.686 - 0.868)	0.021 (0.000 - 0.052)	0.801 (0.714 - 0.888)	0.086
-21	0.826 (0.746 - 0.906)	0.862 (0.790 - 0.934)	0.656 (0.556 - 0.756)	0.016 (0.000 - 0.042)	0.836 (0.758 - 0.914)	0.039
-20	0.800 (0.714 - 0.886)	0.738 (0.644 - 0.832)	0.747 (0.654 - 0.840)	0.026 (0.000 - 0.060)	0.820 (0.738 - 0.902)	0.082
-19	0.795 (0.712 - 0.878)	0.637 (0.538 - 0.736)	0.806 (0.725 - 0.887)	0.036 (0.000 - 0.074)	0.788 (0.704 - 0.872)	0.088
-18	0.820 (0.740 - 0.900)	0.807 (0.725 - 0.889)	0.733 (0.641 - 0.825)	0.020 (0.000 - 0.049)	0.807 (0.725 - 0.889)	0.071
-17	0.848 (0.774 - 0.922)	0.767 (0.680 - 0.854)	0.772 (0.685 - 0.859)	0.024 (0.000 - 0.056)	0.786 (0.701 - 0.871)	0.108
-16	0.819 (0.736 - 0.902)	0.735 (0.640 - 0.830)	0.777 (0.687 - 0.867)	0.025 (0.000 - 0.059)	0.804 (0.719 - 0.889)	0.090
-15	0.787 (0.704 - 0.870)	0.734 (0.645 - 0.823)	0.754 (0.667 - 0.841)	0.029 (0.000 - 0.063)	0.799 (0.718 - 0.880)	0.081
-14	0.804 (0.724 - 0.884)	0.851 (0.779 - 0.923)	0.625 (0.527 - 0.723)	0.020 (0.000 - 0.048)	0.839 (0.765 - 0.913)	0.073
-13	0.831 (0.758 - 0.904)	0.851 (0.782 - 0.920)	0.653 (0.560 - 0.746)	0.020 (0.000 - 0.047)	0.818 (0.743 - 0.893)	0.064

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-12	0.801 (0.723 - 0.879)	0.842 (0.771 - 0.913)	0.621 (0.526 - 0.716)	0.023 (0.000 - 0.052)	0.832 (0.759 - 0.905)	0.061
-11	0.832 (0.760 - 0.904)	0.721 (0.635 - 0.807)	0.797 (0.720 - 0.874)	0.032 (0.000 - 0.066)	0.751 (0.668 - 0.834)	0.110
-10	0.861 (0.794 - 0.928)	0.825 (0.752 - 0.898)	0.736 (0.651 - 0.821)	0.021 (0.000 - 0.049)	0.776 (0.695 - 0.857)	0.063
-9	0.832 (0.759 - 0.905)	0.650 (0.557 - 0.743)	0.857 (0.788 - 0.926)	0.035 (0.000 - 0.071)	0.712 (0.623 - 0.801)	0.156
-8	0.842 (0.774 - 0.910)	0.777 (0.700 - 0.854)	0.758 (0.679 - 0.837)	0.029 (0.000 - 0.060)	0.756 (0.676 - 0.836)	0.084
-7	0.867 (0.804 - 0.930)	0.856 (0.791 - 0.921)	0.787 (0.711 - 0.863)	0.018 (0.000 - 0.043)	0.714 (0.630 - 0.798)	0.102
-6	0.874 (0.812 - 0.936)	0.802 (0.728 - 0.876)	0.834 (0.765 - 0.903)	0.023 (0.000 - 0.051)	0.675 (0.588 - 0.762)	0.133
-5	0.879 (0.819 - 0.939)	0.798 (0.724 - 0.872)	0.804 (0.731 - 0.877)	0.025 (0.000 - 0.054)	0.706 (0.622 - 0.790)	0.105
-4	0.871 (0.809 - 0.933)	0.772 (0.695 - 0.849)	0.822 (0.752 - 0.892)	0.028 (0.000 - 0.058)	0.692 (0.607 - 0.777)	0.107
-3	0.873 (0.812 - 0.934)	0.754 (0.675 - 0.833)	0.841 (0.774 - 0.908)	0.029 (0.000 - 0.060)	0.673 (0.587 - 0.759)	0.136
-2	0.867 (0.805 - 0.929)	0.817 (0.746 - 0.888)	0.788 (0.713 - 0.863)	0.023 (0.000 - 0.050)	0.715 (0.632 - 0.798)	0.107
-1	0.906 (0.850 - 0.962)	0.893 (0.833 - 0.953)	0.788 (0.709 - 0.867)	0.012 (0.000 - 0.033)	0.721 (0.634 - 0.808)	0.104
Mean ± SD	**0.838 ± 0.031**	**0.786 ± 0.067**	**0.755 ± 0.072**	**0.024 ± 0.006**	**0.770 ± 0.056**	
Model	**LR**					
-24	0.838 (0.758 - 0.918)	0.802 (0.715 - 0.889)	0.769 (0.677 - 0.861)	0.018 (0.000 - 0.047)	0.798 (0.711 - 0.885)	0.057
-23	0.830 (0.747 - 0.913)	0.759 (0.665 - 0.853)	0.789 (0.699 - 0.879)	0.021 (0.000 - 0.053)	0.797 (0.708 - 0.886)	0.064
-22	0.839 (0.759 - 0.919)	0.728 (0.631 - 0.825)	0.823 (0.740 - 0.906)	0.023 (0.000 - 0.056)	0.770 (0.678 - 0.862)	0.080
-21	0.552 (0.448 - 0.656)	0.138 (0.066 - 0.210)	0.966 (0.928 - 1.000)	0.065 (0.013 - 0.117)	0.760 (0.670 - 0.850)	0.001
-20	0.529 (0.422 - 0.636)	0.083 (0.024 - 0.142)	0.975 (0.942 - 1.000)	0.200 (0.114 - 0.286)	0.934 (0.881 - 0.987)	0.001
-19	0.585 (0.484 - 0.686)	0.209 (0.125 - 0.293)	0.961 (0.921 - 1.000)	0.063 (0.013 - 0.113)	0.694 (0.599 - 0.789)	0.001
-18	0.796 (0.712 - 0.880)	0.648 (0.548 - 0.748)	0.801 (0.718 - 0.884)	0.034 (0.000 - 0.072)	0.795 (0.711 - 0.879)	0.076
-17	0.824 (0.745 - 0.903)	0.556 (0.453 - 0.659)	0.940 (0.891 - 0.989)	0.037 (0.000 - 0.076)	0.573 (0.471 - 0.675)	0.190
-16	0.802 (0.716 - 0.888)	0.807 (0.722 - 0.892)	0.654 (0.552 - 0.756)	0.021 (0.000 - 0.052)	0.852 (0.776 - 0.928)	0.037
-15	0.800 (0.719 - 0.881)	0.766 (0.680 - 0.852)	0.710 (0.618 - 0.802)	0.027 (0.000 - 0.060)	0.818 (0.740 - 0.896)	0.050
-14	0.810 (0.731 - 0.889)	0.670 (0.575 - 0.765)	0.821 (0.744 - 0.898)	0.033 (0.000 - 0.069)	0.760 (0.674 - 0.846)	0.094
-13	0.827 (0.753 - 0.901)	0.663 (0.571 - 0.755)	0.872 (0.807 - 0.937)	0.034 (0.000 - 0.069)	0.679 (0.588 - 0.770)	0.125
-12	0.761 (0.678 - 0.844)	0.624 (0.530 - 0.718)	0.765 (0.682 - 0.848)	0.043 (0.003 - 0.083)	0.806 (0.729 - 0.883)	0.055
-11	0.840 (0.770 - 0.910)	0.817 (0.743 - 0.891)	0.708 (0.621 - 0.795)	0.024 (0.000 - 0.053)	0.793 (0.715 - 0.871)	0.051
-10	0.856 (0.788 - 0.924)	0.786 (0.707 - 0.865)	0.792 (0.714 - 0.870)	0.024 (0.000 - 0.054)	0.741 (0.656 - 0.826)	0.077
-9	0.857 (0.788 - 0.926)	0.720 (0.632 - 0.808)	0.838 (0.766 - 0.910)	0.029 (0.000 - 0.062)	0.717 (0.629 - 0.805)	0.107
-8	0.841 (0.773 - 0.909)	0.679 (0.593 - 0.765)	0.863 (0.799 - 0.927)	0.036 (0.001 - 0.071)	0.667 (0.580 - 0.754)	0.125
-7	0.867 (0.804 - 0.930)	0.793 (0.718 - 0.868)	0.813 (0.740 - 0.886)	0.025 (0.000 - 0.054)	0.704 (0.619 - 0.789)	0.093
-6	0.887 (0.828 - 0.946)	0.766 (0.687 - 0.845)	0.854 (0.788 - 0.920)	0.027 (0.000 - 0.057)	0.657 (0.569 - 0.745)	0.120

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-5	0.855 (0.790 - 0.920)	0.781 (0.705 - 0.857)	0.814 (0.743 - 0.885)	0.027 (0.000 - 0.057)	0.699 (0.615 - 0.783)	0.089
-4	0.858 (0.794 - 0.922)	0.798 (0.724 - 0.872)	0.793 (0.719 - 0.867)	0.025 (0.000 - 0.054)	0.717 (0.634 - 0.800)	0.080
-3	0.862 (0.799 - 0.925)	0.860 (0.796 - 0.924)	0.789 (0.714 - 0.864)	0.018 (0.000 - 0.042)	0.706 (0.622 - 0.790)	0.075
-2	0.847 (0.781 - 0.913)	0.757 (0.679 - 0.835)	0.822 (0.752 - 0.892)	0.030 (0.000 - 0.061)	0.695 (0.611 - 0.779)	0.085
-1	0.896 (0.837 - 0.955)	0.825 (0.752 - 0.898)	0.841 (0.770 - 0.912)	0.019 (0.000 - 0.045)	0.678 (0.588 - 0.768)	0.078
Mean ± SD	**0.802 ± 0.100**	**0.668 ± 0.216**	**0.824 ± 0.080**	**0.038 ± 0.037**	**0.742 ± 0.076**	
Model	**RF**					
-24	0.750 (0.656 - 0.844)	0.630 (0.525 - 0.735)	0.777 (0.686 - 0.868)	0.034 (0.000 - 0.073)	0.829 (0.747 - 0.911)	0.080
-23	0.770 (0.677 - 0.863)	0.734 (0.637 - 0.831)	0.718 (0.619 - 0.817)	0.026 (0.000 - 0.061)	0.845 (0.765 - 0.925)	0.070
-22	0.811 (0.726 - 0.896)	0.728 (0.631 - 0.825)	0.739 (0.643 - 0.835)	0.026 (0.000 - 0.061)	0.831 (0.749 - 0.913)	0.080
-21	0.780 (0.693 - 0.867)	0.690 (0.593 - 0.787)	0.737 (0.644 - 0.830)	0.032 (0.000 - 0.069)	0.830 (0.751 - 0.909)	0.080
-20	0.776 (0.687 - 0.865)	0.786 (0.698 - 0.874)	0.660 (0.559 - 0.761)	0.024 (0.000 - 0.057)	0.852 (0.776 - 0.928)	0.065
-19	0.784 (0.699 - 0.869)	0.615 (0.515 - 0.715)	0.806 (0.725 - 0.887)	0.038 (0.000 - 0.077)	0.793 (0.710 - 0.876)	0.095
-18	0.781 (0.695 - 0.867)	0.784 (0.698 - 0.870)	0.667 (0.569 - 0.765)	0.025 (0.000 - 0.058)	0.843 (0.767 - 0.919)	0.060
-17	0.819 (0.739 - 0.899)	0.744 (0.654 - 0.834)	0.759 (0.671 - 0.847)	0.027 (0.000 - 0.060)	0.800 (0.717 - 0.883)	0.085
-16	0.816 (0.733 - 0.899)	0.711 (0.613 - 0.809)	0.779 (0.690 - 0.868)	0.027 (0.000 - 0.062)	0.808 (0.723 - 0.893)	0.095
-15	0.773 (0.688 - 0.858)	0.819 (0.741 - 0.897)	0.603 (0.504 - 0.702)	0.025 (0.000 - 0.057)	0.852 (0.780 - 0.924)	0.055
-14	0.759 (0.673 - 0.845)	0.734 (0.645 - 0.823)	0.691 (0.598 - 0.784)	0.031 (0.000 - 0.066)	0.833 (0.758 - 0.908)	0.075
-13	0.812 (0.736 - 0.888)	0.653 (0.560 - 0.746)	0.836 (0.764 - 0.908)	0.036 (0.000 - 0.072)	0.734 (0.648 - 0.820)	0.130
-12	0.792 (0.713 - 0.871)	0.594 (0.498 - 0.690)	0.843 (0.772 - 0.914)	0.042 (0.003 - 0.081)	0.745 (0.660 - 0.830)	0.160
-11	0.804 (0.728 - 0.880)	0.740 (0.656 - 0.824)	0.727 (0.641 - 0.813)	0.032 (0.000 - 0.066)	0.798 (0.721 - 0.875)	0.100
-10	0.842 (0.772 - 0.912)	0.883 (0.821 - 0.945)	0.642 (0.549 - 0.735)	0.160 (0.089 - 0.231)	0.815 (0.740 - 0.890)	0.070
-9	0.815 (0.739 - 0.891)	0.720 (0.632 - 0.808)	0.780 (0.699 - 0.861)	0.031 (0.000 - 0.065)	0.774 (0.692 - 0.856)	0.130
-8	0.816 (0.744 - 0.888)	0.768 (0.690 - 0.846)	0.746 (0.665 - 0.827)	0.030 (0.000 - 0.062)	0.766 (0.688 - 0.844)	0.110
-7	0.832 (0.762 - 0.902)	0.766 (0.687 - 0.845)	0.746 (0.665 - 0.827)	0.030 (0.000 - 0.062)	0.769 (0.691 - 0.847)	0.095
-6	0.847 (0.780 - 0.914)	0.811 (0.738 - 0.884)	0.734 (0.652 - 0.816)	0.025 (0.000 - 0.054)	0.767 (0.688 - 0.846)	0.095
-5	0.854 (0.789 - 0.919)	0.807 (0.735 - 0.879)	0.777 (0.701 - 0.853)	0.025 (0.000 - 0.054)	0.729 (0.647 - 0.811)	0.110
-4	0.845 (0.779 - 0.911)	0.789 (0.714 - 0.864)	0.776 (0.699 - 0.853)	0.027 (0.000 - 0.057)	0.735 (0.654 - 0.816)	0.115
-3	0.838 (0.770 - 0.906)	0.833 (0.765 - 0.901)	0.718 (0.635 - 0.801)	0.023 (0.000 - 0.051)	0.768 (0.691 - 0.845)	0.085
-2	0.848 (0.782 - 0.914)	0.852 (0.787 - 0.917)	0.718 (0.636 - 0.800)	0.021 (0.000 - 0.047)	0.762 (0.684 - 0.840)	0.085
-1	0.881 (0.818 - 0.944)	0.893 (0.833 - 0.953)	0.750 (0.666 - 0.834)	0.013 (0.000 - 0.035)	0.753 (0.670 - 0.836)	0.095
Mean ± SD	**0.810 ± 0.034**	**0.754 ± 0.079**	**0.739 ± 0.057**	**0.034 ± 0.028**	**0.793 ± 0.040**	
Model	**SVM**					

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-24	0.761 (0.668 - 0.854)	0.642 (0.538 - 0.746)	0.783 (0.693 - 0.873)	0.032 (0.000 - 0.070)	0.822 (0.739 - 0.905)	0.065
-23	0.803 (0.715 - 0.891)	0.671 (0.567 - 0.775)	0.833 (0.751 - 0.915)	0.027 (0.000 - 0.063)	0.778 (0.686 - 0.870)	0.081
-22	0.810 (0.725 - 0.895)	0.790 (0.701 - 0.879)	0.728 (0.631 - 0.825)	0.021 (0.000 - 0.052)	0.826 (0.743 - 0.909)	0.061
-21	0.752 (0.661 - 0.843)	0.793 (0.708 - 0.878)	0.625 (0.523 - 0.727)	0.025 (0.000 - 0.058)	0.858 (0.785 - 0.931)	0.050
-20	0.747 (0.654 - 0.840)	0.750 (0.657 - 0.843)	0.679 (0.579 - 0.779)	0.027 (0.000 - 0.062)	0.850 (0.774 - 0.926)	0.060
-19	0.741 (0.651 - 0.831)	0.670 (0.573 - 0.767)	0.733 (0.642 - 0.824)	0.036 (0.000 - 0.074)	0.829 (0.752 - 0.906)	0.064
-18	0.790 (0.705 - 0.875)	0.727 (0.634 - 0.820)	0.746 (0.655 - 0.837)	0.028 (0.000 - 0.062)	0.816 (0.735 - 0.897)	0.070
-17	0.753 (0.664 - 0.842)	0.622 (0.522 - 0.722)	0.825 (0.746 - 0.904)	0.036 (0.000 - 0.074)	0.777 (0.691 - 0.863)	0.086
-16	0.761 (0.669 - 0.853)	0.675 (0.574 - 0.776)	0.743 (0.649 - 0.837)	0.031 (0.000 - 0.068)	0.837 (0.758 - 0.916)	0.071
-15	0.779 (0.695 - 0.863)	0.681 (0.587 - 0.775)	0.796 (0.715 - 0.877)	0.033 (0.000 - 0.069)	0.780 (0.696 - 0.864)	0.078
-14	0.758 (0.671 - 0.845)	0.777 (0.693 - 0.861)	0.637 (0.540 - 0.734)	0.029 (0.000 - 0.063)	0.847 (0.774 - 0.920)	0.054
-13	0.759 (0.676 - 0.842)	0.772 (0.690 - 0.854)	0.629 (0.535 - 0.723)	0.032 (0.000 - 0.066)	0.841 (0.770 - 0.912)	0.056
-12	0.758 (0.674 - 0.842)	0.693 (0.603 - 0.783)	0.730 (0.643 - 0.817)	0.037 (0.000 - 0.074)	0.811 (0.735 - 0.887)	0.065
-11	0.761 (0.679 - 0.843)	0.577 (0.482 - 0.672)	0.842 (0.772 - 0.912)	0.045 (0.005 - 0.085)	0.746 (0.662 - 0.830)	0.092
-10	0.814 (0.739 - 0.889)	0.738 (0.653 - 0.823)	0.760 (0.678 - 0.842)	0.031 (0.000 - 0.064)	0.779 (0.699 - 0.859)	0.083
-9	0.795 (0.716 - 0.874)	0.610 (0.514 - 0.706)	0.850 (0.780 - 0.920)	0.039 (0.001 - 0.077)	0.735 (0.649 - 0.821)	0.112
-8	0.809 (0.736 - 0.882)	0.661 (0.573 - 0.749)	0.852 (0.786 - 0.918)	0.039 (0.003 - 0.075)	0.690 (0.604 - 0.776)	0.101
-7	0.816 (0.744 - 0.888)	0.667 (0.579 - 0.755)	0.849 (0.782 - 0.916)	0.038 (0.002 - 0.074)	0.694 (0.608 - 0.780)	0.105
-6	0.826 (0.755 - 0.897)	0.757 (0.677 - 0.837)	0.787 (0.711 - 0.863)	0.030 (0.000 - 0.062)	0.738 (0.656 - 0.820)	0.083
-5	0.805 (0.732 - 0.878)	0.772 (0.695 - 0.849)	0.742 (0.662 - 0.822)	0.031 (0.000 - 0.063)	0.765 (0.687 - 0.843)	0.070
-4	0.821 (0.751 - 0.891)	0.842 (0.775 - 0.909)	0.712 (0.629 - 0.795)	0.022 (0.000 - 0.049)	0.769 (0.692 - 0.846)	0.069
-3	0.825 (0.755 - 0.895)	0.825 (0.755 - 0.895)	0.725 (0.643 - 0.807)	0.024 (0.000 - 0.052)	0.765 (0.687 - 0.843)	0.071
-2	0.831 (0.763 - 0.899)	0.748 (0.669 - 0.827)	0.826 (0.757 - 0.895)	0.031 (0.000 - 0.063)	0.693 (0.609 - 0.777)	0.089
-1	0.878 (0.815 - 0.941)	0.786 (0.707 - 0.865)	0.833 (0.761 - 0.905)	0.023 (0.000 - 0.052)	0.699 (0.610 - 0.788)	0.091
Mean ± SD	**0.790 ± 0.035**	**0.719 ± 0.071**	**0.761 ± 0.071**	**0.031 ± 0.006**	**0.781 ± 0.054**	

Table 7.3.B Performance measures for each model at each lead time when predicting AKI within 25h since ICU admission, using missing values in testing data.

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**BARTm**					
-24	0.844 (0.761 - 0.927)	0.870 (0.793 - 0.947)	0.699 (0.594 - 0.804)	0.011 (0.000 - 0.035)	0.854 (0.773 - 0.935)	0.061
-23	0.860 (0.787 - 0.933)	0.759 (0.669 - 0.849)	0.797 (0.712 - 0.882)	0.020 (0.000 - 0.049)	0.800 (0.716 - 0.884)	0.096
-22	0.851 (0.776 - 0.926)	0.791 (0.705 - 0.877)	0.738 (0.645 - 0.831)	0.018 (0.000 - 0.046)	0.834 (0.755 - 0.913)	0.094
-21	0.823 (0.743 - 0.903)	0.682 (0.585 - 0.779)	0.829 (0.750 - 0.908)	0.025 (0.000 - 0.058)	0.787 (0.701 - 0.873)	0.140
-20	0.821 (0.737 - 0.905)	0.793 (0.704 - 0.882)	0.718 (0.619 - 0.817)	0.018 (0.000 - 0.047)	0.850 (0.772 - 0.928)	0.070
-19	0.828 (0.750 - 0.906)	0.780 (0.695 - 0.865)	0.715 (0.622 - 0.808)	0.021 (0.000 - 0.050)	0.839 (0.763 - 0.915)	0.078
-18	0.830 (0.750 - 0.910)	0.824 (0.743 - 0.905)	0.714 (0.618 - 0.810)	0.016 (0.000 - 0.043)	0.842 (0.764 - 0.920)	0.078
-17	0.832 (0.759 - 0.905)	0.861 (0.794 - 0.928)	0.695 (0.605 - 0.785)	0.015 (0.000 - 0.039)	0.819 (0.744 - 0.894)	0.066
-16	0.837 (0.765 - 0.909)	0.777 (0.695 - 0.859)	0.775 (0.693 - 0.857)	0.022 (0.000 - 0.051)	0.784 (0.703 - 0.865)	0.084
-15	0.797 (0.717 - 0.877)	0.680 (0.587 - 0.773)	0.795 (0.715 - 0.875)	0.029 (0.000 - 0.062)	0.802 (0.723 - 0.881)	0.101
-14	0.816 (0.740 - 0.892)	0.747 (0.661 - 0.833)	0.779 (0.697 - 0.861)	0.024 (0.000 - 0.054)	0.796 (0.717 - 0.875)	0.115
-13	0.841 (0.770 - 0.912)	0.745 (0.660 - 0.830)	0.766 (0.684 - 0.848)	0.025 (0.000 - 0.055)	0.800 (0.722 - 0.878)	0.113
-12	0.845 (0.774 - 0.916)	0.757 (0.673 - 0.841)	0.811 (0.734 - 0.888)	0.023 (0.000 - 0.052)	0.759 (0.675 - 0.843)	0.115
-11	0.842 (0.774 - 0.910)	0.848 (0.782 - 0.914)	0.679 (0.593 - 0.765)	0.019 (0.000 - 0.044)	0.814 (0.742 - 0.886)	0.067
-10	0.819 (0.747 - 0.891)	0.909 (0.855 - 0.963)	0.583 (0.491 - 0.675)	0.013 (0.000 - 0.034)	0.844 (0.776 - 0.912)	0.050
-9	0.799 (0.727 - 0.871)	0.720 (0.639 - 0.801)	0.743 (0.664 - 0.822)	0.033 (0.001 - 0.065)	0.797 (0.724 - 0.870)	0.105
-8	0.788 (0.711 - 0.865)	0.685 (0.597 - 0.773)	0.753 (0.671 - 0.835)	0.033 (0.000 - 0.067)	0.814 (0.740 - 0.888)	0.110
-7	0.829 (0.759 - 0.899)	0.757 (0.677 - 0.837)	0.746 (0.665 - 0.827)	0.027 (0.000 - 0.057)	0.798 (0.723 - 0.873)	0.088
-6	0.820 (0.749 - 0.891)	0.874 (0.812 - 0.936)	0.636 (0.546 - 0.726)	0.017 (0.000 - 0.041)	0.830 (0.760 - 0.900)	0.052
-5	0.826 (0.758 - 0.894)	0.603 (0.516 - 0.690)	0.900 (0.847 - 0.953)	0.040 (0.005 - 0.075)	0.639 (0.553 - 0.725)	0.206
-4	0.840 (0.784 - 0.896)	0.777 (0.714 - 0.840)	0.753 (0.688 - 0.818)	0.027 (0.002 - 0.052)	0.773 (0.709 - 0.837)	0.098
-3	0.818 (0.759 - 0.877)	0.868 (0.817 - 0.919)	0.607 (0.533 - 0.681)	0.020 (0.000 - 0.041)	0.829 (0.772 - 0.886)	0.051
-2	0.837 (0.770 - 0.904)	0.785 (0.711 - 0.859)	0.758 (0.681 - 0.835)	0.026 (0.000 - 0.055)	0.767 (0.691 - 0.843)	0.074
-1	0.883 (0.825 - 0.941)	0.833 (0.765 - 0.901)	0.785 (0.710 - 0.860)	0.019 (0.000 - 0.044)	0.736 (0.656 - 0.816)	0.088
Mean ± SD	**0.830 ± 0.020**	**0.780 ± 0.074**	**0.741 ± 0.070**	**0.023 ± 0.007**	**0.800 ± 0.046**	
Model	**C5.0**					
-24	0.784 (0.690 - 0.878)	0.675 (0.568 - 0.782)	0.781 (0.686 - 0.876)	0.024 (0.000 - 0.059)	0.846 (0.763 - 0.929)	0.147
-23	0.796 (0.711 - 0.881)	0.655 (0.555 - 0.755)	0.814 (0.732 - 0.896)	0.028 (0.000 - 0.063)	0.809 (0.726 - 0.892)	0.154

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-22	0.808 (0.725 - 0.891)	0.791 (0.705 - 0.877)	0.706 (0.610 - 0.802)	0.019 (0.000 - 0.048)	0.849 (0.773 - 0.925)	0.084
-21	0.812 (0.730 - 0.894)	0.773 (0.685 - 0.861)	0.745 (0.654 - 0.836)	0.020 (0.000 - 0.049)	0.830 (0.752 - 0.908)	0.136
-20	0.777 (0.686 - 0.868)	0.695 (0.594 - 0.796)	0.760 (0.666 - 0.854)	0.025 (0.000 - 0.059)	0.846 (0.767 - 0.925)	0.137
-19	0.773 (0.687 - 0.859)	0.681 (0.585 - 0.777)	0.798 (0.716 - 0.880)	0.027 (0.000 - 0.060)	0.809 (0.728 - 0.890)	0.163
-18	0.769 (0.679 - 0.859)	0.776 (0.687 - 0.865)	0.676 (0.577 - 0.775)	0.021 (0.000 - 0.051)	0.865 (0.792 - 0.938)	0.081
-17	0.799 (0.721 - 0.877)	0.802 (0.724 - 0.880)	0.683 (0.592 - 0.774)	0.022 (0.000 - 0.051)	0.835 (0.763 - 0.907)	0.076
-16	0.786 (0.706 - 0.866)	0.728 (0.641 - 0.815)	0.759 (0.675 - 0.843)	0.028 (0.000 - 0.060)	0.806 (0.728 - 0.884)	0.135
-15	0.768 (0.684 - 0.852)	0.649 (0.554 - 0.744)	0.799 (0.719 - 0.879)	0.032 (0.000 - 0.067)	0.806 (0.727 - 0.885)	0.134
-14	0.739 (0.652 - 0.826)	0.566 (0.468 - 0.664)	0.875 (0.810 - 0.940)	0.036 (0.000 - 0.073)	0.743 (0.657 - 0.829)	0.251
-13	0.779 (0.698 - 0.860)	0.686 (0.596 - 0.776)	0.758 (0.675 - 0.841)	0.031 (0.000 - 0.065)	0.818 (0.743 - 0.893)	0.161
-12	0.824 (0.749 - 0.899)	0.757 (0.673 - 0.841)	0.790 (0.710 - 0.870)	0.024 (0.000 - 0.054)	0.778 (0.697 - 0.859)	0.147
-11	0.775 (0.698 - 0.852)	0.732 (0.650 - 0.814)	0.728 (0.646 - 0.810)	0.031 (0.000 - 0.063)	0.811 (0.738 - 0.884)	0.137
-10	0.801 (0.726 - 0.876)	0.736 (0.654 - 0.818)	0.776 (0.698 - 0.854)	0.028 (0.000 - 0.059)	0.782 (0.705 - 0.859)	0.156
-9	0.808 (0.737 - 0.879)	0.788 (0.714 - 0.862)	0.748 (0.670 - 0.826)	0.025 (0.000 - 0.053)	0.778 (0.703 - 0.853)	0.145
-8	0.799 (0.723 - 0.875)	0.667 (0.578 - 0.756)	0.831 (0.760 - 0.902)	0.032 (0.000 - 0.065)	0.754 (0.672 - 0.836)	0.176
-7	0.783 (0.706 - 0.860)	0.757 (0.677 - 0.837)	0.708 (0.623 - 0.793)	0.028 (0.000 - 0.059)	0.819 (0.747 - 0.891)	0.104
-6	0.789 (0.713 - 0.865)	0.712 (0.628 - 0.796)	0.764 (0.685 - 0.843)	0.031 (0.000 - 0.063)	0.796 (0.721 - 0.871)	0.150
-5	0.801 (0.730 - 0.872)	0.628 (0.542 - 0.714)	0.845 (0.781 - 0.909)	0.040 (0.005 - 0.075)	0.725 (0.645 - 0.805)	0.188
-4	0.814 (0.755 - 0.873)	0.777 (0.714 - 0.840)	0.743 (0.677 - 0.809)	0.027 (0.002 - 0.052)	0.779 (0.716 - 0.842)	0.145
-3	0.824 (0.766 - 0.882)	0.835 (0.779 - 0.891)	0.701 (0.632 - 0.770)	0.022 (0.000 - 0.044)	0.793 (0.732 - 0.854)	0.091
-2	0.806 (0.735 - 0.877)	0.752 (0.674 - 0.830)	0.735 (0.655 - 0.815)	0.031 (0.000 - 0.062)	0.790 (0.717 - 0.863)	0.128
-1	0.847 (0.781 - 0.913)	0.758 (0.680 - 0.836)	0.813 (0.742 - 0.884)	0.027 (0.000 - 0.056)	0.727 (0.646 - 0.808)	0.156
Mean ± SD	**0.794 ± 0.023**	**0.724 ± 0.064**	**0.764 ± 0.051**	**0.027 ± 0.005**	**0.800 ± 0.037**	

Table 7.3.C. Performance measures for each model at each lead time when predicting AKI within 25h since ICU admission, using median imputation in testing data.

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**AdaBoost**					
-24	0.796 (0.704 - 0.888)	0.857 (0.777 - 0.937)	0.600 (0.488 - 0.712)	0.014 (0.000 - 0.041)	0.888 (0.816 - 0.960)	0.145
-23	0.806 (0.723 - 0.889)	0.701 (0.605 - 0.797)	0.807 (0.724 - 0.890)	0.024 (0.000 - 0.056)	0.804 (0.721 - 0.887)	0.220
-22	0.839 (0.761 - 0.917)	0.860 (0.787 - 0.933)	0.684 (0.586 - 0.782)	0.013 (0.000 - 0.037)	0.847 (0.771 - 0.923)	0.172
-21	0.823 (0.743 - 0.903)	0.739 (0.647 - 0.831)	0.755 (0.665 - 0.845)	0.023 (0.000 - 0.054)	0.831 (0.753 - 0.909)	0.196
-20	0.771 (0.679 - 0.863)	0.768 (0.676 - 0.860)	0.686 (0.584 - 0.788)	0.021 (0.000 - 0.052)	0.867 (0.793 - 0.941)	0.172
-19	0.783 (0.698 - 0.868)	0.714 (0.621 - 0.807)	0.732 (0.641 - 0.823)	0.027 (0.000 - 0.060)	0.843 (0.768 - 0.918)	0.195
-18	0.795 (0.709 - 0.881)	0.741 (0.648 - 0.834)	0.754 (0.662 - 0.846)	0.022 (0.000 - 0.053)	0.836 (0.757 - 0.915)	0.200
-17	0.820 (0.745 - 0.895)	0.782 (0.701 - 0.863)	0.744 (0.659 - 0.829)	0.022 (0.000 - 0.051)	0.807 (0.730 - 0.884)	0.198
-16	0.781 (0.700 - 0.862)	0.709 (0.620 - 0.798)	0.748 (0.663 - 0.833)	0.030 (0.000 - 0.063)	0.817 (0.741 - 0.893)	0.196
-15	0.572 (0.474 - 0.670)	0.354 (0.259 - 0.449)	0.830 (0.755 - 0.905)	0.072 (0.000 - 0.123)	0.828 (0.753 - 0.903)	0.172
-14	0.821 (0.745 - 0.897)	0.737 (0.650 - 0.824)	0.835 (0.762 - 0.908)	0.023 (0.000 - 0.053)	0.747 (0.661 - 0.833)	0.228
-13	0.810 (0.734 - 0.886)	0.706 (0.618 - 0.794)	0.774 (0.693 - 0.855)	0.029 (0.000 - 0.062)	0.803 (0.726 - 0.880)	0.219
-12	0.847 (0.776 - 0.918)	0.796 (0.717 - 0.875)	0.766 (0.683 - 0.849)	0.021 (0.000 - 0.049)	0.788 (0.708 - 0.868)	0.221
-11	0.824 (0.753 - 0.895)	0.723 (0.640 - 0.806)	0.790 (0.715 - 0.865)	0.029 (0.000 - 0.060)	0.771 (0.693 - 0.849)	0.228
-10	0.834 (0.764 - 0.904)	0.827 (0.756 - 0.898)	0.694 (0.608 - 0.780)	0.021 (0.000 - 0.048)	0.814 (0.741 - 0.887)	0.182
-9	0.789 (0.715 - 0.863)	0.636 (0.549 - 0.723)	0.834 (0.767 - 0.901)	0.038 (0.004 - 0.072)	0.741 (0.662 - 0.820)	0.250
-8	0.819 (0.746 - 0.892)	0.833 (0.762 - 0.904)	0.638 (0.547 - 0.729)	0.021 (0.000 - 0.048)	0.840 (0.771 - 0.909)	0.151
-7	0.824 (0.753 - 0.895)	0.766 (0.687 - 0.845)	0.776 (0.698 - 0.854)	0.025 (0.000 - 0.054)	0.775 (0.697 - 0.853)	0.222
-6	0.803 (0.729 - 0.877)	0.640 (0.551 - 0.729)	0.827 (0.757 - 0.897)	0.036 (0.001 - 0.071)	0.760 (0.681 - 0.839)	0.245
-5	0.792 (0.720 - 0.864)	0.630 (0.544 - 0.716)	0.852 (0.789 - 0.915)	0.032 (0.001 - 0.063)	0.754 (0.677 - 0.831)	0.249
-4	0.814 (0.755 - 0.873)	0.777 (0.714 - 0.840)	0.710 (0.641 - 0.779)	0.029 (0.004 - 0.054)	0.800 (0.739 - 0.861)	0.180
-3	0.814 (0.755 - 0.873)	0.702 (0.633 - 0.771)	0.804 (0.744 - 0.864)	0.033 (0.006 - 0.060)	0.749 (0.683 - 0.815)	0.237
-2	0.821 (0.752 - 0.890)	0.719 (0.638 - 0.800)	0.816 (0.746 - 0.886)	0.031 (0.000 - 0.062)	0.732 (0.652 - 0.812)	0.242
-1	0.882 (0.823 - 0.941)	0.850 (0.785 - 0.915)	0.812 (0.741 - 0.883)	0.017 (0.000 - 0.041)	0.704 (0.621 - 0.787)	0.205
Mean ± SD	**0.803 ± 0.055**	**0.732 ± 0.104**	**0.761 ± 0.066**	**0.027 ± 0.011**	**0.798 ± 0.046**	
Model	**BARTm**					
-24	0.860 (0.780 - 0.940)	0.779 (0.684 - 0.874)	0.815 (0.726 - 0.904)	0.016 (0.000 - 0.045)	0.801 (0.709 - 0.893)	0.081
-23	0.868 (0.797 - 0.939)	0.655 (0.555 - 0.755)	0.904 (0.842 - 0.966)	0.025 (0.000 - 0.058)	0.687 (0.590 - 0.784)	0.155
-22	0.866 (0.794 - 0.938)	0.756 (0.665 - 0.847)	0.835 (0.757 - 0.913)	0.019 (0.000 - 0.048)	0.768 (0.679 - 0.857)	0.117
-21	0.851 (0.777 - 0.925)	0.784 (0.698 - 0.870)	0.787 (0.701 - 0.873)	0.018 (0.000 - 0.046)	0.801 (0.718 - 0.884)	0.095

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-20	0.824 (0.741 - 0.907)	0.878 (0.806 - 0.950)	0.625 (0.519 - 0.731)	0.012 (0.000 - 0.036)	0.872 (0.799 - 0.945)	0.039
-19	0.840 (0.765 - 0.915)	0.648 (0.550 - 0.746)	0.872 (0.803 - 0.941)	0.027 (0.000 - 0.060)	0.738 (0.648 - 0.828)	0.140
-18	0.853 (0.778 - 0.928)	0.882 (0.813 - 0.951)	0.704 (0.607 - 0.801)	0.011 (0.000 - 0.033)	0.838 (0.760 - 0.916)	0.062
-17	0.842 (0.771 - 0.913)	0.881 (0.818 - 0.944)	0.695 (0.605 - 0.785)	0.013 (0.000 - 0.035)	0.816 (0.740 - 0.892)	0.057
-16	0.838 (0.766 - 0.910)	0.825 (0.751 - 0.899)	0.716 (0.628 - 0.804)	0.019 (0.000 - 0.046)	0.812 (0.735 - 0.889)	0.060
-15	0.609 (0.512 - 0.706)	0.421 (0.323 - 0.519)	0.804 (0.725 - 0.883)	0.067 (0.017 - 0.117)	0.823 (0.747 - 0.899)	0.049
-14	0.836 (0.763 - 0.909)	0.788 (0.707 - 0.869)	0.781 (0.700 - 0.862)	0.020 (0.000 - 0.048)	0.785 (0.704 - 0.866)	0.091
-13	0.862 (0.795 - 0.929)	0.716 (0.628 - 0.804)	0.843 (0.772 - 0.914)	0.026 (0.000 - 0.057)	0.736 (0.650 - 0.822)	0.123
-12	0.863 (0.796 - 0.930)	0.709 (0.620 - 0.798)	0.869 (0.803 - 0.935)	0.026 (0.000 - 0.057)	0.700 (0.610 - 0.790)	0.147
-11	0.858 (0.793 - 0.923)	0.759 (0.680 - 0.838)	0.816 (0.744 - 0.888)	0.025 (0.000 - 0.054)	0.737 (0.655 - 0.819)	0.108
-10	0.856 (0.790 - 0.922)	0.845 (0.777 - 0.913)	0.706 (0.621 - 0.791)	0.018 (0.000 - 0.043)	0.805 (0.731 - 0.879)	0.069
-9	0.856 (0.793 - 0.919)	0.831 (0.763 - 0.899)	0.736 (0.656 - 0.816)	0.021 (0.000 - 0.047)	0.777 (0.702 - 0.852)	0.078
-8	0.837 (0.767 - 0.907)	0.667 (0.578 - 0.756)	0.834 (0.763 - 0.905)	0.032 (0.000 - 0.065)	0.751 (0.669 - 0.833)	0.116
-7	0.840 (0.772 - 0.908)	0.892 (0.834 - 0.950)	0.613 (0.522 - 0.704)	0.015 (0.000 - 0.038)	0.836 (0.767 - 0.905)	0.042
-6	0.842 (0.774 - 0.910)	0.793 (0.718 - 0.868)	0.721 (0.638 - 0.804)	0.024 (0.000 - 0.052)	0.805 (0.731 - 0.879)	0.070
-5	0.819 (0.750 - 0.888)	0.710 (0.629 - 0.791)	0.791 (0.719 - 0.863)	0.027 (0.000 - 0.056)	0.793 (0.721 - 0.865)	0.095
-4	0.859 (0.806 - 0.912)	0.711 (0.642 - 0.780)	0.843 (0.788 - 0.898)	0.031 (0.005 - 0.057)	0.702 (0.633 - 0.771)	0.124
-3	0.845 (0.790 - 0.900)	0.785 (0.723 - 0.847)	0.733 (0.666 - 0.800)	0.027 (0.002 - 0.052)	0.785 (0.723 - 0.847)	0.069
-2	0.846 (0.781 - 0.911)	0.694 (0.611 - 0.777)	0.819 (0.750 - 0.888)	0.034 (0.001 - 0.067)	0.736 (0.656 - 0.816)	0.117
-1	0.892 (0.836 - 0.948)	0.825 (0.756 - 0.894)	0.791 (0.717 - 0.865)	0.020 (0.000 - 0.045)	0.732 (0.651 - 0.813)	0.075
Mean ± SD	**0.840 ± 0.052**	**0.760 ± 0.103**	**0.777 ± 0.076**	**0.024 ± 0.011**	**0.777 ± 0.048**	
Model	**C5.0**					
-24	0.785 (0.691 - 0.879)	0.688 (0.582 - 0.794)	0.768 (0.671 - 0.865)	0.023 (0.000 - 0.057)	0.851 (0.769 - 0.933)	0.146
-23	0.798 (0.714 - 0.882)	0.667 (0.568 - 0.766)	0.824 (0.744 - 0.904)	0.026 (0.000 - 0.059)	0.798 (0.714 - 0.882)	0.154
-22	0.807 (0.724 - 0.890)	0.791 (0.705 - 0.877)	0.704 (0.608 - 0.800)	0.019 (0.000 - 0.048)	0.850 (0.775 - 0.925)	0.084
-21	0.815 (0.734 - 0.896)	0.784 (0.698 - 0.870)	0.739 (0.647 - 0.831)	0.019 (0.000 - 0.048)	0.831 (0.753 - 0.909)	0.136
-20	0.777 (0.686 - 0.868)	0.695 (0.594 - 0.796)	0.765 (0.672 - 0.858)	0.024 (0.000 - 0.058)	0.843 (0.763 - 0.923)	0.137
-19	0.772 (0.686 - 0.858)	0.692 (0.597 - 0.787)	0.788 (0.704 - 0.872)	0.027 (0.000 - 0.060)	0.814 (0.734 - 0.894)	0.161
-18	0.767 (0.677 - 0.857)	0.671 (0.571 - 0.771)	0.780 (0.692 - 0.868)	0.027 (0.000 - 0.061)	0.834 (0.755 - 0.913)	0.157
-17	0.795 (0.716 - 0.874)	0.802 (0.724 - 0.880)	0.673 (0.582 - 0.764)	0.022 (0.000 - 0.051)	0.839 (0.767 - 0.911)	0.076
-16	0.783 (0.702 - 0.864)	0.728 (0.641 - 0.815)	0.758 (0.674 - 0.842)	0.028 (0.000 - 0.060)	0.807 (0.730 - 0.884)	0.134
-15	0.649 (0.554 - 0.744)	0.525 (0.426 - 0.624)	0.729 (0.641 - 0.817)	0.061 (0.013 - 0.109)	0.838 (0.765 - 0.911)	0.072
-14	0.761 (0.677 - 0.845)	0.586 (0.489 - 0.683)	0.860 (0.792 - 0.928)	0.035 (0.000 - 0.071)	0.758 (0.674 - 0.842)	0.238

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-13	0.772 (0.691 - 0.853)	0.686 (0.596 - 0.776)	0.749 (0.665 - 0.833)	0.032 (0.000 - 0.066)	0.823 (0.749 - 0.897)	0.161
-12	0.823 (0.748 - 0.898)	0.728 (0.641 - 0.815)	0.815 (0.739 - 0.891)	0.026 (0.000 - 0.057)	0.763 (0.680 - 0.846)	0.163
-11	0.779 (0.702 - 0.856)	0.750 (0.670 - 0.830)	0.716 (0.632 - 0.800)	0.029 (0.000 - 0.060)	0.814 (0.742 - 0.886)	0.136
-10	0.796 (0.721 - 0.871)	0.736 (0.654 - 0.818)	0.771 (0.692 - 0.850)	0.028 (0.000 - 0.059)	0.786 (0.709 - 0.863)	0.157
-9	0.801 (0.729 - 0.873)	0.763 (0.686 - 0.840)	0.750 (0.672 - 0.828)	0.028 (0.000 - 0.058)	0.782 (0.708 - 0.856)	0.154
-8	0.799 (0.723 - 0.875)	0.667 (0.578 - 0.756)	0.825 (0.753 - 0.897)	0.032 (0.000 - 0.065)	0.760 (0.679 - 0.841)	0.176
-7	0.787 (0.711 - 0.863)	0.757 (0.677 - 0.837)	0.711 (0.627 - 0.795)	0.028 (0.000 - 0.059)	0.818 (0.746 - 0.890)	0.104
-6	0.788 (0.712 - 0.864)	0.667 (0.579 - 0.755)	0.814 (0.742 - 0.886)	0.034 (0.000 - 0.068)	0.767 (0.688 - 0.846)	0.171
-5	0.777 (0.703 - 0.851)	0.740 (0.662 - 0.818)	0.746 (0.668 - 0.824)	0.026 (0.000 - 0.054)	0.817 (0.748 - 0.886)	0.132
-4	0.811 (0.752 - 0.870)	0.760 (0.695 - 0.825)	0.756 (0.691 - 0.821)	0.029 (0.004 - 0.054)	0.775 (0.712 - 0.838)	0.152
-3	0.824 (0.766 - 0.882)	0.835 (0.779 - 0.891)	0.706 (0.637 - 0.775)	0.021 (0.000 - 0.043)	0.790 (0.728 - 0.852)	0.092
-2	0.808 (0.737 - 0.879)	0.752 (0.674 - 0.830)	0.730 (0.650 - 0.810)	0.031 (0.000 - 0.062)	0.794 (0.721 - 0.867)	0.116
-1	0.850 (0.785 - 0.915)	0.783 (0.708 - 0.858)	0.805 (0.733 - 0.877)	0.024 (0.000 - 0.052)	0.728 (0.647 - 0.809)	0.156
Mean ± SD	**0.789 ± 0.036**	**0.719 ± 0.069**	**0.762 ± 0.046**	**0.028 ± 0.008**	**0.803 ± 0.033**	
Model	**GBM**					
-24	0.847 (0.764 - 0.930)	0.805 (0.714 - 0.896)	0.785 (0.691 - 0.879)	0.014 (0.000 - 0.041)	0.819 (0.731 - 0.907)	0.072
-23	0.860 (0.787 - 0.933)	0.747 (0.656 - 0.838)	0.827 (0.748 - 0.906)	0.020 (0.000 - 0.049)	0.776 (0.688 - 0.864)	0.107
-22	0.862 (0.789 - 0.935)	0.849 (0.773 - 0.925)	0.747 (0.655 - 0.839)	0.013 (0.000 - 0.037)	0.818 (0.736 - 0.900)	0.087
-21	0.846 (0.771 - 0.921)	0.784 (0.698 - 0.870)	0.784 (0.698 - 0.870)	0.018 (0.000 - 0.046)	0.803 (0.720 - 0.886)	0.093
-20	0.821 (0.737 - 0.905)	0.854 (0.777 - 0.931)	0.665 (0.562 - 0.768)	0.014 (0.000 - 0.040)	0.862 (0.786 - 0.938)	0.053
-19	0.790 (0.706 - 0.874)	0.626 (0.527 - 0.725)	0.829 (0.752 - 0.906)	0.031 (0.000 - 0.067)	0.796 (0.713 - 0.879)	0.132
-18	0.827 (0.747 - 0.907)	0.788 (0.701 - 0.875)	0.745 (0.652 - 0.838)	0.018 (0.000 - 0.046)	0.832 (0.753 - 0.911)	0.069
-17	0.842 (0.771 - 0.913)	0.822 (0.747 - 0.897)	0.720 (0.632 - 0.808)	0.019 (0.000 - 0.046)	0.813 (0.737 - 0.889)	0.066
-16	0.832 (0.759 - 0.905)	0.796 (0.717 - 0.875)	0.727 (0.640 - 0.814)	0.022 (0.000 - 0.051)	0.811 (0.734 - 0.888)	0.059
-15	0.584 (0.486 - 0.682)	0.333 (0.239 - 0.427)	0.888 (0.825 - 0.951)	0.070 (0.019 - 0.121)	0.770 (0.686 - 0.854)	0.116
-14	0.851 (0.781 - 0.921)	0.727 (0.639 - 0.815)	0.836 (0.763 - 0.909)	0.024 (0.000 - 0.054)	0.748 (0.662 - 0.834)	0.117
-13	0.853 (0.784 - 0.922)	0.725 (0.638 - 0.812)	0.821 (0.747 - 0.895)	0.026 (0.000 - 0.057)	0.759 (0.676 - 0.842)	0.107
-12	0.860 (0.792 - 0.928)	0.709 (0.620 - 0.798)	0.849 (0.779 - 0.919)	0.026 (0.000 - 0.057)	0.730 (0.643 - 0.817)	0.153
-11	0.854 (0.789 - 0.919)	0.786 (0.710 - 0.862)	0.772 (0.694 - 0.850)	0.023 (0.000 - 0.051)	0.770 (0.692 - 0.848)	0.129
-10	0.850 (0.783 - 0.917)	0.755 (0.675 - 0.835)	0.803 (0.729 - 0.877)	0.025 (0.000 - 0.054)	0.755 (0.675 - 0.835)	0.104
-9	0.829 (0.761 - 0.897)	0.729 (0.649 - 0.809)	0.824 (0.755 - 0.893)	0.029 (0.000 - 0.059)	0.726 (0.646 - 0.806)	0.110
-8	0.834 (0.763 - 0.905)	0.769 (0.689 - 0.849)	0.750 (0.668 - 0.832)	0.025 (0.000 - 0.055)	0.797 (0.721 - 0.873)	0.107
-7	0.845 (0.778 - 0.912)	0.820 (0.749 - 0.891)	0.714 (0.630 - 0.798)	0.021 (0.000 - 0.048)	0.804 (0.730 - 0.878)	0.072

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-6	0.843 (0.775 - 0.911)	0.694 (0.608 - 0.780)	0.841 (0.773 - 0.909)	0.030 (0.000 - 0.062)	0.729 (0.646 - 0.812)	0.143
-5	0.835 (0.769 - 0.901)	0.740 (0.662 - 0.818)	0.786 (0.713 - 0.859)	0.025 (0.000 - 0.053)	0.790 (0.717 - 0.863)	0.163
-4	0.842 (0.787 - 0.897)	0.661 (0.589 - 0.733)	0.849 (0.795 - 0.903)	0.036 (0.008 - 0.064)	0.710 (0.641 - 0.779)	0.091
-3	0.841 (0.786 - 0.896)	0.736 (0.669 - 0.803)	0.796 (0.735 - 0.857)	0.030 (0.004 - 0.056)	0.748 (0.682 - 0.814)	0.113
-2	0.849 (0.784 - 0.914)	0.661 (0.576 - 0.746)	0.880 (0.821 - 0.939)	0.035 (0.002 - 0.068)	0.661 (0.576 - 0.746)	0.188
-1	0.888 (0.831 - 0.945)	0.858 (0.794 - 0.922)	0.778 (0.702 - 0.854)	0.017 (0.000 - 0.041)	0.736 (0.656 - 0.816)	0.889
Mean ± SD	**0.833 ± 0.056**	**0.741 ± 0.107**	**0.792 ± 0.055**	**0.025 ± 0.011**	**0.773 ± 0.045**	
Model	**LR**					
-24	0.787 (0.693 - 0.881)	0.636 (0.526 - 0.746)	0.825 (0.738 - 0.912)	0.025 (0.000 - 0.061)	0.823 (0.735 - 0.911)	0.058
-23	0.827 (0.748 - 0.906)	0.839 (0.762 - 0.916)	0.716 (0.621 - 0.811)	0.015 (0.000 - 0.041)	0.835 (0.757 - 0.913)	0.040
-22	0.679 (0.580 - 0.778)	0.558 (0.453 - 0.663)	0.812 (0.729 - 0.895)	0.035 (0.000 - 0.074)	0.836 (0.758 - 0.914)	0.055
-21	0.622 (0.521 - 0.723)	0.284 (0.190 - 0.378)	0.960 (0.919 - 1.001)	0.048 (0.003 - 0.093)	0.675 (0.577 - 0.773)	0.001
-20	0.589 (0.481 - 0.697)	0.207 (0.118 - 0.296)	0.971 (0.934 - 1.008)	0.049 (0.002 - 0.096)	0.691 (0.590 - 0.792)	0.001
-19	0.620 (0.520 - 0.720)	0.286 (0.193 - 0.379)	0.955 (0.912 - 0.998)	0.050 (0.005 - 0.095)	0.694 (0.599 - 0.789)	0.001
-18	0.637 (0.535 - 0.739)	0.482 (0.376 - 0.588)	0.837 (0.758 - 0.916)	0.039 (0.000 - 0.080)	0.839 (0.761 - 0.917)	0.070
-17	0.663 (0.571 - 0.755)	0.911 (0.855 - 0.967)	0.478 (0.381 - 0.575)	0.014 (0.000 - 0.037)	0.880 (0.817 - 0.943)	0.034
-16	0.765 (0.682 - 0.848)	0.816 (0.740 - 0.892)	0.676 (0.584 - 0.768)	0.021 (0.000 - 0.049)	0.833 (0.760 - 0.906)	0.044
-15	0.661 (0.567 - 0.755)	0.498 (0.398 - 0.598)	0.811 (0.733 - 0.889)	0.058 (0.011 - 0.105)	0.791 (0.710 - 0.872)	0.050
-14	0.833 (0.760 - 0.906)	0.838 (0.765 - 0.911)	0.724 (0.636 - 0.812)	0.017 (0.000 - 0.042)	0.812 (0.735 - 0.889)	0.050
-13	0.718 (0.631 - 0.805)	0.863 (0.796 - 0.930)	0.552 (0.455 - 0.649)	0.019 (0.000 - 0.045)	0.869 (0.804 - 0.934)	0.069
-12	0.857 (0.788 - 0.926)	0.777 (0.695 - 0.859)	0.848 (0.778 - 0.918)	0.020 (0.000 - 0.047)	0.712 (0.623 - 0.801)	0.094
-11	0.762 (0.683 - 0.841)	0.714 (0.630 - 0.798)	0.753 (0.673 - 0.833)	0.032 (0.000 - 0.065)	0.800 (0.726 - 0.874)	0.073
-10	0.857 (0.792 - 0.922)	0.727 (0.644 - 0.810)	0.836 (0.767 - 0.905)	0.027 (0.000 - 0.057)	0.728 (0.645 - 0.811)	0.098
-9	0.857 (0.794 - 0.920)	0.839 (0.773 - 0.905)	0.738 (0.659 - 0.817)	0.020 (0.000 - 0.045)	0.774 (0.699 - 0.849)	0.053
-8	0.829 (0.758 - 0.900)	0.806 (0.731 - 0.881)	0.715 (0.629 - 0.801)	0.022 (0.000 - 0.050)	0.810 (0.736 - 0.884)	0.045
-7	0.753 (0.673 - 0.833)	0.784 (0.707 - 0.861)	0.679 (0.592 - 0.766)	0.026 (0.000 - 0.056)	0.828 (0.758 - 0.898)	0.045
-6	0.755 (0.675 - 0.835)	0.847 (0.780 - 0.914)	0.624 (0.534 - 0.714)	0.020 (0.000 - 0.046)	0.839 (0.771 - 0.907)	0.030
-5	0.839 (0.774 - 0.904)	0.730 (0.651 - 0.809)	0.796 (0.724 - 0.868)	0.025 (0.000 - 0.053)	0.784 (0.711 - 0.857)	0.073
-4	0.791 (0.729 - 0.853)	0.810 (0.751 - 0.869)	0.700 (0.630 - 0.770)	0.025 (0.001 - 0.049)	0.798 (0.737 - 0.859)	0.063
-3	0.809 (0.749 - 0.869)	0.810 (0.751 - 0.869)	0.710 (0.641 - 0.779)	0.024 (0.001 - 0.047)	0.793 (0.732 - 0.854)	0.063
-2	0.853 (0.789 - 0.917)	0.752 (0.674 - 0.830)	0.802 (0.730 - 0.874)	0.028 (0.000 - 0.058)	0.738 (0.659 - 0.817)	0.079
-1	0.901 (0.847 - 0.955)	0.783 (0.708 - 0.858)	0.900 (0.845 - 0.955)	0.022 (0.000 - 0.049)	0.580 (0.490 - 0.670)	0.109
Mean ± SD	**0.761 ± 0.091**	**0.692 ± 0.200**	**0.767 ± 0.122**	**0.028 ± 0.012**	**0.782 ± 0.071**	

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**RF**					
-24	0.816 (0.727 - 0.905)	0.662 (0.553 - 0.771)	0.804 (0.713 - 0.895)	0.024 (0.000 - 0.059)	0.834 (0.749 - 0.919)	0.110
-23	0.834 (0.756 - 0.912)	0.770 (0.682 - 0.858)	0.775 (0.687 - 0.863)	0.019 (0.000 - 0.048)	0.813 (0.731 - 0.895)	0.095
-22	0.833 (0.754 - 0.912)	0.779 (0.691 - 0.867)	0.753 (0.662 - 0.844)	0.019 (0.000 - 0.048)	0.827 (0.747 - 0.907)	0.105
-21	0.833 (0.755 - 0.911)	0.795 (0.711 - 0.879)	0.689 (0.592 - 0.786)	0.020 (0.000 - 0.049)	0.852 (0.778 - 0.926)	0.070
-20	0.808 (0.722 - 0.894)	0.780 (0.689 - 0.871)	0.698 (0.597 - 0.799)	0.019 (0.000 - 0.049)	0.861 (0.785 - 0.937)	0.070
-19	0.816 (0.736 - 0.896)	0.769 (0.682 - 0.856)	0.720 (0.628 - 0.812)	0.022 (0.000 - 0.052)	0.839 (0.763 - 0.915)	0.075
-18	0.830 (0.750 - 0.910)	0.753 (0.661 - 0.845)	0.773 (0.684 - 0.862)	0.020 (0.000 - 0.050)	0.823 (0.742 - 0.904)	0.090
-17	0.819 (0.744 - 0.894)	0.802 (0.724 - 0.880)	0.672 (0.580 - 0.764)	0.023 (0.000 - 0.052)	0.840 (0.769 - 0.911)	0.065
-16	0.819 (0.744 - 0.894)	0.621 (0.526 - 0.716)	0.878 (0.814 - 0.942)	0.033 (0.000 - 0.068)	0.712 (0.623 - 0.801)	0.125
-15	0.601 (0.504 - 0.698)	0.428 (0.330 - 0.526)	0.776 (0.693 - 0.859)	0.069 (0.019 - 0.119)	0.839 (0.766 - 0.912)	0.055
-14	0.820 (0.744 - 0.896)	0.747 (0.661 - 0.833)	0.773 (0.690 - 0.856)	0.024 (0.000 - 0.054)	0.799 (0.720 - 0.878)	0.110
-13	0.848 (0.778 - 0.918)	0.814 (0.738 - 0.890)	0.723 (0.636 - 0.810)	0.020 (0.000 - 0.047)	0.813 (0.737 - 0.889)	0.080
-12	0.841 (0.769 - 0.913)	0.825 (0.751 - 0.899)	0.748 (0.663 - 0.833)	0.018 (0.000 - 0.044)	0.794 (0.715 - 0.873)	0.110
-11	0.829 (0.759 - 0.899)	0.750 (0.670 - 0.830)	0.797 (0.723 - 0.871)	0.026 (0.000 - 0.055)	0.758 (0.679 - 0.837)	0.130
-10	0.812 (0.739 - 0.885)	0.664 (0.576 - 0.752)	0.820 (0.748 - 0.892)	0.034 (0.000 - 0.068)	0.762 (0.682 - 0.842)	0.145
-9	0.811 (0.740 - 0.882)	0.780 (0.705 - 0.855)	0.730 (0.650 - 0.810)	0.027 (0.000 - 0.056)	0.792 (0.719 - 0.865)	0.105
-8	0.796 (0.720 - 0.872)	0.741 (0.658 - 0.824)	0.724 (0.639 - 0.809)	0.029 (0.000 - 0.061)	0.819 (0.746 - 0.892)	0.095
-7	0.818 (0.746 - 0.890)	0.802 (0.728 - 0.876)	0.720 (0.636 - 0.804)	0.023 (0.000 - 0.051)	0.804 (0.730 - 0.878)	0.100
-6	0.811 (0.738 - 0.884)	0.793 (0.718 - 0.868)	0.711 (0.627 - 0.795)	0.024 (0.000 - 0.052)	0.811 (0.738 - 0.884)	0.095
-5	0.830 (0.763 - 0.897)	0.870 (0.810 - 0.930)	0.641 (0.556 - 0.726)	0.015 (0.000 - 0.037)	0.843 (0.778 - 0.908)	0.060
-4	0.808 (0.748 - 0.868)	0.736 (0.669 - 0.803)	0.739 (0.672 - 0.806)	0.032 (0.005 - 0.059)	0.792 (0.730 - 0.854)	0.110
-3	0.810 (0.751 - 0.869)	0.694 (0.624 - 0.764)	0.779 (0.716 - 0.842)	0.035 (0.007 - 0.063)	0.773 (0.709 - 0.837)	0.135
-2	0.805 (0.734 - 0.876)	0.694 (0.611 - 0.777)	0.822 (0.753 - 0.891)	0.034 (0.001 - 0.067)	0.733 (0.653 - 0.813)	0.170
-1	0.853 (0.789 - 0.917)	0.792 (0.718 - 0.866)	0.779 (0.703 - 0.855)	0.024 (0.000 - 0.052)	0.751 (0.672 - 0.830)	0.095
Mean ± SD	**0.813 ± 0.047**	**0.744 ± 0.088**	**0.752 ± 0.053**	**0.026 ± 0.011**	**0.804 ± 0.039**	
Model	**SVM**					
-24	0.762 (0.664 - 0.860)	0.688 (0.582 - 0.794)	0.760 (0.662 - 0.858)	0.024 (0.000 - 0.059)	0.855 (0.774 - 0.936)	0.066
-23	0.818 (0.737 - 0.899)	0.667 (0.568 - 0.766)	0.877 (0.808 - 0.946)	0.025 (0.000 - 0.058)	0.734 (0.641 - 0.827)	0.095
-22	0.804 (0.720 - 0.888)	0.709 (0.613 - 0.805)	0.796 (0.711 - 0.881)	0.024 (0.000 - 0.056)	0.813 (0.731 - 0.895)	0.075
-21	0.792 (0.707 - 0.877)	0.693 (0.597 - 0.789)	0.773 (0.685 - 0.861)	0.026 (0.000 - 0.059)	0.829 (0.750 - 0.908)	0.075
-20	0.777 (0.686 - 0.868)	0.695 (0.594 - 0.796)	0.751 (0.656 - 0.846)	0.025 (0.000 - 0.059)	0.851 (0.773 - 0.929)	0.074
-19	0.774 (0.688 - 0.860)	0.769 (0.682 - 0.856)	0.717 (0.624 - 0.810)	0.022 (0.000 - 0.052)	0.840 (0.765 - 0.915)	0.065

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-18	0.774 (0.685 - 0.863)	0.706 (0.609 - 0.803)	0.724 (0.629 - 0.819)	0.026 (0.000 - 0.060)	0.857 (0.783 - 0.931)	0.071
-17	0.799 (0.721 - 0.877)	0.782 (0.701 - 0.863)	0.700 (0.611 - 0.789)	0.024 (0.000 - 0.054)	0.830 (0.757 - 0.903)	0.063
-16	0.787 (0.707 - 0.867)	0.670 (0.578 - 0.762)	0.799 (0.720 - 0.878)	0.032 (0.000 - 0.066)	0.790 (0.710 - 0.870)	0.079
-15	0.592 (0.494 - 0.690)	0.451 (0.352 - 0.550)	0.728 (0.639 - 0.817)	0.070 (0.019 - 0.121)	0.858 (0.789 - 0.927)	0.048
-14	0.808 (0.730 - 0.886)	0.758 (0.674 - 0.842)	0.726 (0.638 - 0.814)	0.025 (0.000 - 0.056)	0.826 (0.751 - 0.901)	0.065
-13	0.808 (0.732 - 0.884)	0.716 (0.628 - 0.804)	0.771 (0.689 - 0.853)	0.028 (0.000 - 0.060)	0.803 (0.726 - 0.880)	0.075
-12	0.785 (0.704 - 0.866)	0.660 (0.567 - 0.753)	0.829 (0.755 - 0.903)	0.031 (0.000 - 0.065)	0.766 (0.683 - 0.849)	0.091
-11	0.812 (0.740 - 0.884)	0.777 (0.700 - 0.854)	0.701 (0.616 - 0.786)	0.027 (0.000 - 0.057)	0.817 (0.745 - 0.889)	0.063
-10	0.802 (0.728 - 0.876)	0.700 (0.614 - 0.786)	0.786 (0.709 - 0.863)	0.031 (0.000 - 0.063)	0.784 (0.707 - 0.861)	0.081
-9	0.798 (0.726 - 0.870)	0.644 (0.558 - 0.730)	0.811 (0.740 - 0.882)	0.038 (0.004 - 0.072)	0.763 (0.686 - 0.840)	0.091
-8	0.775 (0.696 - 0.854)	0.787 (0.709 - 0.865)	0.607 (0.514 - 0.700)	0.028 (0.000 - 0.059)	0.858 (0.792 - 0.924)	0.050
-7	0.809 (0.736 - 0.882)	0.631 (0.541 - 0.721)	0.813 (0.740 - 0.886)	0.037 (0.002 - 0.072)	0.777 (0.700 - 0.854)	0.089
-6	0.803 (0.729 - 0.877)	0.874 (0.812 - 0.936)	0.577 (0.485 - 0.669)	0.018 (0.000 - 0.043)	0.850 (0.784 - 0.916)	0.051
-5	0.805 (0.734 - 0.876)	0.750 (0.673 - 0.827)	0.724 (0.644 - 0.804)	0.026 (0.000 - 0.054)	0.827 (0.760 - 0.894)	0.068
-4	0.798 (0.737 - 0.859)	0.653 (0.581 - 0.725)	0.825 (0.767 - 0.883)	0.038 (0.009 - 0.067)	0.741 (0.675 - 0.807)	0.090
-3	0.807 (0.747 - 0.867)	0.736 (0.669 - 0.803)	0.774 (0.711 - 0.837)	0.031 (0.005 - 0.057)	0.767 (0.703 - 0.831)	0.075
-2	0.824 (0.755 - 0.893)	0.694 (0.611 - 0.777)	0.826 (0.758 - 0.894)	0.033 (0.001 - 0.065)	0.729 (0.649 - 0.809)	0.083
-1	0.886 (0.828 - 0.944)	0.750 (0.671 - 0.829)	0.873 (0.812 - 0.934)	0.026 (0.000 - 0.055)	0.646 (0.559 - 0.733)	0.098
Mean ± SD	**0.792 ± 0.049**	**0.707 ± 0.078**	**0.761 ± 0.072**	**0.030 ± 0.010**	**0.800 ± 0.053**	

Table 7.3.D. Performance measures for each model at each lead time when predicting AKI within 25h since ICU admission, using 0 imputation in testing data.

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**AdaBoost**					
-24	0.782 (0.687 - 0.877)	0.740 (0.639 - 0.841)	0.685 (0.578 - 0.792)	0.022 (0.000 - 0.056)	0.878 (0.803 - 0.953)	0.170
-23	0.807 (0.724 - 0.890)	0.713 (0.618 - 0.808)	0.812 (0.730 - 0.894)	0.023 (0.000 - 0.054)	0.797 (0.712 - 0.882)	0.220
-22	0.839 (0.761 - 0.917)	0.872 (0.801 - 0.943)	0.668 (0.568 - 0.768)	0.012 (0.000 - 0.035)	0.852 (0.777 - 0.927)	0.172
-21	0.812 (0.730 - 0.894)	0.648 (0.548 - 0.748)	0.839 (0.762 - 0.916)	0.028 (0.000 - 0.062)	0.786 (0.700 - 0.872)	0.231
-20	0.771 (0.679 - 0.863)	0.768 (0.676 - 0.860)	0.682 (0.580 - 0.784)	0.021 (0.000 - 0.052)	0.868 (0.794 - 0.942)	0.172
-19	0.792 (0.709 - 0.875)	0.714 (0.621 - 0.807)	0.735 (0.644 - 0.826)	0.027 (0.000 - 0.060)	0.841 (0.766 - 0.916)	0.195
-18	0.789 (0.702 - 0.876)	0.765 (0.675 - 0.855)	0.714 (0.618 - 0.810)	0.021 (0.000 - 0.051)	0.852 (0.777 - 0.927)	0.200
-17	0.813 (0.737 - 0.889)	0.782 (0.701 - 0.863)	0.745 (0.660 - 0.830)	0.022 (0.000 - 0.051)	0.807 (0.730 - 0.884)	0.198
-16	0.781 (0.700 - 0.862)	0.767 (0.684 - 0.850)	0.690 (0.599 - 0.781)	0.026 (0.000 - 0.057)	0.835 (0.762 - 0.908)	0.175
-15	0.770 (0.686 - 0.854)	0.649 (0.554 - 0.744)	0.769 (0.685 - 0.853)	0.033 (0.000 - 0.069)	0.826 (0.751 - 0.901)	0.203
-14	0.813 (0.736 - 0.890)	0.727 (0.639 - 0.815)	0.810 (0.733 - 0.887)	0.025 (0.000 - 0.056)	0.774 (0.692 - 0.856)	0.225
-13	0.807 (0.730 - 0.884)	0.686 (0.596 - 0.776)	0.780 (0.700 - 0.860)	0.031 (0.000 - 0.065)	0.803 (0.726 - 0.880)	0.220
-12	0.836 (0.763 - 0.909)	0.738 (0.652 - 0.824)	0.793 (0.714 - 0.872)	0.025 (0.000 - 0.056)	0.780 (0.699 - 0.861)	0.227
-11	0.814 (0.742 - 0.886)	0.714 (0.630 - 0.798)	0.785 (0.709 - 0.861)	0.030 (0.000 - 0.062)	0.777 (0.700 - 0.854)	0.227
-10	0.814 (0.741 - 0.887)	0.818 (0.746 - 0.890)	0.674 (0.586 - 0.762)	0.022 (0.000 - 0.049)	0.825 (0.754 - 0.896)	0.192
-9	0.774 (0.699 - 0.849)	0.712 (0.630 - 0.794)	0.725 (0.644 - 0.806)	0.035 (0.002 - 0.068)	0.809 (0.738 - 0.880)	0.219
-8	0.807 (0.732 - 0.882)	0.667 (0.578 - 0.756)	0.783 (0.705 - 0.861)	0.034 (0.000 - 0.068)	0.797 (0.721 - 0.873)	0.223
-7	0.814 (0.742 - 0.886)	0.748 (0.667 - 0.829)	0.747 (0.666 - 0.828)	0.028 (0.000 - 0.059)	0.799 (0.724 - 0.874)	0.222
-6	0.782 (0.705 - 0.859)	0.559 (0.467 - 0.651)	0.860 (0.795 - 0.925)	0.042 (0.005 - 0.079)	0.747 (0.666 - 0.828)	0.281
-5	0.795 (0.723 - 0.867)	0.653 (0.568 - 0.738)	0.801 (0.730 - 0.872)	0.039 (0.005 - 0.073)	0.765 (0.689 - 0.841)	0.253
-4	0.785 (0.723 - 0.847)	0.645 (0.572 - 0.718)	0.784 (0.722 - 0.846)	0.041 (0.011 - 0.071)	0.782 (0.719 - 0.845)	0.247
-3	0.795 (0.734 - 0.856)	0.694 (0.624 - 0.764)	0.756 (0.691 - 0.821)	0.036 (0.008 - 0.064)	0.790 (0.728 - 0.852)	0.237
-2	0.805 (0.734 - 0.876)	0.711 (0.629 - 0.793)	0.792 (0.719 - 0.865)	0.033 (0.001 - 0.065)	0.758 (0.681 - 0.835)	0.242
-1	0.873 (0.812 - 0.934)	0.842 (0.776 - 0.908)	0.795 (0.722 - 0.868)	0.018 (0.000 - 0.042)	0.725 (0.644 - 0.806)	0.206
Mean ± SD	**0.803 ± 0.024**	**0.722 ± 0.069**	**0.759 ± 0.053**	**0.028 ± 0.008**	**0.803 ± 0.038**	
Model	**BARTm**					
-24	0.874 (0.798 - 0.950)	0.870 (0.793 - 0.947)	0.766 (0.669 - 0.863)	0.010 (0.000 - 0.033)	0.820 (0.732 - 0.908)	0.062
-23	0.867 (0.796 - 0.938)	0.759 (0.669 - 0.849)	0.818 (0.737 - 0.899)	0.019 (0.000 - 0.048)	0.781 (0.694 - 0.868)	0.085
-22	0.868 (0.796 - 0.940)	0.860 (0.787 - 0.933)	0.719 (0.624 - 0.814)	0.013 (0.000 - 0.037)	0.831 (0.752 - 0.910)	0.057
-21	0.846 (0.771 - 0.921)	0.818 (0.737 - 0.899)	0.732 (0.639 - 0.825)	0.017 (0.000 - 0.044)	0.829 (0.750 - 0.908)	0.058

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-20	0.828 (0.745 - 0.911)	0.854 (0.777 - 0.931)	0.700 (0.600 - 0.800)	0.013 (0.000 - 0.038)	0.848 (0.769 - 0.927)	0.048
-19	0.838 (0.762 - 0.914)	0.813 (0.733 - 0.893)	0.733 (0.642 - 0.824)	0.018 (0.000 - 0.045)	0.824 (0.746 - 0.902)	0.064
-18	0.842 (0.764 - 0.920)	0.812 (0.729 - 0.895)	0.770 (0.681 - 0.859)	0.016 (0.000 - 0.043)	0.814 (0.731 - 0.897)	0.078
-17	0.845 (0.774 - 0.916)	0.832 (0.759 - 0.905)	0.768 (0.686 - 0.850)	0.017 (0.000 - 0.042)	0.781 (0.700 - 0.862)	0.065
-16	0.841 (0.769 - 0.913)	0.874 (0.809 - 0.939)	0.689 (0.598 - 0.780)	0.014 (0.000 - 0.037)	0.817 (0.741 - 0.893)	0.047
-15	0.816 (0.739 - 0.893)	0.773 (0.690 - 0.856)	0.748 (0.662 - 0.834)	0.022 (0.000 - 0.051)	0.814 (0.737 - 0.891)	0.058
-14	0.848 (0.777 - 0.919)	0.717 (0.628 - 0.806)	0.850 (0.780 - 0.920)	0.025 (0.000 - 0.056)	0.733 (0.646 - 0.820)	0.115
-13	0.868 (0.802 - 0.934)	0.863 (0.796 - 0.930)	0.721 (0.634 - 0.808)	0.015 (0.000 - 0.039)	0.805 (0.728 - 0.882)	0.069
-12	0.852 (0.782 - 0.922)	0.746 (0.661 - 0.831)	0.838 (0.766 - 0.910)	0.023 (0.000 - 0.052)	0.733 (0.646 - 0.820)	0.108
-11	0.855 (0.790 - 0.920)	0.768 (0.690 - 0.846)	0.803 (0.729 - 0.877)	0.024 (0.000 - 0.052)	0.748 (0.668 - 0.828)	0.088
-10	0.850 (0.783 - 0.917)	0.773 (0.695 - 0.851)	0.792 (0.716 - 0.868)	0.024 (0.000 - 0.053)	0.761 (0.681 - 0.841)	0.089
-9	0.841 (0.775 - 0.907)	0.763 (0.686 - 0.840)	0.777 (0.702 - 0.852)	0.027 (0.000 - 0.056)	0.762 (0.685 - 0.839)	0.090
-8	0.818 (0.745 - 0.891)	0.769 (0.689 - 0.849)	0.726 (0.641 - 0.811)	0.026 (0.000 - 0.056)	0.812 (0.738 - 0.886)	0.063
-7	0.843 (0.775 - 0.911)	0.820 (0.749 - 0.891)	0.725 (0.642 - 0.808)	0.021 (0.000 - 0.048)	0.798 (0.723 - 0.873)	0.061
-6	0.848 (0.781 - 0.915)	0.847 (0.780 - 0.914)	0.701 (0.616 - 0.786)	0.018 (0.000 - 0.043)	0.806 (0.732 - 0.880)	0.052
-5	0.841 (0.776 - 0.906)	0.744 (0.666 - 0.822)	0.784 (0.711 - 0.857)	0.030 (0.000 - 0.060)	0.756 (0.679 - 0.833)	0.089
-4	0.862 (0.810 - 0.914)	0.843 (0.788 - 0.898)	0.716 (0.648 - 0.784)	0.020 (0.000 - 0.041)	0.783 (0.720 - 0.846)	0.058
-3	0.846 (0.791 - 0.901)	0.843 (0.788 - 0.898)	0.691 (0.621 - 0.761)	0.021 (0.000 - 0.043)	0.797 (0.736 - 0.858)	0.054
-2	0.844 (0.779 - 0.909)	0.752 (0.674 - 0.830)	0.807 (0.736 - 0.878)	0.028 (0.000 - 0.058)	0.733 (0.653 - 0.813)	0.085
-1	0.899 (0.844 - 0.954)	0.850 (0.785 - 0.915)	0.807 (0.735 - 0.879)	0.017 (0.000 - 0.041)	0.710 (0.627 - 0.793)	0.069
Mean ± SD	**0.849 ± 0.018**	**0.807 ± 0.048**	**0.758 ± 0.047**	**0.020 ± 0.005**	**0.787 ± 0.037**	
Model	**C5.0**					
-24	0.787 (0.693 - 0.881)	0.610 (0.498 - 0.722)	0.854 (0.773 - 0.935)	0.026 (0.000 - 0.063)	0.803 (0.712 - 0.894)	0.159
-23	0.797 (0.712 - 0.882)	0.667 (0.568 - 0.766)	0.820 (0.739 - 0.901)	0.027 (0.000 - 0.061)	0.801 (0.717 - 0.885)	0.154
-22	0.792 (0.706 - 0.878)	0.756 (0.665 - 0.847)	0.701 (0.604 - 0.798)	0.023 (0.000 - 0.055)	0.857 (0.783 - 0.931)	0.083
-21	0.808 (0.726 - 0.890)	0.636 (0.535 - 0.737)	0.859 (0.786 - 0.932)	0.028 (0.000 - 0.062)	0.766 (0.678 - 0.854)	0.180
-20	0.786 (0.696 - 0.876)	0.902 (0.837 - 0.967)	0.560 (0.451 - 0.669)	0.011 (0.000 - 0.034)	0.886 (0.816 - 0.956)	0.066
-19	0.780 (0.695 - 0.865)	0.681 (0.585 - 0.777)	0.814 (0.734 - 0.894)	0.027 (0.000 - 0.060)	0.795 (0.712 - 0.878)	0.163
-18	0.774 (0.685 - 0.863)	0.682 (0.583 - 0.781)	0.792 (0.706 - 0.878)	0.025 (0.000 - 0.058)	0.824 (0.743 - 0.905)	0.157
-17	0.805 (0.728 - 0.882)	0.792 (0.713 - 0.871)	0.694 (0.604 - 0.784)	0.023 (0.000 - 0.052)	0.832 (0.759 - 0.905)	0.076
-16	0.790 (0.710 - 0.870)	0.748 (0.663 - 0.833)	0.758 (0.674 - 0.842)	0.026 (0.000 - 0.057)	0.803 (0.725 - 0.881)	0.095
-15	0.755 (0.669 - 0.841)	0.577 (0.479 - 0.675)	0.854 (0.784 - 0.924)	0.036 (0.000 - 0.073)	0.772 (0.689 - 0.855)	0.157
-14	0.716 (0.627 - 0.805)	0.515 (0.417 - 0.613)	0.881 (0.817 - 0.945)	0.040 (0.001 - 0.079)	0.752 (0.667 - 0.837)	0.238

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-13	0.786 (0.706 - 0.866)	0.696 (0.607 - 0.785)	0.775 (0.694 - 0.856)	0.030 (0.000 - 0.063)	0.805 (0.728 - 0.882)	0.161
-12	0.824 (0.749 - 0.899)	0.738 (0.652 - 0.824)	0.803 (0.725 - 0.881)	0.025 (0.000 - 0.056)	0.771 (0.689 - 0.853)	0.147
-11	0.783 (0.707 - 0.859)	0.732 (0.650 - 0.814)	0.756 (0.676 - 0.836)	0.030 (0.000 - 0.062)	0.794 (0.719 - 0.869)	0.133
-10	0.795 (0.720 - 0.870)	0.664 (0.576 - 0.752)	0.816 (0.744 - 0.888)	0.034 (0.000 - 0.068)	0.767 (0.688 - 0.846)	0.157
-9	0.771 (0.695 - 0.847)	0.771 (0.695 - 0.847)	0.682 (0.598 - 0.766)	0.030 (0.000 - 0.061)	0.819 (0.750 - 0.888)	0.086
-8	0.774 (0.695 - 0.853)	0.704 (0.618 - 0.790)	0.755 (0.674 - 0.836)	0.031 (0.000 - 0.064)	0.808 (0.733 - 0.883)	0.149
-7	0.781 (0.704 - 0.858)	0.748 (0.667 - 0.829)	0.722 (0.639 - 0.805)	0.029 (0.000 - 0.060)	0.814 (0.742 - 0.886)	0.098
-6	0.783 (0.706 - 0.860)	0.649 (0.560 - 0.738)	0.827 (0.757 - 0.897)	0.035 (0.001 - 0.069)	0.758 (0.678 - 0.838)	0.171
-5	0.792 (0.720 - 0.864)	0.636 (0.550 - 0.722)	0.865 (0.804 - 0.926)	0.038 (0.004 - 0.072)	0.694 (0.612 - 0.776)	0.188
-4	0.797 (0.736 - 0.858)	0.727 (0.659 - 0.795)	0.747 (0.681 - 0.813)	0.033 (0.006 - 0.060)	0.788 (0.726 - 0.850)	0.152
-3	0.807 (0.747 - 0.867)	0.727 (0.659 - 0.795)	0.784 (0.722 - 0.846)	0.031 (0.005 - 0.057)	0.760 (0.695 - 0.825)	0.164
-2	0.792 (0.719 - 0.865)	0.752 (0.674 - 0.830)	0.715 (0.634 - 0.796)	0.031 (0.000 - 0.062)	0.802 (0.730 - 0.874)	0.095
-1	0.833 (0.765 - 0.901)	0.767 (0.690 - 0.844)	0.791 (0.717 - 0.865)	0.027 (0.000 - 0.056)	0.747 (0.668 - 0.826)	0.156
Mean ± SD	**0.788 ± 0.023**	**0.703 ± 0.079**	**0.776 ± 0.073**	**0.029 ± 0.006**	**0.792 ± 0.039**	
Model	**GBM**					
-24	0.846 (0.763 - 0.929)	0.805 (0.714 - 0.896)	0.782 (0.687 - 0.877)	0.014 (0.000 - 0.041)	0.821 (0.733 - 0.909)	0.072
-23	0.860 (0.787 - 0.933)	0.759 (0.669 - 0.849)	0.834 (0.756 - 0.912)	0.019 (0.000 - 0.048)	0.766 (0.677 - 0.855)	0.103
-22	0.862 (0.789 - 0.935)	0.849 (0.773 - 0.925)	0.749 (0.657 - 0.841)	0.013 (0.000 - 0.037)	0.817 (0.735 - 0.899)	0.087
-21	0.843 (0.767 - 0.919)	0.864 (0.792 - 0.936)	0.705 (0.610 - 0.800)	0.013 (0.000 - 0.037)	0.835 (0.757 - 0.913)	0.073
-20	0.817 (0.732 - 0.902)	0.866 (0.791 - 0.941)	0.651 (0.547 - 0.755)	0.013 (0.000 - 0.038)	0.865 (0.790 - 0.940)	0.079
-19	0.788 (0.704 - 0.872)	0.681 (0.585 - 0.777)	0.788 (0.704 - 0.872)	0.028 (0.000 - 0.062)	0.816 (0.736 - 0.896)	0.101
-18	0.815 (0.732 - 0.898)	0.812 (0.729 - 0.895)	0.692 (0.594 - 0.790)	0.017 (0.000 - 0.044)	0.854 (0.779 - 0.929)	0.151
-17	0.833 (0.760 - 0.906)	0.842 (0.771 - 0.913)	0.676 (0.585 - 0.767)	0.018 (0.000 - 0.044)	0.831 (0.758 - 0.904)	0.066
-16	0.818 (0.742 - 0.894)	0.816 (0.740 - 0.892)	0.690 (0.599 - 0.781)	0.021 (0.000 - 0.049)	0.826 (0.752 - 0.900)	0.061
-15	0.802 (0.723 - 0.881)	0.825 (0.749 - 0.901)	0.627 (0.531 - 0.723)	0.020 (0.000 - 0.048)	0.858 (0.789 - 0.927)	0.081
-14	0.842 (0.770 - 0.914)	0.737 (0.650 - 0.824)	0.818 (0.742 - 0.894)	0.024 (0.000 - 0.054)	0.765 (0.681 - 0.849)	0.121
-13	0.844 (0.774 - 0.914)	0.676 (0.585 - 0.767)	0.863 (0.796 - 0.930)	0.029 (0.000 - 0.062)	0.721 (0.634 - 0.808)	0.154
-12	0.846 (0.775 - 0.917)	0.709 (0.620 - 0.798)	0.831 (0.758 - 0.904)	0.027 (0.000 - 0.059)	0.751 (0.666 - 0.836)	0.158
-11	0.854 (0.789 - 0.919)	0.732 (0.650 - 0.814)	0.835 (0.766 - 0.904)	0.027 (0.000 - 0.057)	0.723 (0.640 - 0.806)	0.129
-10	0.842 (0.774 - 0.910)	0.700 (0.614 - 0.786)	0.849 (0.782 - 0.916)	0.029 (0.000 - 0.060)	0.719 (0.635 - 0.803)	0.138
-9	0.805 (0.734 - 0.876)	0.780 (0.705 - 0.855)	0.722 (0.641 - 0.803)	0.027 (0.000 - 0.056)	0.796 (0.723 - 0.869)	0.104
-8	0.821 (0.748 - 0.894)	0.778 (0.699 - 0.857)	0.717 (0.632 - 0.802)	0.025 (0.000 - 0.055)	0.715 (0.629 - 0.801)	0.106
-7	0.842 (0.774 - 0.910)	0.802 (0.728 - 0.876)	0.741 (0.660 - 0.822)	0.022 (0.000 - 0.049)	0.792 (0.716 - 0.868)	0.093

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-6	0.828 (0.758 - 0.898)	0.694 (0.608 - 0.780)	0.826 (0.755 - 0.897)	0.031 (0.000 - 0.063)	0.747 (0.666 - 0.828)	0.143
-5	0.821 (0.753 - 0.889)	0.669 (0.585 - 0.753)	0.837 (0.771 - 0.903)	0.036 (0.003 - 0.069)	0.723 (0.643 - 0.803)	0.146
-4	0.824 (0.766 - 0.882)	0.628 (0.555 - 0.701)	0.861 (0.809 - 0.913)	0.039 (0.010 - 0.068)	0.703 (0.634 - 0.772)	0.153
-3	0.827 (0.770 - 0.884)	0.669 (0.598 - 0.740)	0.846 (0.791 - 0.901)	0.035 (0.007 - 0.063)	0.711 (0.642 - 0.780)	0.139
-2	0.833 (0.766 - 0.900)	0.678 (0.594 - 0.762)	0.853 (0.789 - 0.917)	0.034 (0.001 - 0.067)	0.700 (0.617 - 0.783)	0.178
-1	0.885 (0.827 - 0.943)	0.858 (0.794 - 0.922)	0.775 (0.699 - 0.851)	0.017 (0.000 - 0.041)	0.739 (0.659 - 0.819)	0.099
Mean ± SD	**0.833 ± 0.021**	**0.760 ± 0.074**	**0.774 ± 0.074**	**0.024 ± 0.008**	**0.775 ± 0.055**	
Model	**LR**					
-24	0.806 (0.715 - 0.897)	0.649 (0.540 - 0.758)	0.826 (0.739 - 0.913)	0.024 (0.000 - 0.059)	0.819 (0.731 - 0.907)	0.072
-23	0.844 (0.768 - 0.920)	0.851 (0.776 - 0.926)	0.745 (0.653 - 0.837)	0.013 (0.000 - 0.037)	0.817 (0.736 - 0.898)	0.048
-22	0.801 (0.717 - 0.885)	0.756 (0.665 - 0.847)	0.772 (0.683 - 0.861)	0.020 (0.000 - 0.050)	0.820 (0.739 - 0.901)	0.048
-21	0.602 (0.500 - 0.704)	0.239 (0.150 - 0.328)	0.966 (0.928 - 1.004)	0.051 (0.005 - 0.097)	0.677 (0.579 - 0.775)	0.001
-20	0.570 (0.462 - 0.678)	0.171 (0.088 - 0.254)	0.970 (0.933 - 1.007)	0.051 (0.003 - 0.099)	0.736 (0.639 - 0.833)	0.001
-19	0.613 (0.513 - 0.713)	0.308 (0.213 - 0.403)	0.918 (0.862 - 0.974)	0.050 (0.005 - 0.095)	0.791 (0.707 - 0.875)	0.001
-18	0.778 (0.690 - 0.866)	0.776 (0.687 - 0.865)	0.722 (0.627 - 0.817)	0.020 (0.000 - 0.050)	0.846 (0.769 - 0.923)	0.040
-17	0.823 (0.749 - 0.897)	0.842 (0.771 - 0.913)	0.721 (0.634 - 0.808)	0.017 (0.000 - 0.042)	0.809 (0.732 - 0.886)	0.039
-16	0.825 (0.751 - 0.899)	0.777 (0.695 - 0.859)	0.767 (0.684 - 0.850)	0.023 (0.000 - 0.052)	0.790 (0.710 - 0.870)	0.049
-15	0.781 (0.699 - 0.863)	0.732 (0.644 - 0.820)	0.725 (0.636 - 0.814)	0.027 (0.000 - 0.059)	0.834 (0.760 - 0.908)	0.040
-14	0.837 (0.764 - 0.910)	0.828 (0.754 - 0.902)	0.747 (0.661 - 0.833)	0.017 (0.000 - 0.042)	0.800 (0.721 - 0.879)	0.050
-13	0.851 (0.782 - 0.920)	0.882 (0.819 - 0.945)	0.695 (0.606 - 0.784)	0.013 (0.000 - 0.035)	0.815 (0.740 - 0.890)	0.039
-12	0.849 (0.779 - 0.919)	0.718 (0.630 - 0.806)	0.863 (0.796 - 0.930)	0.025 (0.000 - 0.056)	0.708 (0.619 - 0.797)	0.094
-11	0.843 (0.776 - 0.910)	0.741 (0.660 - 0.822)	0.782 (0.706 - 0.858)	0.028 (0.000 - 0.059)	0.773 (0.695 - 0.851)	0.051
-10	0.845 (0.777 - 0.913)	0.845 (0.777 - 0.913)	0.724 (0.640 - 0.808)	0.018 (0.000 - 0.043)	0.794 (0.718 - 0.870)	0.040
-9	0.865 (0.803 - 0.927)	0.814 (0.744 - 0.884)	0.770 (0.694 - 0.846)	0.022 (0.000 - 0.048)	0.756 (0.679 - 0.833)	0.055
-8	0.830 (0.759 - 0.901)	0.889 (0.829 - 0.949)	0.637 (0.546 - 0.728)	0.014 (0.000 - 0.036)	0.832 (0.761 - 0.903)	0.026
-7	0.812 (0.739 - 0.885)	0.874 (0.812 - 0.936)	0.653 (0.564 - 0.742)	0.016 (0.000 - 0.039)	0.823 (0.752 - 0.894)	0.024
-6	0.824 (0.753 - 0.895)	0.820 (0.749 - 0.891)	0.715 (0.631 - 0.799)	0.021 (0.000 - 0.048)	0.803 (0.729 - 0.877)	0.037
-5	0.830 (0.763 - 0.897)	0.785 (0.712 - 0.858)	0.761 (0.685 - 0.837)	0.026 (0.000 - 0.054)	0.765 (0.689 - 0.841)	0.069
-4	0.848 (0.794 - 0.902)	0.769 (0.705 - 0.833)	0.801 (0.740 - 0.862)	0.026 (0.002 - 0.050)	0.734 (0.667 - 0.801)	0.075
-3	0.856 (0.803 - 0.909)	0.868 (0.817 - 0.919)	0.716 (0.648 - 0.784)	0.017 (0.000 - 0.037)	0.778 (0.715 - 0.841)	0.042
-2	0.854 (0.790 - 0.918)	0.810 (0.739 - 0.881)	0.751 (0.673 - 0.829)	0.023 (0.000 - 0.050)	0.767 (0.691 - 0.843)	0.044
-1	0.893 (0.837 - 0.949)	0.742 (0.662 - 0.822)	0.911 (0.859 - 0.963)	0.026 (0.000 - 0.055)	0.564 (0.474 - 0.654)	0.098
Mean ± SD	**0.803 ± 0.085**	**0.729 ± 0.199**	**0.777 ± 0.090**	**0.025 ± 0.011**	**0.777 ± 0.062**	

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**RF**					
-24	0.803 (0.712 - 0.894)	0.688 (0.582 - 0.794)	0.788 (0.694 - 0.882)	0.023 (0.000 - 0.057)	0.839 (0.755 - 0.923)	0.115
-23	0.819 (0.738 - 0.900)	0.828 (0.749 - 0.907)	0.702 (0.606 - 0.798)	0.016 (0.000 - 0.042)	0.843 (0.767 - 0.919)	0.085
-22	0.800 (0.715 - 0.885)	0.802 (0.718 - 0.886)	0.688 (0.590 - 0.786)	0.019 (0.000 - 0.048)	0.855 (0.781 - 0.929)	0.105
-21	0.813 (0.732 - 0.894)	0.659 (0.560 - 0.758)	0.823 (0.743 - 0.903)	0.027 (0.000 - 0.061)	0.799 (0.715 - 0.883)	0.115
-20	0.791 (0.702 - 0.880)	0.756 (0.662 - 0.850)	0.702 (0.602 - 0.802)	0.021 (0.000 - 0.052)	0.863 (0.788 - 0.938)	0.085
-19	0.790 (0.706 - 0.874)	0.802 (0.720 - 0.884)	0.647 (0.549 - 0.745)	0.021 (0.000 - 0.050)	0.863 (0.792 - 0.934)	0.075
-18	0.804 (0.720 - 0.888)	0.753 (0.661 - 0.845)	0.704 (0.607 - 0.801)	0.022 (0.000 - 0.053)	0.858 (0.784 - 0.932)	0.090
-17	0.807 (0.730 - 0.884)	0.634 (0.540 - 0.728)	0.850 (0.780 - 0.920)	0.033 (0.000 - 0.068)	0.752 (0.668 - 0.836)	0.130
-16	0.804 (0.726 - 0.882)	0.670 (0.578 - 0.762)	0.812 (0.735 - 0.889)	0.031 (0.000 - 0.065)	0.779 (0.698 - 0.860)	0.115
-15	0.801 (0.722 - 0.880)	0.825 (0.749 - 0.901)	0.611 (0.514 - 0.708)	0.021 (0.000 - 0.050)	0.863 (0.795 - 0.931)	0.065
-14	0.798 (0.719 - 0.877)	0.768 (0.685 - 0.851)	0.712 (0.623 - 0.801)	0.024 (0.000 - 0.054)	0.831 (0.757 - 0.905)	0.110
-13	0.825 (0.751 - 0.899)	0.833 (0.761 - 0.905)	0.653 (0.561 - 0.745)	0.020 (0.000 - 0.047)	0.841 (0.770 - 0.912)	0.080
-12	0.830 (0.756 - 0.904)	0.864 (0.797 - 0.931)	0.676 (0.584 - 0.768)	0.016 (0.000 - 0.041)	0.826 (0.752 - 0.900)	0.105
-11	0.817 (0.745 - 0.889)	0.696 (0.611 - 0.781)	0.827 (0.757 - 0.897)	0.031 (0.000 - 0.063)	0.742 (0.661 - 0.823)	0.170
-10	0.798 (0.723 - 0.873)	0.691 (0.605 - 0.777)	0.796 (0.721 - 0.871)	0.032 (0.000 - 0.065)	0.777 (0.699 - 0.855)	0.160
-9	0.775 (0.700 - 0.850)	0.746 (0.667 - 0.825)	0.705 (0.623 - 0.787)	0.032 (0.000 - 0.064)	0.813 (0.743 - 0.883)	0.130
-8	0.761 (0.680 - 0.842)	0.759 (0.678 - 0.840)	0.675 (0.586 - 0.764)	0.029 (0.000 - 0.061)	0.838 (0.768 - 0.908)	0.105
-7	0.780 (0.703 - 0.857)	0.775 (0.697 - 0.853)	0.706 (0.621 - 0.791)	0.026 (0.000 - 0.056)	0.817 (0.745 - 0.889)	0.125
-6	0.774 (0.696 - 0.852)	0.757 (0.677 - 0.837)	0.695 (0.609 - 0.781)	0.029 (0.000 - 0.060)	0.826 (0.755 - 0.897)	0.125
-5	0.779 (0.705 - 0.853)	0.702 (0.621 - 0.783)	0.735 (0.656 - 0.814)	0.036 (0.003 - 0.069)	0.801 (0.730 - 0.872)	0.145
-4	0.769 (0.705 - 0.833)	0.752 (0.687 - 0.817)	0.673 (0.602 - 0.744)	0.033 (0.006 - 0.060)	0.823 (0.765 - 0.881)	0.115
-3	0.765 (0.701 - 0.829)	0.653 (0.581 - 0.725)	0.762 (0.697 - 0.827)	0.041 (0.011 - 0.071)	0.796 (0.735 - 0.857)	0.165
-2	0.763 (0.686 - 0.840)	0.711 (0.629 - 0.793)	0.740 (0.661 - 0.819)	0.035 (0.002 - 0.068)	0.797 (0.724 - 0.870)	0.160
-1	0.816 (0.745 - 0.887)	0.808 (0.736 - 0.880)	0.683 (0.598 - 0.768)	0.025 (0.000 - 0.053)	0.809 (0.737 - 0.881)	0.095
Mean ± SD	**0.795 ± 0.020**	**0.747 ± 0.064**	**0.724 ± 0.063**	**0.027 ± 0.007**	**0.819 ± 0.034**	
Model	**SVM**					
-24	0.718 (0.615 - 0.821)	0.675 (0.568 - 0.782)	0.676 (0.569 - 0.783)	0.028 (0.000 - 0.066)	0.890 (0.818 - 0.962)	0.066
-23	0.776 (0.688 - 0.864)	0.770 (0.682 - 0.858)	0.696 (0.599 - 0.793)	0.022 (0.000 - 0.053)	0.855 (0.781 - 0.929)	0.072
-22	0.754 (0.663 - 0.845)	0.802 (0.718 - 0.886)	0.616 (0.513 - 0.719)	0.021 (0.000 - 0.051)	0.879 (0.810 - 0.948)	0.060
-21	0.755 (0.665 - 0.845)	0.716 (0.622 - 0.810)	0.688 (0.591 - 0.785)	0.027 (0.000 - 0.061)	0.865 (0.794 - 0.936)	0.071
-20	0.731 (0.634 - 0.828)	0.683 (0.581 - 0.785)	0.687 (0.585 - 0.789)	0.028 (0.000 - 0.064)	0.880 (0.809 - 0.951)	0.074
-19	0.730 (0.639 - 0.821)	0.736 (0.645 - 0.827)	0.677 (0.581 - 0.773)	0.027 (0.000 - 0.060)	0.862 (0.791 - 0.933)	0.072

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-18	0.747 (0.655 - 0.839)	0.765 (0.675 - 0.855)	0.632 (0.529 - 0.735)	0.024 (0.000 - 0.057)	0.881 (0.812 - 0.950)	0.067
-17	0.768 (0.686 - 0.850)	0.812 (0.736 - 0.888)	0.635 (0.541 - 0.729)	0.023 (0.000 - 0.052)	0.852 (0.783 - 0.921)	0.063
-16	0.740 (0.654 - 0.826)	0.650 (0.557 - 0.743)	0.741 (0.655 - 0.827)	0.036 (0.000 - 0.073)	0.833 (0.760 - 0.906)	0.081
-15	0.721 (0.632 - 0.810)	0.619 (0.522 - 0.716)	0.725 (0.636 - 0.814)	0.038 (0.000 - 0.076)	0.856 (0.786 - 0.926)	0.077
-14	0.767 (0.684 - 0.850)	0.788 (0.707 - 0.869)	0.657 (0.563 - 0.751)	0.024 (0.000 - 0.054)	0.851 (0.781 - 0.921)	0.065
-13	0.759 (0.676 - 0.842)	0.735 (0.649 - 0.821)	0.694 (0.605 - 0.783)	0.029 (0.000 - 0.062)	0.841 (0.770 - 0.912)	0.075
-12	0.732 (0.645 - 0.819)	0.777 (0.695 - 0.859)	0.645 (0.551 - 0.739)	0.027 (0.000 - 0.059)	0.852 (0.782 - 0.922)	0.067
-11	0.765 (0.686 - 0.844)	0.795 (0.720 - 0.870)	0.622 (0.532 - 0.712)	0.028 (0.000 - 0.059)	0.846 (0.779 - 0.913)	0.063
-10	0.745 (0.664 - 0.826)	0.755 (0.675 - 0.835)	0.653 (0.564 - 0.742)	0.031 (0.000 - 0.063)	0.845 (0.777 - 0.913)	0.073
-9	0.749 (0.671 - 0.827)	0.678 (0.594 - 0.762)	0.728 (0.648 - 0.808)	0.039 (0.004 - 0.074)	0.815 (0.745 - 0.885)	0.090
-8	0.714 (0.628 - 0.800)	0.657 (0.567 - 0.747)	0.676 (0.587 - 0.765)	0.040 (0.003 - 0.077)	0.656 (0.566 - 0.746)	0.068
-7	0.755 (0.675 - 0.835)	0.892 (0.834 - 0.950)	0.512 (0.419 - 0.605)	0.018 (0.000 - 0.043)	0.865 (0.801 - 0.929)	0.050
-6	0.742 (0.661 - 0.823)	0.874 (0.812 - 0.936)	0.513 (0.420 - 0.606)	0.020 (0.000 - 0.046)	0.867 (0.804 - 0.930)	0.051
-5	0.737 (0.659 - 0.815)	0.653 (0.568 - 0.738)	0.725 (0.645 - 0.805)	0.043 (0.007 - 0.079)	0.818 (0.749 - 0.887)	0.087
-4	0.744 (0.678 - 0.810)	0.727 (0.659 - 0.795)	0.665 (0.593 - 0.737)	0.037 (0.008 - 0.066)	0.831 (0.774 - 0.888)	0.070
-3	0.753 (0.688 - 0.818)	0.760 (0.695 - 0.825)	0.690 (0.620 - 0.760)	0.031 (0.005 - 0.057)	0.813 (0.754 - 0.872)	0.075
-2	0.766 (0.690 - 0.842)	0.686 (0.602 - 0.770)	0.742 (0.663 - 0.821)	0.038 (0.004 - 0.072)	0.801 (0.729 - 0.873)	0.083
-1	0.832 (0.764 - 0.900)	0.758 (0.680 - 0.836)	0.775 (0.699 - 0.851)	0.028 (0.000 - 0.058)	0.762 (0.685 - 0.839)	0.092
Mean ± SD	**0.750 ± 0.024**	**0.740 ± 0.070**	**0.670 ± 0.063**	**0.029 ± 0.007**	**0.838 ± 0.048**	

Table 7.3.E. Performance measures for each model at each lead time when predicting AKI within 25h since ICU admission, using missForest imputation in testing data.

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**AdaBoost**					
-24	0.793 (0.700 - 0.886)	0.740 (0.639 - 0.841)	0.707 (0.603 - 0.811)	0.021 (0.000 - 0.054)	0.870 (0.793 - 0.947)	0.174
-23	0.812 (0.730 - 0.894)	0.736 (0.643 - 0.829)	0.776 (0.688 - 0.864)	0.022 (0.000 - 0.053)	0.820 (0.739 - 0.901)	0.204
-22	0.834 (0.755 - 0.913)	0.884 (0.816 - 0.952)	0.664 (0.564 - 0.764)	0.011 (0.000 - 0.033)	0.852 (0.777 - 0.927)	0.172
-21	0.818 (0.737 - 0.899)	0.636 (0.535 - 0.737)	0.842 (0.766 - 0.918)	0.028 (0.000 - 0.062)	0.786 (0.700 - 0.872)	0.243
-20	0.783 (0.693 - 0.873)	0.780 (0.689 - 0.871)	0.682 (0.580 - 0.784)	0.020 (0.000 - 0.051)	0.867 (0.793 - 0.941)	0.173
-19	0.782 (0.697 - 0.867)	0.769 (0.682 - 0.856)	0.677 (0.581 - 0.773)	0.023 (0.000 - 0.054)	0.857 (0.785 - 0.929)	0.172
-18	0.788 (0.701 - 0.875)	0.729 (0.635 - 0.823)	0.748 (0.656 - 0.840)	0.023 (0.000 - 0.055)	0.841 (0.763 - 0.919)	0.200
-17	0.814 (0.738 - 0.890)	0.832 (0.759 - 0.905)	0.693 (0.603 - 0.783)	0.019 (0.000 - 0.046)	0.825 (0.751 - 0.899)	0.175
-16	0.781 (0.700 - 0.862)	0.718 (0.630 - 0.806)	0.748 (0.663 - 0.833)	0.029 (0.000 - 0.062)	0.815 (0.739 - 0.891)	0.196
-15	0.778 (0.695 - 0.861)	0.691 (0.599 - 0.783)	0.762 (0.677 - 0.847)	0.029 (0.000 - 0.062)	0.822 (0.746 - 0.898)	0.200
-14	0.819 (0.743 - 0.895)	0.737 (0.650 - 0.824)	0.829 (0.755 - 0.903)	0.024 (0.000 - 0.054)	0.753 (0.668 - 0.838)	0.228
-13	0.809 (0.733 - 0.885)	0.706 (0.618 - 0.794)	0.768 (0.686 - 0.850)	0.029 (0.000 - 0.062)	0.807 (0.730 - 0.884)	0.219
-12	0.842 (0.771 - 0.913)	0.748 (0.663 - 0.833)	0.794 (0.715 - 0.873)	0.025 (0.000 - 0.056)	0.777 (0.695 - 0.859)	0.227
-11	0.818 (0.747 - 0.889)	0.723 (0.640 - 0.806)	0.782 (0.706 - 0.858)	0.030 (0.000 - 0.062)	0.777 (0.700 - 0.854)	0.228
-10	0.831 (0.761 - 0.901)	0.827 (0.756 - 0.898)	0.691 (0.605 - 0.777)	0.021 (0.000 - 0.048)	0.815 (0.742 - 0.888)	0.182
-9	0.784 (0.710 - 0.858)	0.627 (0.540 - 0.714)	0.832 (0.765 - 0.899)	0.039 (0.004 - 0.074)	0.747 (0.669 - 0.825)	0.250
-8	0.813 (0.739 - 0.887)	0.926 (0.876 - 0.976)	0.536 (0.442 - 0.630)	0.011 (0.000 - 0.031)	0.858 (0.792 - 0.924)	0.126
-7	0.820 (0.749 - 0.891)	0.766 (0.687 - 0.845)	0.771 (0.693 - 0.849)	0.025 (0.000 - 0.054)	0.778 (0.701 - 0.855)	0.222
-6	0.799 (0.724 - 0.874)	0.640 (0.551 - 0.729)	0.822 (0.751 - 0.893)	0.036 (0.001 - 0.071)	0.766 (0.687 - 0.845)	0.245
-5	0.804 (0.733 - 0.875)	0.612 (0.525 - 0.699)	0.863 (0.802 - 0.924)	0.040 (0.005 - 0.075)	0.705 (0.624 - 0.786)	0.297
-4	0.812 (0.753 - 0.871)	0.686 (0.616 - 0.756)	0.798 (0.737 - 0.859)	0.035 (0.007 - 0.063)	0.759 (0.694 - 0.824)	0.224
-3	0.813 (0.754 - 0.872)	0.711 (0.642 - 0.780)	0.803 (0.743 - 0.863)	0.033 (0.006 - 0.060)	0.748 (0.682 - 0.814)	0.237
-2	0.824 (0.755 - 0.893)	0.678 (0.594 - 0.762)	0.853 (0.789 - 0.917)	0.034 (0.001 - 0.067)	0.700 (0.617 - 0.783)	0.272
-1	0.883 (0.825 - 0.941)	0.858 (0.794 - 0.922)	0.807 (0.735 - 0.879)	0.016 (0.000 - 0.039)	0.708 (0.625 - 0.791)	0.205
Mean ± SD	**0.811 ± 0.024**	**0.740 ± 0.081**	**0.760 ± 0.076**	**0.026 ± 0.008**	**0.794 ± 0.052**	
Model	**BARTm**					
-24	0.867 (0.789 - 0.945)	0.857 (0.777 - 0.937)	0.728 (0.626 - 0.830)	0.011 (0.000 - 0.035)	0.843 (0.760 - 0.926)	0.061
-23	0.864 (0.792 - 0.936)	0.793 (0.708 - 0.878)	0.772 (0.684 - 0.860)	0.018 (0.000 - 0.046)	0.811 (0.729 - 0.893)	0.085
-22	0.866 (0.794 - 0.938)	0.907 (0.846 - 0.968)	0.669 (0.570 - 0.768)	0.009 (0.000 - 0.029)	0.847 (0.771 - 0.923)	0.057
-21	0.843 (0.767 - 0.919)	0.670 (0.572 - 0.768)	0.861 (0.789 - 0.933)	0.025 (0.000 - 0.058)	0.754 (0.664 - 0.844)	0.144

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-20	0.826 (0.743 - 0.909)	0.817 (0.732 - 0.902)	0.721 (0.623 - 0.819)	0.016 (0.000 - 0.043)	0.845 (0.766 - 0.924)	0.065
-19	0.836 (0.760 - 0.912)	0.813 (0.733 - 0.893)	0.712 (0.619 - 0.805)	0.018 (0.000 - 0.045)	0.835 (0.759 - 0.911)	0.064
-18	0.842 (0.764 - 0.920)	0.824 (0.743 - 0.905)	0.742 (0.649 - 0.835)	0.015 (0.000 - 0.041)	0.828 (0.748 - 0.908)	0.074
-17	0.841 (0.770 - 0.912)	0.851 (0.782 - 0.920)	0.727 (0.640 - 0.814)	0.016 (0.000 - 0.040)	0.804 (0.727 - 0.881)	0.066
-16	0.839 (0.767 - 0.911)	0.825 (0.751 - 0.899)	0.738 (0.652 - 0.824)	0.019 (0.000 - 0.046)	0.799 (0.720 - 0.878)	0.066
-15	0.808 (0.730 - 0.886)	0.691 (0.599 - 0.783)	0.799 (0.719 - 0.879)	0.028 (0.000 - 0.061)	0.796 (0.716 - 0.876)	0.089
-14	0.843 (0.771 - 0.915)	0.747 (0.661 - 0.833)	0.829 (0.755 - 0.903)	0.023 (0.000 - 0.053)	0.751 (0.666 - 0.836)	0.115
-13	0.862 (0.795 - 0.929)	0.745 (0.660 - 0.830)	0.805 (0.728 - 0.882)	0.024 (0.000 - 0.054)	0.770 (0.688 - 0.852)	0.133
-12	0.851 (0.781 - 0.921)	0.777 (0.695 - 0.859)	0.806 (0.728 - 0.884)	0.021 (0.000 - 0.049)	0.760 (0.676 - 0.844)	0.108
-11	0.850 (0.784 - 0.916)	0.839 (0.771 - 0.907)	0.709 (0.625 - 0.793)	0.019 (0.000 - 0.044)	0.800 (0.726 - 0.874)	0.067
-10	0.841 (0.773 - 0.909)	0.736 (0.654 - 0.818)	0.780 (0.703 - 0.857)	0.028 (0.000 - 0.059)	0.779 (0.701 - 0.857)	0.111
-9	0.835 (0.768 - 0.902)	0.686 (0.602 - 0.770)	0.825 (0.756 - 0.894)	0.033 (0.001 - 0.065)	0.736 (0.656 - 0.816)	0.127
-8	0.817 (0.744 - 0.890)	0.731 (0.647 - 0.815)	0.751 (0.669 - 0.833)	0.029 (0.000 - 0.061)	0.804 (0.729 - 0.879)	0.080
-7	0.835 (0.766 - 0.904)	0.748 (0.667 - 0.829)	0.761 (0.682 - 0.840)	0.027 (0.000 - 0.057)	0.790 (0.714 - 0.866)	0.088
-6	0.845 (0.778 - 0.912)	0.865 (0.801 - 0.929)	0.665 (0.577 - 0.753)	0.017 (0.000 - 0.041)	0.820 (0.749 - 0.891)	0.052
-5	0.832 (0.765 - 0.899)	0.686 (0.603 - 0.769)	0.828 (0.761 - 0.895)	0.034 (0.002 - 0.066)	0.729 (0.650 - 0.808)	0.133
-4	0.859 (0.806 - 0.912)	0.760 (0.695 - 0.825)	0.793 (0.732 - 0.854)	0.027 (0.002 - 0.052)	0.744 (0.678 - 0.810)	0.098
-3	0.839 (0.783 - 0.895)	0.860 (0.807 - 0.913)	0.652 (0.580 - 0.724)	0.020 (0.000 - 0.041)	0.812 (0.753 - 0.871)	0.054
-2	0.839 (0.773 - 0.905)	0.793 (0.720 - 0.866)	0.744 (0.665 - 0.823)	0.025 (0.000 - 0.053)	0.776 (0.701 - 0.851)	0.077
-1	0.897 (0.842 - 0.952)	0.825 (0.756 - 0.894)	0.827 (0.758 - 0.896)	0.019 (0.000 - 0.044)	0.693 (0.609 - 0.777)	0.087
Mean ± SD	**0.845 ± 0.018**	**0.785 ± 0.065**	**0.760 ± 0.057**	**0.022 ± 0.006**	**0.789 ± 0.040**	
Model	**C5.0**					
-24	0.777 (0.682 - 0.872)	0.610 (0.498 - 0.722)	0.829 (0.743 - 0.915)	0.027 (0.000 - 0.064)	0.826 (0.739 - 0.913)	0.164
-23	0.802 (0.718 - 0.886)	0.736 (0.643 - 0.829)	0.789 (0.703 - 0.875)	0.022 (0.000 - 0.053)	0.811 (0.729 - 0.893)	0.153
-22	0.803 (0.719 - 0.887)	0.791 (0.705 - 0.877)	0.696 (0.599 - 0.793)	0.019 (0.000 - 0.048)	0.853 (0.778 - 0.928)	0.083
-21	0.811 (0.729 - 0.893)	0.784 (0.698 - 0.870)	0.732 (0.639 - 0.825)	0.020 (0.000 - 0.049)	0.835 (0.757 - 0.913)	0.136
-20	0.771 (0.679 - 0.863)	0.695 (0.594 - 0.796)	0.745 (0.649 - 0.841)	0.025 (0.000 - 0.059)	0.854 (0.777 - 0.931)	0.136
-19	0.773 (0.687 - 0.859)	0.692 (0.597 - 0.787)	0.788 (0.704 - 0.872)	0.027 (0.000 - 0.060)	0.814 (0.734 - 0.894)	0.161
-18	0.765 (0.675 - 0.855)	0.671 (0.571 - 0.771)	0.776 (0.687 - 0.865)	0.027 (0.000 - 0.061)	0.837 (0.758 - 0.916)	0.157
-17	0.791 (0.712 - 0.870)	0.802 (0.724 - 0.880)	0.666 (0.574 - 0.758)	0.023 (0.000 - 0.052)	0.842 (0.771 - 0.913)	0.076
-16	0.780 (0.699 - 0.861)	0.728 (0.641 - 0.815)	0.752 (0.667 - 0.837)	0.028 (0.000 - 0.060)	0.810 (0.733 - 0.887)	0.134
-15	0.764 (0.679 - 0.849)	0.619 (0.522 - 0.716)	0.817 (0.740 - 0.894)	0.034 (0.000 - 0.070)	0.799 (0.719 - 0.879)	0.150
-14	0.760 (0.676 - 0.844)	0.586 (0.489 - 0.683)	0.859 (0.790 - 0.928)	0.035 (0.000 - 0.071)	0.759 (0.675 - 0.843)	0.238

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-13	0.771 (0.689 - 0.853)	0.686 (0.596 - 0.776)	0.748 (0.664 - 0.832)	0.032 (0.000 - 0.066)	0.824 (0.750 - 0.898)	0.161
-12	0.819 (0.744 - 0.894)	0.728 (0.641 - 0.815)	0.809 (0.732 - 0.886)	0.026 (0.000 - 0.057)	0.769 (0.686 - 0.852)	0.163
-11	0.774 (0.697 - 0.851)	0.750 (0.670 - 0.830)	0.707 (0.623 - 0.791)	0.030 (0.000 - 0.062)	0.819 (0.748 - 0.890)	0.136
-10	0.793 (0.717 - 0.869)	0.745 (0.664 - 0.826)	0.755 (0.675 - 0.835)	0.028 (0.000 - 0.059)	0.796 (0.721 - 0.871)	0.157
-9	0.799 (0.727 - 0.871)	0.754 (0.676 - 0.832)	0.753 (0.675 - 0.831)	0.029 (0.000 - 0.059)	0.782 (0.708 - 0.856)	0.154
-8	0.795 (0.719 - 0.871)	0.667 (0.578 - 0.756)	0.821 (0.748 - 0.894)	0.032 (0.000 - 0.065)	0.765 (0.685 - 0.845)	0.176
-7	0.779 (0.702 - 0.856)	0.757 (0.677 - 0.837)	0.696 (0.610 - 0.782)	0.029 (0.000 - 0.060)	0.825 (0.754 - 0.896)	0.104
-6	0.779 (0.702 - 0.856)	0.658 (0.570 - 0.746)	0.804 (0.730 - 0.878)	0.035 (0.001 - 0.069)	0.778 (0.701 - 0.855)	0.171
-5	0.797 (0.725 - 0.869)	0.628 (0.542 - 0.714)	0.839 (0.774 - 0.904)	0.040 (0.005 - 0.075)	0.732 (0.653 - 0.811)	0.188
-4	0.812 (0.753 - 0.871)	0.760 (0.695 - 0.825)	0.754 (0.689 - 0.819)	0.029 (0.004 - 0.054)	0.776 (0.713 - 0.839)	0.152
-3	0.817 (0.758 - 0.876)	0.835 (0.779 - 0.891)	0.690 (0.620 - 0.760)	0.022 (0.000 - 0.044)	0.799 (0.738 - 0.860)	0.092
-2	0.805 (0.734 - 0.876)	0.752 (0.674 - 0.830)	0.725 (0.644 - 0.806)	0.031 (0.000 - 0.062)	0.796 (0.723 - 0.869)	0.116
-1	0.848 (0.783 - 0.913)	0.783 (0.708 - 0.858)	0.799 (0.726 - 0.872)	0.025 (0.000 - 0.053)	0.734 (0.654 - 0.814)	0.156
Mean ± SD	**0.791 ± 0.021**	**0.717 ± 0.066**	**0.765 ± 0.052**	**0.028 ± 0.005**	**0.801 ± 0.034**	
Model	**GBM**					
-24	0.844 (0.761 - 0.927)	0.805 (0.714 - 0.896)	0.774 (0.678 - 0.870)	0.015 (0.000 - 0.043)	0.826 (0.739 - 0.913)	0.072
-23	0.860 (0.787 - 0.933)	0.782 (0.695 - 0.869)	0.807 (0.724 - 0.890)	0.018 (0.000 - 0.046)	0.787 (0.701 - 0.873)	0.107
-22	0.862 (0.789 - 0.935)	0.849 (0.773 - 0.925)	0.743 (0.651 - 0.835)	0.013 (0.000 - 0.037)	0.821 (0.740 - 0.902)	0.087
-21	0.843 (0.767 - 0.919)	0.784 (0.698 - 0.870)	0.778 (0.691 - 0.865)	0.018 (0.000 - 0.046)	0.807 (0.725 - 0.889)	0.093
-20	0.817 (0.732 - 0.902)	0.854 (0.777 - 0.931)	0.648 (0.543 - 0.753)	0.014 (0.000 - 0.040)	0.868 (0.794 - 0.942)	0.079
-19	0.791 (0.707 - 0.875)	0.626 (0.527 - 0.725)	0.829 (0.752 - 0.906)	0.031 (0.000 - 0.067)	0.796 (0.713 - 0.879)	0.133
-18	0.826 (0.745 - 0.907)	0.788 (0.701 - 0.875)	0.743 (0.650 - 0.836)	0.018 (0.000 - 0.046)	0.834 (0.755 - 0.913)	0.069
-17	0.839 (0.767 - 0.911)	0.822 (0.747 - 0.897)	0.714 (0.626 - 0.802)	0.019 (0.000 - 0.046)	0.817 (0.742 - 0.892)	0.066
-16	0.831 (0.758 - 0.904)	0.796 (0.717 - 0.875)	0.726 (0.639 - 0.813)	0.022 (0.000 - 0.051)	0.812 (0.735 - 0.889)	0.092
-15	0.800 (0.720 - 0.880)	0.866 (0.798 - 0.934)	0.587 (0.489 - 0.685)	0.017 (0.000 - 0.043)	0.865 (0.797 - 0.933)	0.109
-14	0.848 (0.777 - 0.919)	0.727 (0.639 - 0.815)	0.830 (0.756 - 0.904)	0.024 (0.000 - 0.054)	0.754 (0.669 - 0.839)	0.117
-13	0.851 (0.782 - 0.920)	0.676 (0.585 - 0.767)	0.871 (0.806 - 0.936)	0.028 (0.000 - 0.060)	0.709 (0.621 - 0.797)	0.154
-12	0.857 (0.788 - 0.926)	0.825 (0.751 - 0.899)	0.733 (0.646 - 0.820)	0.018 (0.000 - 0.044)	0.803 (0.725 - 0.881)	0.161
-11	0.850 (0.784 - 0.916)	0.786 (0.710 - 0.862)	0.769 (0.691 - 0.847)	0.024 (0.000 - 0.052)	0.773 (0.695 - 0.851)	0.129
-10	0.846 (0.779 - 0.913)	0.755 (0.675 - 0.835)	0.799 (0.724 - 0.874)	0.025 (0.000 - 0.054)	0.759 (0.679 - 0.839)	0.104
-9	0.827 (0.759 - 0.895)	0.788 (0.714 - 0.862)	0.764 (0.687 - 0.841)	0.025 (0.000 - 0.053)	0.767 (0.691 - 0.843)	0.110
-8	0.830 (0.759 - 0.901)	0.769 (0.689 - 0.849)	0.747 (0.665 - 0.829)	0.025 (0.000 - 0.055)	0.800 (0.724 - 0.876)	0.107
-7	0.842 (0.774 - 0.910)	0.820 (0.749 - 0.891)	0.712 (0.628 - 0.796)	0.021 (0.000 - 0.048)	0.805 (0.731 - 0.879)	0.071

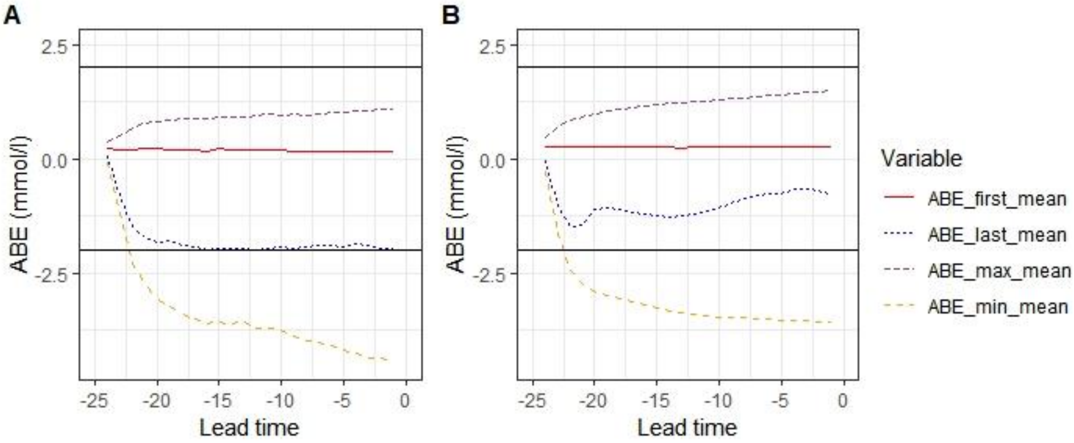
Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-6	0.840 (0.772 - 0.908)	0.694 (0.608 - 0.780)	0.837 (0.768 - 0.906)	0.030 (0.000 - 0.062)	0.734 (0.652 - 0.816)	0.144
-5	0.835 (0.769 - 0.901)	0.678 (0.595 - 0.761)	0.841 (0.776 - 0.906)	0.035 (0.002 - 0.068)	0.715 (0.635 - 0.795)	0.163
-4	0.840 (0.784 - 0.896)	0.661 (0.589 - 0.733)	0.852 (0.798 - 0.906)	0.036 (0.008 - 0.064)	0.705 (0.636 - 0.774)	0.153
-3	0.839 (0.783 - 0.895)	0.719 (0.651 - 0.787)	0.808 (0.748 - 0.868)	0.032 (0.005 - 0.059)	0.741 (0.675 - 0.807)	0.134
-2	0.846 (0.781 - 0.911)	0.661 (0.576 - 0.746)	0.880 (0.821 - 0.939)	0.035 (0.002 - 0.068)	0.661 (0.576 - 0.746)	0.188
-1	0.886 (0.828 - 0.944)	0.858 (0.794 - 0.922)	0.773 (0.697 - 0.849)	0.017 (0.000 - 0.041)	0.741 (0.661 - 0.821)	0.093
Mean ± SD	**0.840 ± 0.020**	**0.766 ± 0.070**	**0.774 ± 0.069**	**0.023 ± 0.007**	**0.779 ± 0.052**	
Model	**LR**					
-24	0.616 (0.504 - 0.728)	0.571 (0.457 - 0.685)	0.724 (0.621 - 0.827)	0.034 (0.000 - 0.076)	0.891 (0.820 - 0.962)	0.058
-23	0.664 (0.565 - 0.763)	0.701 (0.605 - 0.797)	0.678 (0.580 - 0.776)	0.029 (0.000 - 0.064)	0.873 (0.803 - 0.943)	0.044
-22	0.638 (0.536 - 0.740)	0.605 (0.502 - 0.708)	0.754 (0.663 - 0.845)	0.033 (0.000 - 0.071)	0.860 (0.787 - 0.933)	0.064
-21	0.623 (0.522 - 0.724)	0.284 (0.190 - 0.378)	0.962 (0.922 - 1.000)	0.048 (0.003 - 0.093)	0.667 (0.569 - 0.765)	0.001
-20	0.580 (0.472 - 0.688)	0.195 (0.108 - 0.282)	0.965 (0.925 - 1.000)	0.050 (0.002 - 0.098)	0.742 (0.646 - 0.838)	0.001
-19	0.610 (0.510 - 0.710)	0.264 (0.173 - 0.355)	0.956 (0.914 - 0.998)	0.051 (0.006 - 0.096)	0.704 (0.610 - 0.798)	0.001
-18	0.723 (0.628 - 0.818)	0.753 (0.661 - 0.845)	0.694 (0.596 - 0.792)	0.023 (0.000 - 0.055)	0.862 (0.789 - 0.935)	0.070
-17	0.620 (0.525 - 0.715)	0.683 (0.592 - 0.774)	0.656 (0.563 - 0.749)	0.036 (0.000 - 0.072)	0.866 (0.800 - 0.932)	0.034
-16	0.784 (0.703 - 0.865)	0.748 (0.663 - 0.833)	0.724 (0.636 - 0.812)	0.027 (0.000 - 0.059)	0.822 (0.747 - 0.897)	0.063
-15	0.707 (0.616 - 0.798)	0.649 (0.554 - 0.744)	0.695 (0.603 - 0.787)	0.036 (0.000 - 0.073)	0.863 (0.795 - 0.931)	0.060
-14	0.796 (0.717 - 0.875)	0.798 (0.719 - 0.877)	0.730 (0.643 - 0.817)	0.021 (0.000 - 0.049)	0.816 (0.740 - 0.892)	0.054
-13	0.555 (0.459 - 0.651)	0.667 (0.576 - 0.758)	0.590 (0.495 - 0.685)	0.042 (0.003 - 0.081)	0.887 (0.826 - 0.948)	0.059
-12	0.803 (0.725 - 0.881)	0.728 (0.641 - 0.815)	0.783 (0.702 - 0.864)	0.027 (0.000 - 0.059)	0.791 (0.711 - 0.871)	0.094
-11	0.807 (0.734 - 0.880)	0.768 (0.690 - 0.846)	0.743 (0.662 - 0.824)	0.026 (0.000 - 0.055)	0.795 (0.720 - 0.870)	0.055
-10	0.826 (0.755 - 0.897)	0.882 (0.822 - 0.942)	0.661 (0.573 - 0.749)	0.015 (0.000 - 0.038)	0.820 (0.748 - 0.892)	0.040
-9	0.845 (0.780 - 0.910)	0.822 (0.753 - 0.891)	0.731 (0.651 - 0.811)	0.022 (0.000 - 0.048)	0.782 (0.708 - 0.856)	0.055
-8	0.822 (0.750 - 0.894)	0.815 (0.741 - 0.889)	0.703 (0.616 - 0.790)	0.021 (0.000 - 0.048)	0.815 (0.741 - 0.889)	0.045
-7	0.801 (0.727 - 0.875)	0.802 (0.728 - 0.876)	0.713 (0.629 - 0.797)	0.023 (0.000 - 0.051)	0.808 (0.735 - 0.881)	0.045
-6	0.775 (0.697 - 0.853)	0.811 (0.738 - 0.884)	0.677 (0.590 - 0.764)	0.023 (0.000 - 0.051)	0.824 (0.753 - 0.895)	0.037
-5	0.826 (0.758 - 0.894)	0.793 (0.721 - 0.865)	0.743 (0.665 - 0.821)	0.025 (0.000 - 0.053)	0.776 (0.702 - 0.850)	0.067
-4	0.819 (0.761 - 0.877)	0.810 (0.751 - 0.869)	0.732 (0.665 - 0.799)	0.024 (0.001 - 0.047)	0.780 (0.717 - 0.843)	0.063
-3	0.825 (0.767 - 0.883)	0.810 (0.751 - 0.869)	0.733 (0.666 - 0.800)	0.024 (0.001 - 0.047)	0.779 (0.716 - 0.842)	0.061
-2	0.837 (0.770 - 0.904)	0.661 (0.576 - 0.746)	0.866 (0.805 - 0.927)	0.035 (0.002 - 0.068)	0.685 (0.601 - 0.769)	0.124
-1	0.897 (0.842 - 0.952)	0.783 (0.708 - 0.858)	0.886 (0.828 - 0.944)	0.022 (0.000 - 0.049)	0.610 (0.521 - 0.699)	0.100
Mean ± SD	**0.742 ± 0.101**	**0.683 ± 0.185**	**0.754 ± 0.100**	**0.030 ± 0.010**	**0.797 ± 0.072**	

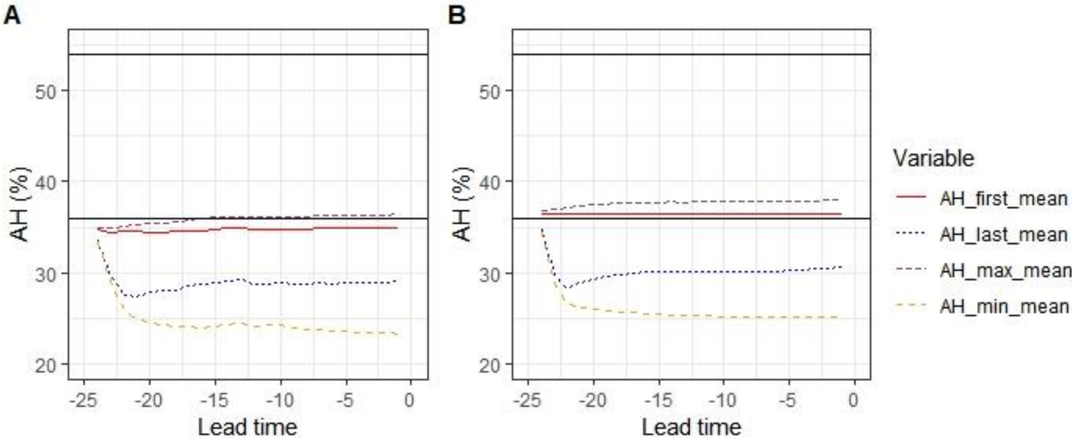
Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**RF**					
-24	0.815 (0.726 - 0.904)	0.636 (0.526 - 0.746)	0.843 (0.760 - 0.926)	0.025 (0.000 - 0.061)	0.807 (0.716 - 0.898)	0.140
-23	0.837 (0.759 - 0.915)	0.793 (0.708 - 0.878)	0.751 (0.660 - 0.842)	0.018 (0.000 - 0.046)	0.824 (0.744 - 0.904)	0.095
-22	0.832 (0.753 - 0.911)	0.791 (0.705 - 0.877)	0.728 (0.634 - 0.822)	0.019 (0.000 - 0.048)	0.839 (0.761 - 0.917)	0.095
-21	0.829 (0.750 - 0.908)	0.659 (0.560 - 0.758)	0.825 (0.746 - 0.904)	0.027 (0.000 - 0.061)	0.796 (0.712 - 0.880)	0.105
-20	0.802 (0.715 - 0.889)	0.805 (0.718 - 0.892)	0.676 (0.573 - 0.779)	0.018 (0.000 - 0.047)	0.865 (0.790 - 0.940)	0.070
-19	0.813 (0.733 - 0.893)	0.758 (0.670 - 0.846)	0.715 (0.622 - 0.808)	0.023 (0.000 - 0.054)	0.843 (0.768 - 0.918)	0.075
-18	0.826 (0.745 - 0.907)	0.741 (0.648 - 0.834)	0.770 (0.681 - 0.859)	0.021 (0.000 - 0.051)	0.827 (0.747 - 0.907)	0.090
-17	0.816 (0.740 - 0.892)	0.812 (0.736 - 0.888)	0.669 (0.577 - 0.761)	0.022 (0.000 - 0.051)	0.839 (0.767 - 0.911)	0.065
-16	0.817 (0.741 - 0.893)	0.621 (0.526 - 0.716)	0.876 (0.811 - 0.941)	0.033 (0.000 - 0.068)	0.714 (0.625 - 0.803)	0.125
-15	0.814 (0.737 - 0.891)	0.639 (0.543 - 0.735)	0.822 (0.746 - 0.898)	0.032 (0.000 - 0.067)	0.788 (0.707 - 0.869)	0.110
-14	0.817 (0.741 - 0.893)	0.747 (0.661 - 0.833)	0.769 (0.686 - 0.852)	0.024 (0.000 - 0.054)	0.803 (0.725 - 0.881)	0.110
-13	0.846 (0.776 - 0.916)	0.814 (0.738 - 0.890)	0.720 (0.633 - 0.807)	0.020 (0.000 - 0.047)	0.814 (0.738 - 0.890)	0.080
-12	0.838 (0.766 - 0.910)	0.825 (0.751 - 0.899)	0.740 (0.654 - 0.826)	0.018 (0.000 - 0.044)	0.799 (0.720 - 0.878)	0.110
-11	0.825 (0.755 - 0.895)	0.750 (0.670 - 0.830)	0.792 (0.717 - 0.867)	0.027 (0.000 - 0.057)	0.763 (0.684 - 0.842)	0.130
-10	0.809 (0.736 - 0.882)	0.664 (0.576 - 0.752)	0.819 (0.747 - 0.891)	0.034 (0.000 - 0.068)	0.764 (0.685 - 0.843)	0.145
-9	0.809 (0.738 - 0.880)	0.720 (0.639 - 0.801)	0.781 (0.706 - 0.856)	0.032 (0.000 - 0.064)	0.769 (0.693 - 0.845)	0.130
-8	0.791 (0.714 - 0.868)	0.741 (0.658 - 0.824)	0.726 (0.641 - 0.811)	0.029 (0.000 - 0.061)	0.817 (0.744 - 0.890)	0.100
-7	0.814 (0.742 - 0.886)	0.757 (0.677 - 0.837)	0.766 (0.687 - 0.845)	0.026 (0.000 - 0.056)	0.784 (0.707 - 0.861)	0.125
-6	0.809 (0.736 - 0.882)	0.730 (0.647 - 0.813)	0.761 (0.682 - 0.840)	0.029 (0.000 - 0.060)	0.794 (0.719 - 0.869)	0.125
-5	0.800 (0.729 - 0.871)	0.760 (0.684 - 0.836)	0.720 (0.640 - 0.800)	0.030 (0.000 - 0.060)	0.797 (0.725 - 0.869)	0.095
-4	0.805 (0.745 - 0.865)	0.736 (0.669 - 0.803)	0.738 (0.671 - 0.805)	0.032 (0.005 - 0.059)	0.792 (0.730 - 0.854)	0.110
-3	0.807 (0.747 - 0.867)	0.719 (0.651 - 0.787)	0.753 (0.688 - 0.818)	0.034 (0.007 - 0.061)	0.786 (0.724 - 0.848)	0.120
-2	0.802 (0.730 - 0.874)	0.702 (0.619 - 0.785)	0.813 (0.743 - 0.883)	0.033 (0.001 - 0.065)	0.740 (0.661 - 0.819)	0.160
-1	0.853 (0.789 - 0.917)	0.792 (0.718 - 0.866)	0.780 (0.705 - 0.855)	0.024 (0.000 - 0.052)	0.750 (0.671 - 0.829)	0.095
Mean ± SD	**0.818 ± 0.015**	**0.738 ± 0.059**	**0.765 ± 0.051**	**0.026 ± 0.006**	**0.796 ± 0.035**	
Model	**SVM**					
-24	0.763 (0.665 - 0.861)	0.714 (0.610 - 0.818)	0.754 (0.655 - 0.853)	0.020 (0.000 - 0.052)	0.854 (0.773 - 0.935)	0.066
-23	0.820 (0.739 - 0.901)	0.667 (0.568 - 0.766)	0.872 (0.802 - 0.942)	0.025 (0.000 - 0.058)	0.741 (0.649 - 0.833)	0.095
-22	0.800 (0.715 - 0.885)	0.709 (0.613 - 0.805)	0.795 (0.710 - 0.880)	0.024 (0.000 - 0.056)	0.814 (0.732 - 0.896)	0.076
-21	0.789 (0.704 - 0.874)	0.693 (0.597 - 0.789)	0.765 (0.676 - 0.854)	0.026 (0.000 - 0.059)	0.833 (0.755 - 0.911)	0.075
-20	0.778 (0.687 - 0.869)	0.756 (0.662 - 0.850)	0.693 (0.592 - 0.794)	0.022 (0.000 - 0.054)	0.866 (0.791 - 0.941)	0.063
-19	0.773 (0.687 - 0.859)	0.769 (0.682 - 0.856)	0.716 (0.623 - 0.809)	0.022 (0.000 - 0.052)	0.841 (0.766 - 0.916)	0.065

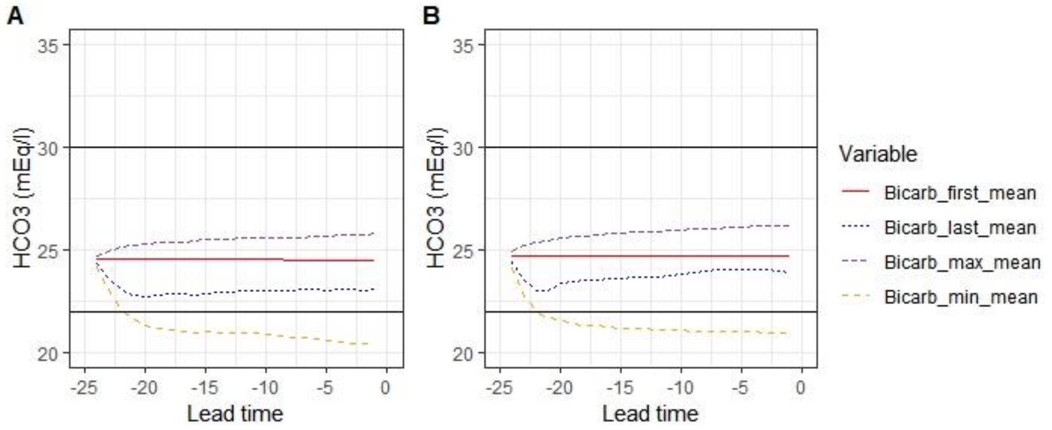
Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-18	0.770 (0.681 - 0.859)	0.741 (0.648 - 0.834)	0.683 (0.584 - 0.782)	0.024 (0.000 - 0.057)	0.868 (0.796 - 0.940)	0.067
-17	0.795 (0.716 - 0.874)	0.782 (0.701 - 0.863)	0.695 (0.605 - 0.785)	0.024 (0.000 - 0.054)	0.833 (0.760 - 0.906)	0.063
-16	0.785 (0.704 - 0.866)	0.699 (0.609 - 0.789)	0.766 (0.683 - 0.849)	0.030 (0.000 - 0.063)	0.807 (0.730 - 0.884)	0.073
-15	0.749 (0.663 - 0.835)	0.598 (0.500 - 0.696)	0.811 (0.733 - 0.889)	0.036 (0.000 - 0.073)	0.809 (0.731 - 0.887)	0.077
-14	0.806 (0.728 - 0.884)	0.758 (0.674 - 0.842)	0.727 (0.639 - 0.815)	0.025 (0.000 - 0.056)	0.826 (0.751 - 0.901)	0.065
-13	0.806 (0.729 - 0.883)	0.725 (0.638 - 0.812)	0.766 (0.684 - 0.848)	0.027 (0.000 - 0.058)	0.804 (0.727 - 0.881)	0.075
-12	0.783 (0.702 - 0.864)	0.660 (0.567 - 0.753)	0.826 (0.752 - 0.900)	0.032 (0.000 - 0.066)	0.769 (0.686 - 0.852)	0.091
-11	0.810 (0.737 - 0.883)	0.777 (0.700 - 0.854)	0.700 (0.615 - 0.785)	0.027 (0.000 - 0.057)	0.817 (0.745 - 0.889)	0.063
-10	0.800 (0.725 - 0.875)	0.709 (0.624 - 0.794)	0.783 (0.706 - 0.860)	0.030 (0.000 - 0.062)	0.783 (0.706 - 0.860)	0.081
-9	0.795 (0.722 - 0.868)	0.686 (0.602 - 0.770)	0.766 (0.690 - 0.842)	0.036 (0.002 - 0.070)	0.789 (0.715 - 0.863)	0.079
-8	0.771 (0.691 - 0.851)	0.630 (0.539 - 0.721)	0.754 (0.672 - 0.836)	0.039 (0.002 - 0.076)	0.825 (0.753 - 0.897)	0.070
-7	0.805 (0.731 - 0.879)	0.928 (0.880 - 0.976)	0.511 (0.418 - 0.604)	0.012 (0.000 - 0.032)	0.861 (0.797 - 0.925)	0.045
-6	0.799 (0.724 - 0.874)	0.568 (0.476 - 0.660)	0.873 (0.811 - 0.935)	0.040 (0.004 - 0.076)	0.724 (0.641 - 0.807)	0.113
-5	0.786 (0.713 - 0.859)	0.603 (0.516 - 0.690)	0.841 (0.776 - 0.906)	0.042 (0.006 - 0.078)	0.738 (0.660 - 0.816)	0.098
-4	0.794 (0.733 - 0.855)	0.653 (0.581 - 0.725)	0.824 (0.766 - 0.882)	0.038 (0.009 - 0.067)	0.743 (0.677 - 0.809)	0.090
-3	0.804 (0.744 - 0.864)	0.727 (0.659 - 0.795)	0.773 (0.709 - 0.837)	0.032 (0.005 - 0.059)	0.770 (0.706 - 0.834)	0.075
-2	0.820 (0.751 - 0.889)	0.694 (0.611 - 0.777)	0.826 (0.758 - 0.894)	0.033 (0.001 - 0.065)	0.729 (0.649 - 0.809)	0.083
-1	0.887 (0.829 - 0.945)	0.775 (0.699 - 0.851)	0.851 (0.786 - 0.916)	0.024 (0.000 - 0.052)	0.675 (0.590 - 0.760)	0.087
Mean ± SD	**0.795 ± 0.026**	**0.709 ± 0.075**	**0.765 ± 0.078**	**0.029 ± 0.007**	**0.797 ± 0.051**	

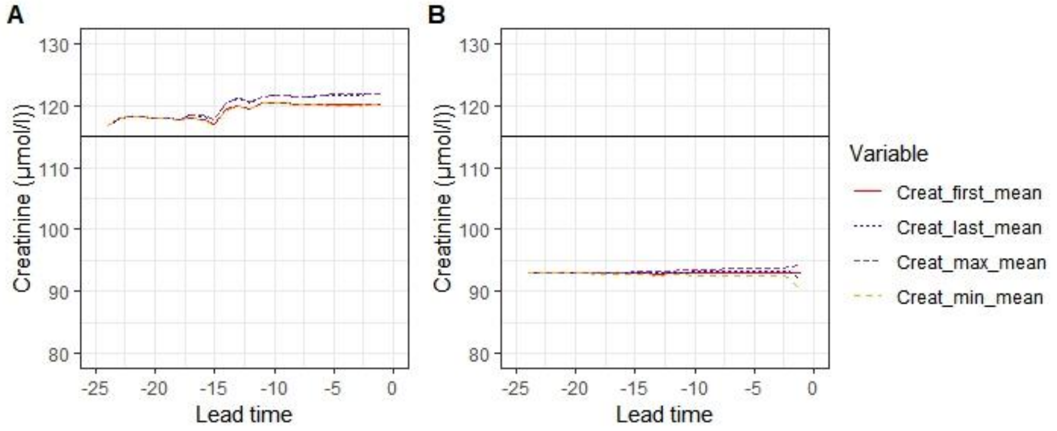
Appendix 7.4: How Laboratory Values Change as the Lead Times Change for AKI vs non-AKI Patients

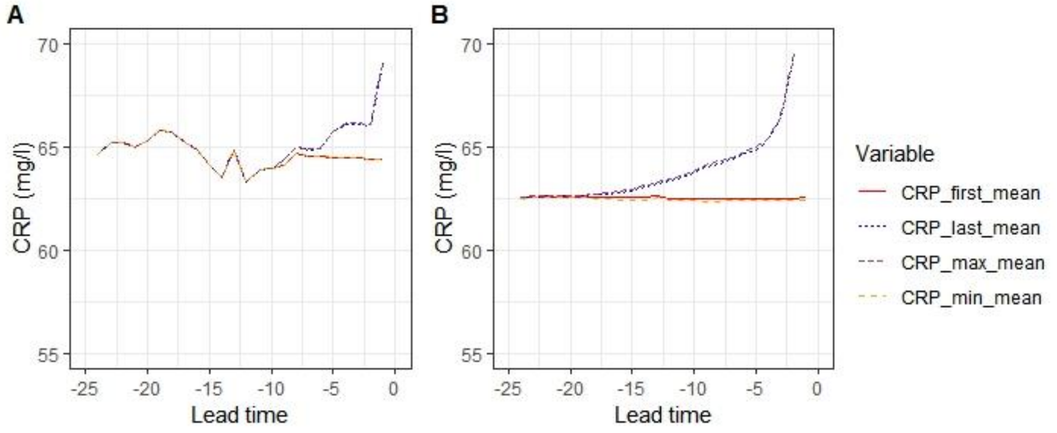
The changing of laboratory variable summary values as the lead time changes for patients with acute kidney injury (AKI) (A) and for patients without AKI (B). The horizontal lines indicate the normal range for each respective laboratory value. Abbreviations: ABE, arterial base excess; AH, arterial haematocrit; CRP, C-Reactive Protein; HCO3, bicarbonate; H+, Hydrogen ion.

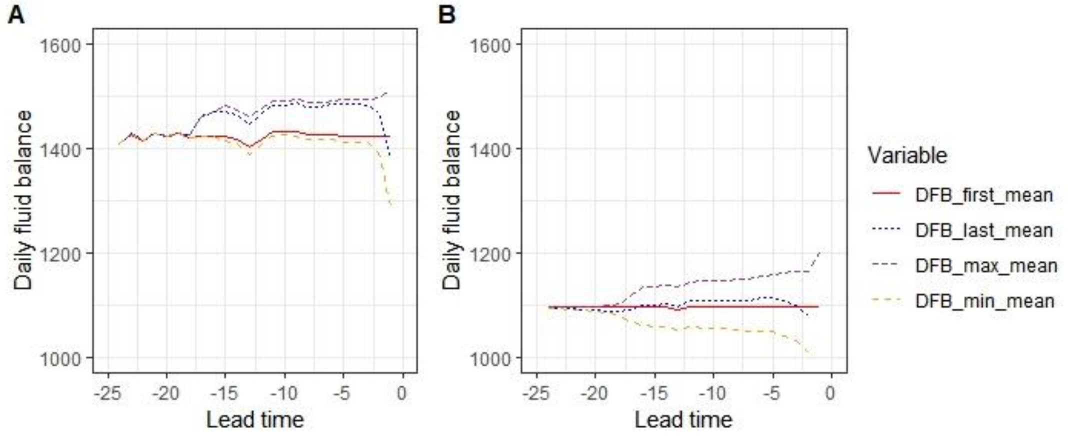


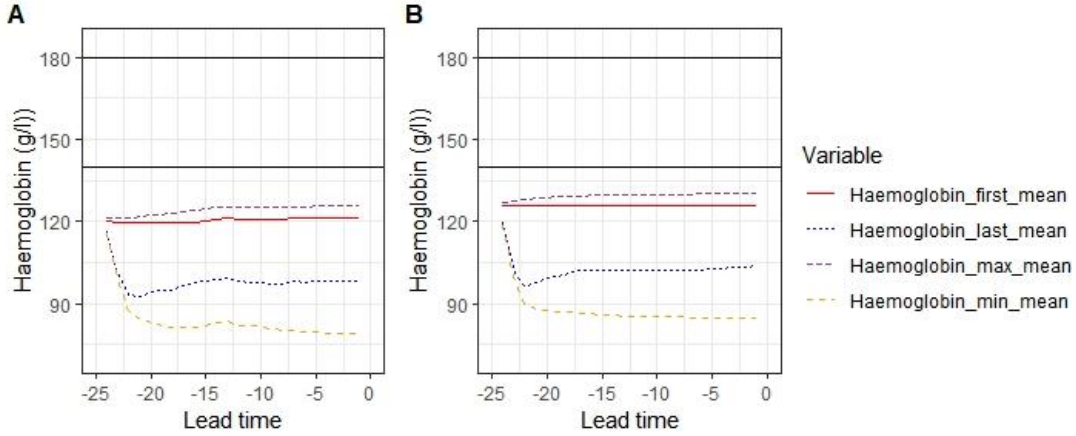


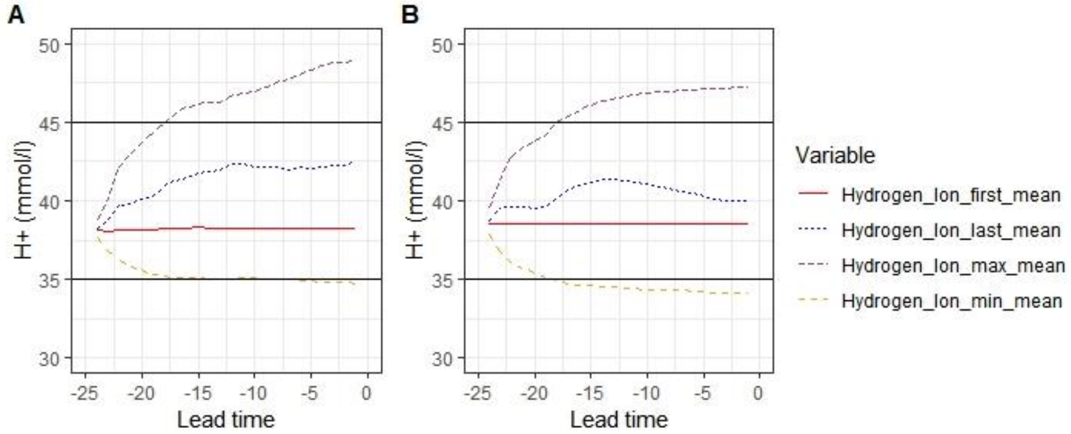

A
B
HCO3 (mEq/l)
35
30
25
20
-25
-20
-15
-10
-5
0
Lead time
Variable
Bicarb_first_mean
Bicarb_last_mean
Bicarb_max_mean
Bicarb_min_mean

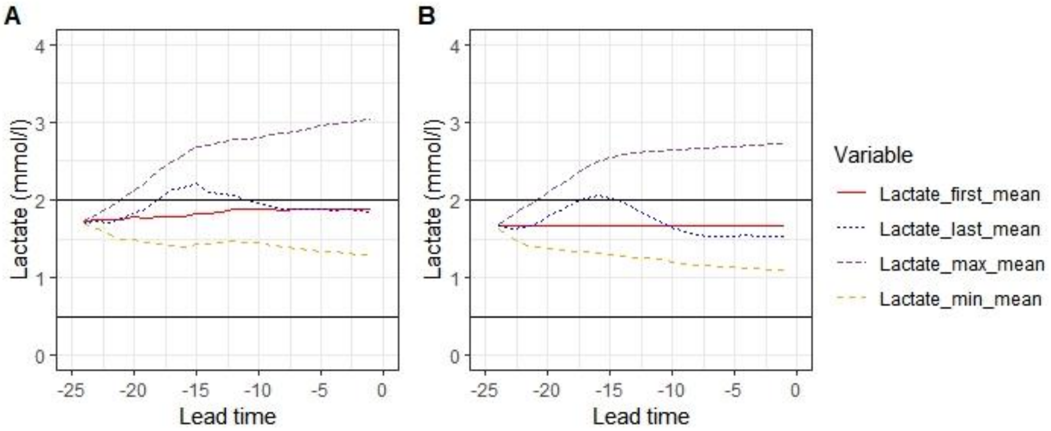

A
B
Creatinine (µmol/l))
130
120
110
100
90
80
-25
-20
-15
-10
-5
0
Lead time
Variable
Creat_first_mean
Creat_last_mean
Creat_max_mean
Creat_min_mean

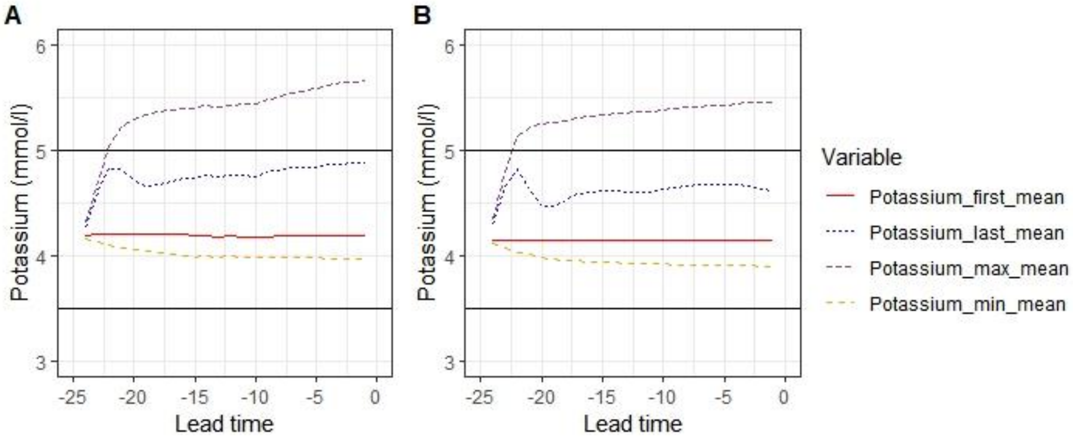

A
B
CRP (mg/l)
70
65
60
55
-25
-20
-15
-10
-5
0
Lead time
Variable
CRP_first_mean
CRP_last_mean
CRP_max_mean
CRP_min_mean

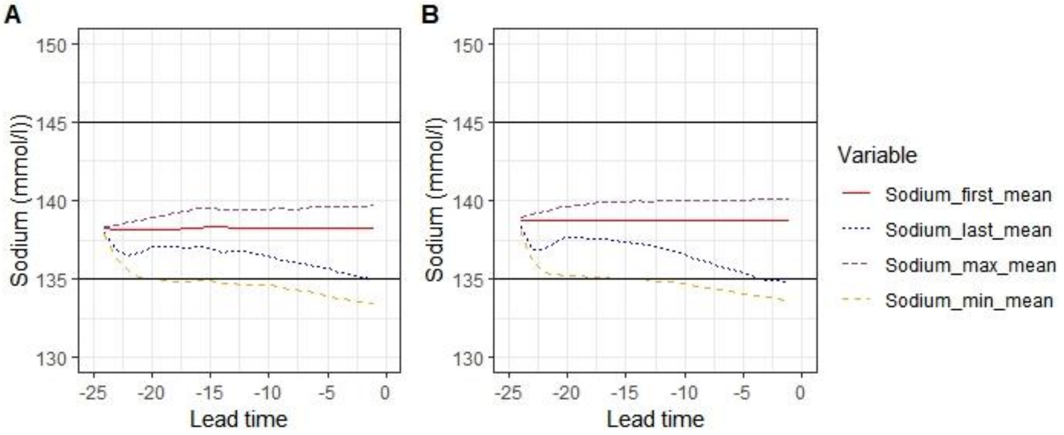

A
B
Daily fluid balance
1600
1400
1200
1000
-25
-20
-15
-10
-5
0
Lead time
Variable
DFB_first_mean
DFB_last_mean
DFB_max_mean
DFB_min_mean

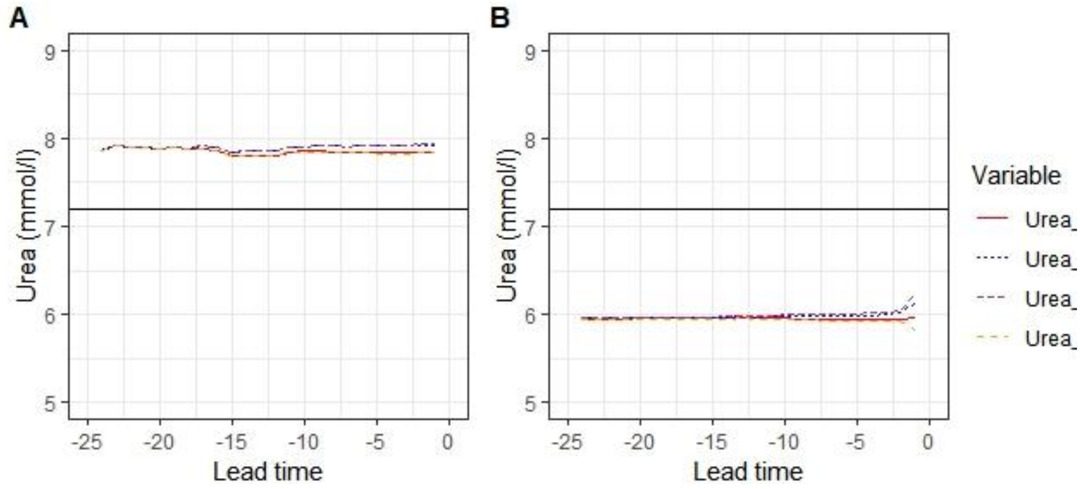

A
B
Haemoglobin (g/l))
180
150
120
90
-25
-20
-15
-10
-5
0
Lead time
Variable
Haemoglobin_first_mean
Haemoglobin_last_mean
Haemoglobin_max_mean
Haemoglobin_min_mean

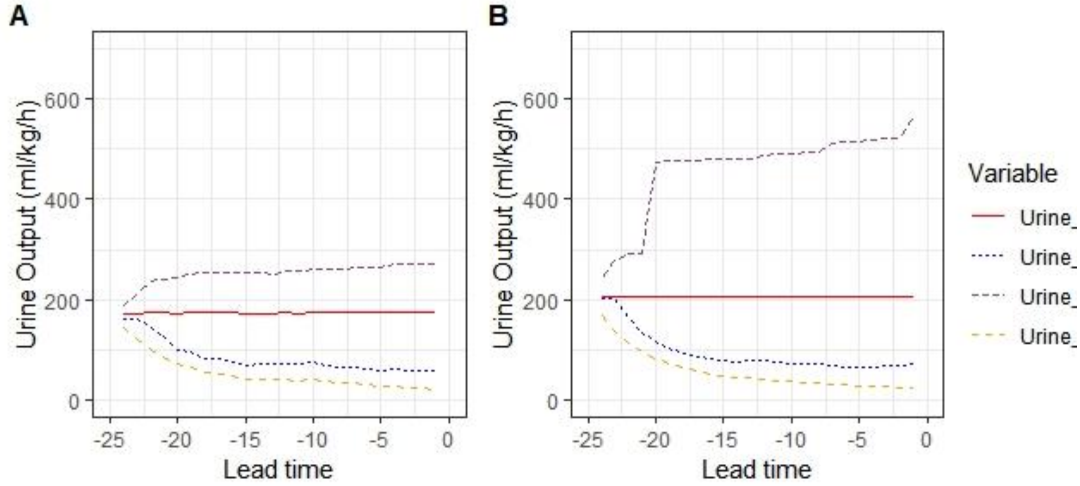

A
B
H+ (mmol/l)
50
45
40
35
30
-25
-20
-15
-10
-5
0
Lead time
Variable
Hydrogen_Ion_first_mean
Hydrogen_Ion_last_mean
Hydrogen_Ion_max_mean
Hydrogen_Ion_min_mean


A
B
Lactate (mmol/l)
0
1
2
3
4
-25
-20
-15
-10
-5
0
Lead time
Variable
Lactate_first_mean
Lactate_last_mean
Lactate_max_mean
Lactate_min_mean


A
B
Potassium (mmol/l)
3
4
5
6
-25
-20
-15
-10
-5
0
Lead time
Variable
Potassium_first_mean
Potassium_last_mean
Potassium_max_mean
Potassium_min_mean


A
B
Sodium (mmol/l))
Sodium (mmol/l)
130
135
140
145
150
-25
-20
-15
-10
-5
0
Lead time
Variable
Sodium_first_mean
Sodium_last_mean
Sodium_max_mean
Sodium_min_mean


A
B
Urea (mmol/l)
9
8
7
6
5
-25
-20
-15
-10
-5
0
Lead time
Variable
Urea_
Urea_
Urea_
Urea_


A
B
Urine Output (ml/kg/h)
600
400
200
0
-25
-20
-15
-10
-5
0
Lead time
Variable
Urine_
Urine_
Urine_
Urine_

Appendix 8.1: Descriptive Statistics when Predicting Delirium using Preoperative and Laboratory ICU Variables

Table 8.1.A. Patient demographics for patients included in the analysis.

Demographic	Total Population N=3322 Mean ± SD or %	Delirium = No N=2905 (87.53%) Mean ± SD or %	Delirium = Yes N=417 (12.47%) Mean ± SD or %	Delirium Yes vs No P-value
Age	65.82 ± 11.19	65.25 ± 11.29	69.77 ± 9.61	<0.0001
Sex				0.0006
Female	28.56%	27.54%	35.73%	
Male	71.44%	72.46%	64.27%	
BMI				0.7988
18.5-25.0	20.96%	20.84%	21.82%	
25.1-30.0	38.52%	38.44%	39.09%	
Over 30.1	40.52%	40.72%	39.09%	
Type II Diabetes				0.1002
No	74.85%	75.33%	71.46%	
Yes	25.15%	24.67%	28.54%	
Smoking Status				0.8382
Never smoked	34.00%	33.99%	34.05%	
Ex-smoker	34.27%	34.40%	33.33%	
Current smoker	15.91%	15.99%	15.35%	
Unknown	15.82%	15.61%	17.27%	
Procedure				<0.0001
CABG	51.97%	54.60%	33.57%	
Valve	33.13%	32.63%	36.69%	
CABG and Valve	14.89%	12.78%	29.74%	
Logistic EuroSCORE	5.58 ± 6.62	5.19 ± 6.09	8.32 ± 9.13	<0.0001
ICU Hours	51.77 ± 109.90	38.01 ± 66.93	148.36 ± 234.21	<0.0001
Total Days in Hospital	11.76 ± 9.47	10.99 ± 8.89	17.16 ± 12.95	<0.0001
Outcome				0.2265
Alive	98.80%	98.91%	98.08%	
Dead	1.20%	1.09%	1.92%	

Table 8.1.B. Patient characteristics for the preoperatively recorded data for all patients and patients without delirium vs with delirium.

Preoperative Variable	Total
	N=3322
	Percentage
		Delirium = No
		N=2905
		(87.53%)
		Percentage
			Delirium = Yes
			N=417
			(12.47%)
			Percentage
				Delirium Yes vs No
				P-value

Surgical Priority				0.0001
Elective	48.53%	48.86%	46.28%	
Emergency	1.56%	1.20%	4.08%	
Priority	26.76%	26.96%	25.42%	
Urgent	23.15%	22.99%	24.22%	
Critical Pre-op. State				<0.0001
No	97.34%	97.81%	94.01%	
Yes	2.66%	2.19%	5.99%	
Previous Cardiac Surgery				0.2485
No	96.53%	96.69%	95.44%	
Yes	3.47%	3.31%	4.56%	
Previous Percutaneous Coronary Intervention				0.2253
No	86.21%	85.92%	88.25%	
Yes	13.79%	14.08%	11.75%	
Extracardiac Arteriopathy				0.7490
No	89.77%	89.85%	89.21%	
Yes	10.23%	10.15%	10.79%	
Left Ventricular Function				0.2112
Good	76.61%	76.73%	75.78%	
Moderate	20.04%	20.12%	19.42%	
Poor	3.35%	3.14%	4.80%	
NYHA Grade				<0.0001
I	21.44%	22.45%	14.39%	
II	47.82%	48.55%	42.69%	
III	26.02%	24.94%	33.57%	
IV	4.72%	4.07%	9.35%	
Angina Status				0.1452
0	38.01%	37.31%	42.93%	
I	13.22%	13.53%	11.03%	
II	29.01%	29.45%	25.90%	
III	14.03%	14.08%	13.67%	
IV	5.74%	5.64%	6.47%	
Rhythm				0.0029
Normal	80.95%	81.76%	75.30%	
Abnormal	14.65%	13.87%	20.14%	
Unknown	4.40%	4.37%	4.56%	
Renal Function Before Surgery				<0.0001
Normal	53.11%	55.35%	37.41%	
Moderately Impaired	37.44%	36.73%	42.45%	
Severely Impaired	9.45%	7.93%	20.14%	
Preoperative Creatinine	90.79 ± 46.34	89.31 ± 42.01	101.17 ± 68.74	0.0007
Neurological Dysfunction				0.8421
No	99.28%	99.56%	99.76%	

Preoperative Variable	Total
	N=3322
	
	Percentage
		Delirium = No
		N=2905
		(87.53%)
		Percentage
			Delirium = Yes
			N=417
			(12.47%)
			Percentage
				Delirium Yes vs No
				
				P-value

Yes	0.42%	0.44%	0.24%	
Previous Myocardial Infarction				0.2398
No	64.21%	63.82%	66.91%	
Yes	35.79%	36.18%	33.09%	
Left Main Stem Disease				0.0348
No	53.74%	52.92%	65.47%	
Yes	14.47%	14.86%	11.75%	
Unknown	31.79%	32.22%	28.78%	
Pulmonary Disease				0.4612
No	85.02%	85.21%	83.69%	
Yes	14.98%	14.79%	16.31%	
Hypertension History				0.6735
No	27.60%	72.26%	26.62%	
Yes	72.40%	27.74%	73.38%	
Congestive Cardiac Failure				<0.0001
No	86.93%	87.94%	79.86%	
Yes	13.07%	12.06%	20.14%	
Active Endocarditis				0.8331
No	98.32%	98.36%	98.08%	
Yes	1.68%	1.64%	1.92%	

Table 8.1.C. Descriptive statistics for laboratory variables for total data and for patients without delirium vs with delirium.

Laboratory Variable (Unit)	Total
	N=3322
	Mean ± SD
		Delirium = No
		N=2905 (87.53%)
		Mean ± SD
			Delirium = Yes
			N=417 (12.47%)
			Mean ± SD
				Delirium Yes vs No
				P-value

Arterial Base Excess (mmol/l)	-0.38 ± 3.31	-0.47 ± 3.04	-0.10 ± 3.96	<0.0001
Arterial Haematocrit (%)	29.16 ± 4.28	29.49 ± 4.44	28.24 ± 3.64	<0.0001
Bicarbonate (mEq/l)	24.23 ± 3.31	24.18 ± 3.01	24.37 ± 4.06	<0.0001
C-Reactive Protein (µmol/L)	135.88 ± 96.35	138.56 ± 93.79	129.59 ± 101.85	<0.0001
Creatinine (µmol/L)	113.02 ± 81.69	109.23 ± 77.12	122.45 ± 91.44	<0.0001
Daily Fluid Balance	268.09 ± 923.75	294.55 ± 932.83	205.01 ± 898.67	<0.0001
Haemoglobin (g/l)	98.98 ± 14.67	100.11 ± 15.19	95.54 ± 12.34	<0.0001
Hydrogen Ion (mmol/l)	39.65 ± 6.44	39.88 ± 6.17	38.96 ± 7.10	<0.0001
Lactate (mmol/l)	1.64 ± 1.04	1.65 ± 1.03	1.63 ± 1.05	0.0268
Potassium (mmol/l)	4.52 ± 0.77	4.56 ± 0.81	4.39 ± 0.65	<0.0001
Sodium (mmol/l)	135.72 ± 4.84	135.24 ± 4.39	137.09 ± 5.73	<0.0001
Urea (mmol/l)	9.61 ± 6.31	8.53 ± 5.20	12.14 ± 7.79	<0.0001
Urine Output (l per day)	100.58 ± 1198.79	105.92 ± 1454.53	89.72 ± 237.84	<0.0001
Medicines				
Dobutamine (dose)	3.36 ± 6.83	3.26 ± 6.03	3.65 ± 8.93	<0.0001
N (%)	1019 (30.47%)	846 (28.90%)	173 (41.49%)	<0.0001
Dopamine (dose)	3.64 ± 6.96	3.72 ± 7.61	3.29 ± 2.48	0.0085
N (%)	156 (4.66%)	128 (4.37%)	28 (6.71%)	<0.0001
Noradrenaline (dose)	4.12 ± 6.92	3.72 ± 6.77	5.05 ± 7.16	<0.0001
N (%)	1460 (43.66%)	1241 (42.40%)	219 (52.52%)	<0.0001
Vasopressin (dose)	5.14 ± 2.57	5.11 ± 2.57	5.17 ± 2.58	0.3464
N (%)	112 (3.35%)	70 (2.39%)	42 (10.07%)	<0.0001

Appendix 8.2: Comparison of Laboratory Variables in Test Data Based on Imputation Methods

Table 8.2.A. Mean and standard deviation (SD) for each laboratory variable in the test data when predicting delirium within 21 hours since ICU admission. The p-values are calculated based on t-test where the comparison is made in variable means based on the original data (without imputation) and the corresponding imputation method.

	Without imputation	**With 0 imputation**		**With median imputation**		**With missForest imputation**	
Variable	**Mean +- SD**	**Mean +- SD**	**P-value**	**Mean +- SD**	**P-value**	**Mean +- SD**	**P-value**
Arterial Base Excess							
Min	-3.78 ± 2.57	-3.78 ± 2.57	1.000	-3.78 ± 2.57	1.000	-3.79 ± 2.58	1.000
Max	1.42 ± 2.54	1.42 ± 2.54	1.000	1.42 ± 2.54	1.000	1.40 ± 2.53	1.000
First	0.18 ± 2.54	0.18 ± 2.54	1.000	0.18 ± 2.54	1.000	0.16 ± 2.54	1.000
Last	-0.78 ± 2.20	-0.78 ± 2.20	1.000	-0.78 ± 2.20	1.000	-0.78 ± 2.50	1.000
Arterial Haematocrit							
Min	25.35 ± 4.34	25.35 ± 4.34	1.000	25.35 ± 4.34	1.000	25.38 ± 4.39	1.000
Max	37.72 ± 5.37	37.72 ± 5.37	1.000	37.72 ± 5.37	1.000	37.73 ± 5.34	1.000
First	36.51 ± 6.72	36.51 ± 6.72	1.000	36.51 ± 6.72	1.000	36.52 ± 6.69	1.000
Last	30.16 ± 3.97	30.16 ± 3.97	1.000	30.16 ± 3.97	1.000	30.20 ± 4.01	1.000
Creatinine							
Min	101.02 ± 44.75	100.98 ± 44.66	0.8641	100.63 ± 45.10	0.9867	100.75 ± 44.34	0.9922
Max	104.08 ± 49.44	104.04 ± 49.34	0.8731	103.68 ± 49.77	0.9867	103.82 ± 49.00	0.9877
First	101.50 ± 45.80	101.46 ± 45.72	0.8666	101.10 ± 46.15	0.9864	101.23 ± 45.38	0.9869
Last	103.60 ± 48.45	103.56 ± 48.36	0.8712	103.19 ± 48.79	0.9858	103.36 ± 48.03	0.9858
C-Reactive Protein							
Min	59.45 ± 33.76	59.39 ± 33.52	0.6274	58.60 ± 34.26	0.9713	59.31 ± 33.43	0.9783
Max	63.76 ± 35.37	63.67 ± 35.13	0.6199	62.85 ± 35.93	0.9587	63.54 ± 35.02	0.9861
First	59.51 ± 33.80	59.45 ± 33.56	0.6275	58.66 ± 34.30	0.9709	59.37 ± 33.48	0.9776
Last	63.70 ± 35.34	63.61 ± 35.09	0.6199	62.79 ± 35.90	0.9590	63.48 ± 35.02	0.9852
Daily Fluid Balance							
Min	1094.21 ± 778.58	1094.21 ± 778.58	1.000	1094.21 ± 778.58	1.000	1081.11 ± 782.04	1.000

	Without imputation	With 0 imputation		With median imputation		With missForest imputation	
Variable	**Mean +- SD**	**Mean +- SD**	**P-value**	**Mean +- SD**	**P-value**	**Mean +- SD**	**P-value**
Max	1221.04 ± 715.74	1221.04 ± 715.74	1.000	1221.04 ± 715.74	1.000	1211.96 ± 720.57	1.000
First	1164.94 ± 753.92	1164.94 ± 753.92	1.000	1164.94 ± 753.92	1.000	1153.47 ± 757.63	1.000
Last	1151.32 ± 745.39	1151.32 ± 745.39	1.000	1151.32 ± 745.39	1.000	1140.58 ± 750.29	1.000
Bicarbonate							
Min	20.95 ± 2.25	20.95 ± 2.24	0.0596	20.68 ± 3.27	0.9860	20.95 ± 2.25	0.9538
Max	26.28 ± 2.61	26.27 ± 2.60	0.0467	25.93 ± 3.96	0.9949	26.26 ± 2.59	0.9768
First	24.73 ± 2.54	24.73 ± 2.52	0.0522	24.40 ± 3.78	0.9923	24.69 ± 2.53	0.9349
Last	24.10 ± 2.24	24.10 ± 2.22	0.0387	23.78 ± 3.53	0.9997	24.10 ± 2.22	0.9878
Haemoglobin							
Min	85.09 ± 14.91	85.09 ± 14.91	1.000	85.09 ± 14.91	1.000	85.22 ± 15.10	1.000
Max	130.80 ± 17.01	130.80 ± 17.01	1.000	130.80 ± 17.01	1.000	130.81 ± 16.94	1.000
First	127.73 ± 20.92	127.73 ± 20.92	1.000	127.73 ± 20.92	1.000	127.76 ± 20.83	1.000
Last	102.03 ± 13.38	102.03 ± 13.38	1.000	102.03 ± 13.38	1.000	102.14 ± 13.56	1.000
Hydrogen Ion							
Min	34.84 ± 3.46	34.84 ± 3.46	1.000	34.84 ± 3.46	1.000	34.87 ± 3.46	1.000
Max	48.43 ± 12.12	48.43 ± 12.12	1.000	48.43 ± 12.12	1.000	48.43 ± 12.01	1.000
First	38.93 ± 4.62	38.93 ± 4.62	1.000	38.93 ± 4.62	1.000	38.95 ± 4.60	1.000
Last	40.95 ± 4.63	40.95 ± 4.63	1.000	40.95 ± 4.63	1.000	40.93 ± 4.65	1.000
Lactate							
Min	1.22 ± 0.55	1.22 ± 0.55	1.000	1.22 ± 0.55	1.000	1.22 ± 0.55	1.000
Max	2.90 ± 1.60	2.90 ± 1.60	1.000	2.90 ± 1.60	1.000	2.90 ± 1.60	1.000
First	1.75 ± 0.92	1.75 ± 0.92	1.000	1.75 ± 0.92	1.000	1.76 ± 0.92	1.000
Last	1.75 ± 0.97	1.75 ± 0.97	1.000	1.75 ± 0.97	1.000	1.74 ± 0.97	1.000
Potassium							
Min	3.93 ± 0.36	3.93 ± 0.36	1.000	3.93 ± 0.36	1.000	3.93 ± 0.36	1.000
Max	5.52 ± 0.68	5.52 ± 0.68	1.000	5.52 ± 0.68	1.000	5.52 ± 0.68	1.000
First	4.16 ± 0.56	4.16 ± 0.56	1.000	4.16 ± 0.56	1.000	4.16 ± 0.55	1.000
Last	4.71 ± 0.48	4.71 ± 0.48	1.000	4.71 ± 0.48	1.000	4.71 ± 0.48	1.000
Sodium							
Min	134.03 ± 2.77	134.03 ± 2.77	1.000	134.03 ± 2.77	1.000	134.03 ± 2.78	1.000
Max	140.35 ± 3.28	140.35 ± 3.28	1.000	140.35 ± 3.28	1.000	140.35 ± 3.28	1.000
First	139.03 ± 2.92	139.03 ± 2.92	1.000	139.03 ± 2.92	1.000	139.01 ± 2.92	1.000

Variable	Without imputation Mean +- SD	With 0 imputation Mean +- SD	P-value	With median imputation Mean +- SD	P-value	With missForest imputation Mean +- SD	P-value
Last	135.45 ± 3.37	135.45 ± 3.37	1.000	135.45 ± 3.37	1.000	135.47 ± 3.38	1.000
Urea							
Min	6.37 ± 2.60	6.37 ± 2.60	0.8529	6.35 ± 2.63	0.9895	6.35 ± 2.58	0.9937
Max	6.54 ± 2.77	6.54 ± 2.77	0.8584	6.51 ± 2.80	0.9864	6.52 ± 2.75	0.9939
First	6.39 ± 2.63	6.39 ± 2.63	0.8540	6.37 ± 2.65	0.9891	6.37 ± 2.61	0.9954
Last	6.52 ± 2.74	6.52 ± 2.74	0.8574	6.49 ± 2.77	0.9868	6.50 ± 2.72	0.9872
Urine							
Min	27.10 ± 23.38	27.10 ± 23.38	1.000	27.10 ± 23.38	1.000	27.92 ± 28.87	1.000
Max	318.05 ± 143.39	318.05 ± 143.39	1.000	318.05 ± 143.39	1.000	317.68 ± 142.86	1.000
First	205.61 ± 144.92	205.61 ± 144.92	1.000	205.61 ± 144.92	1.000	204.20 ± 144.73	1.000
Last	66.64 ± 62.45	66.64 ± 62.45	1.000	66.64 ± 62.45	1.000	67.53 ± 64.41	1.000

Appendix 8.3: Performance Measures for Each Model for Each Lead Time

Table 8.3.A. Performance measures for each model at each lead time before delirium when predicting delirium within 21h since ICU admission, using complete data.

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**AdaBoost**					
-13	0.867 (0.771 - 0.963)	0.856 (0.757 - 0.955)	0.713 (0.585 - 0.841)	0.049 (0.000 - 0.110)	0.569 (0.429 - 0.709)	0.265
-12	0.894 (0.829 - 0.959)	0.812 (0.729 - 0.895)	0.832 (0.753 - 0.911)	0.033 (0.000 - 0.071)	0.575 (0.471 - 0.679)	0.299
-11	0.883 (0.815 - 0.951)	0.744 (0.651 - 0.837)	0.863 (0.790 - 0.936)	0.036 (0.000 - 0.076)	0.590 (0.485 - 0.695)	0.251
-10	0.880 (0.819 - 0.941)	0.831 (0.761 - 0.901)	0.797 (0.721 - 0.873)	0.023 (0.000 - 0.051)	0.683 (0.596 - 0.770)	0.231
-9	0.907 (0.848 - 0.966)	0.852 (0.780 - 0.924)	0.826 (0.749 - 0.903)	0.023 (0.000 - 0.053)	0.607 (0.508 - 0.706)	0.269
-8	0.909 (0.845 - 0.973)	0.871 (0.797 - 0.945)	0.833 (0.750 - 0.916)	0.023 (0.000 - 0.056)	0.553 (0.443 - 0.663)	0.298
-7	0.936 (0.885 - 0.987)	0.828 (0.749 - 0.907)	0.903 (0.841 - 0.965)	0.024 (0.000 - 0.056)	0.471 (0.366 - 0.576)	0.308
-6	0.932 (0.883 - 0.981)	0.821 (0.746 - 0.896)	0.903 (0.845 - 0.961)	0.023 (0.000 - 0.052)	0.504 (0.406 - 0.602)	0.248
-5	0.949 (0.903 - 0.995)	0.862 (0.790 - 0.934)	0.900 (0.837 - 0.963)	0.022 (0.000 - 0.053)	0.445 (0.341 - 0.549)	0.252
-4	0.956 (0.910 - 1.000)	0.908 (0.843 - 0.973)	0.908 (0.843 - 0.973)	0.017 (0.000 - 0.046)	0.369 (0.261 - 0.477)	0.245
-3	0.988 (0.965 - 1.000)	0.953 (0.908 - 0.998)	0.950 (0.904 - 0.996)	0.060 (0.010 - 0.110)	0.289 (0.193 - 0.385)	0.281
-2	0.988 (0.967 - 1.000)	0.919 (0.866 - 0.972)	0.997 (0.986 - 1.000)	0.011 (0.000 - 0.031)	0.025 (0.000 - 0.055)	0.251
-1	0.992 (0.975 - 1.000)	0.958 (0.919 - 0.997)	0.982 (0.956 - 1.000)	0.003 (0.000 - 0.014)	0.207 (0.128 - 0.286)	0.297
Mean ± SD	**0.929 ± 0.043**	**0.863 ± 0.060**	**0.877 ± 0.078**	**0.023 ± 0.015**	**0.453 ± 0.144**	
Model	**BARTm**					
-13	0.997 (0.982 - 1.000)	0.958 (0.901 - 1.000)	0.989 (0.959 - 1.000)	0.003 (0.000 - 0.018)	0.132 (0.036 - 0.228)	0.168
-12	0.984 (0.957 - 1.000)	0.930 (0.876 - 0.984)	0.989 (0.967 - 1.000)	0.009 (0.000 - 0.029)	0.080 (0.023 - 0.137)	0.148
-11	0.987 (0.963 - 1.000)	0.965 (0.926 - 1.000)	0.948 (0.901 - 0.995)	0.005 (0.000 - 0.020)	0.293 (0.196 - 0.390)	0.116
-10	0.967 (0.933 - 1.000)	0.890 (0.831 - 0.949)	0.923 (0.873 - 0.973)	0.020 (0.000 - 0.046)	0.336 (0.247 - 0.425)	0.135
-9	0.937 (0.888 - 0.986)	0.862 (0.792 - 0.932)	0.905 (0.846 - 0.964)	0.022 (0.000 - 0.052)	0.434 (0.334 - 0.534)	0.155
-8	0.934 (0.879 - 0.989)	0.846 (0.766 - 0.926)	0.897 (0.830 - 0.964)	0.020 (0.000 - 0.051)	0.511 (0.400 - 0.622)	0.190
-7	0.926 (0.871 - 0.981)	0.920 (0.863 - 0.977)	0.827 (0.748 - 0.906)	0.013 (0.000 - 0.037)	0.590 (0.487 - 0.693)	0.115
-6	0.937 (0.890 - 0.984)	0.891 (0.830 - 0.952)	0.837 (0.765 - 0.909)	0.020 (0.000 - 0.047)	0.541 (0.444 - 0.638)	0.120
-5	0.904 (0.842 - 0.966)	0.750 (0.660 - 0.840)	0.907 (0.846 - 0.968)	0.035 (0.000 - 0.073)	0.484 (0.380 - 0.588)	0.221
-4	0.905 (0.840 - 0.970)	0.857 (0.779 - 0.935)	0.829 (0.745 - 0.913)	0.019 (0.000 - 0.049)	0.637 (0.530 - 0.744)	0.145
-3	0.906 (0.844 - 0.968)	0.872 (0.801 - 0.943)	0.785 (0.698 - 0.872)	0.020 (0.000 - 0.050)	0.659 (0.559 - 0.759)	0.110

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-2	0.896 (0.836 - 0.956)	0.792 (0.713 - 0.871)	0.870 (0.804 - 0.936)	0.035 (0.000 - 0.071)	0.518 (0.421 - 0.615)	0.176
-1	0.898 (0.839 - 0.957)	0.842 (0.771 - 0.913)	0.785 (0.705 - 0.865)	0.030 (0.000 - 0.063)	0.626 (0.532 - 0.720)	0.117
Mean ± SD	**0.937 ± 0.036**	**0.875 ± 0.062**	**0.884 ± 0.069**	**0.019 ± 0.010**	**0.449 ± 0.188**	
Model	**C5.0**					
-13	0.991 (0.964 - 1.000)	0.979 (0.938 - 1.000)	0.971 (0.924 - 1.000)	0.002 (0.000 - 0.015)	0.288 (0.160 - 0.416)	0.262
-12	0.979 (0.949 - 1.000)	0.907 (0.846 - 0.968)	0.969 (0.932 - 1.000)	0.013 (0.000 - 0.037)	0.204 (0.119 - 0.289)	0.222
-11	0.980 (0.950 - 1.000)	0.929 (0.874 - 0.984)	0.959 (0.917 - 1.000)	0.009 (0.000 - 0.029)	0.255 (0.162 - 0.348)	0.274
-10	0.957 (0.919 - 0.995)	0.872 (0.809 - 0.935)	0.916 (0.864 - 0.968)	0.024 (0.000 - 0.053)	0.358 (0.268 - 0.448)	0.200
-9	0.917 (0.861 - 0.973)	0.809 (0.730 - 0.888)	0.890 (0.827 - 0.953)	0.030 (0.000 - 0.064)	0.486 (0.385 - 0.587)	0.254
-8	0.922 (0.862 - 0.982)	0.885 (0.814 - 0.956)	0.816 (0.730 - 0.902)	0.016 (0.000 - 0.044)	0.641 (0.535 - 0.747)	0.177
-7	0.905 (0.843 - 0.967)	0.782 (0.695 - 0.869)	0.905 (0.843 - 0.967)	0.031 (0.000 - 0.067)	0.481 (0.376 - 0.586)	0.286
-6	0.890 (0.829 - 0.951)	0.832 (0.759 - 0.905)	0.820 (0.745 - 0.895)	0.031 (0.000 - 0.065)	0.582 (0.486 - 0.678)	0.187
-5	0.873 (0.803 - 0.943)	0.773 (0.685 - 0.861)	0.820 (0.740 - 0.900)	0.035 (0.000 - 0.073)	0.638 (0.538 - 0.738)	0.203
-4	0.864 (0.787 - 0.941)	0.831 (0.747 - 0.915)	0.769 (0.675 - 0.863)	0.024 (0.000 - 0.058)	0.710 (0.609 - 0.811)	0.192
-3	0.860 (0.787 - 0.933)	0.616 (0.513 - 0.719)	0.951 (0.905 - 0.997)	0.049 (0.003 - 0.095)	0.384 (0.281 - 0.487)	0.307
-2	0.872 (0.807 - 0.937)	0.733 (0.647 - 0.819)	0.847 (0.777 - 0.917)	0.046 (0.005 - 0.087)	0.577 (0.481 - 0.673)	0.206
-1	0.881 (0.818 - 0.944)	0.733 (0.647 - 0.819)	0.868 (0.802 - 0.934)	0.045 (0.005 - 0.085)	0.540 (0.443 - 0.637)	0.244
Mean ± SD	**0.915 ± 0.047**	**0.822 ± 0.097**	**0.885 ± 0.067**	**0.027 ± 0.015**	**0.473 ± 0.162**	
Model	**GBM**					
-13	0.993 (0.969 - 1.000)	0.979 (0.938 - 1.000)	0.961 (0.906 - 1.000)	0.002 (0.000 - 0.015)	0.356 (0.221 - 0.491)	0.099
-12	0.982 (0.954 - 1.000)	0.919 (0.861 - 0.977)	0.958 (0.916 - 1.000)	0.011 (0.000 - 0.033)	0.255 (0.163 - 0.347)	0.070
-11	0.990 (0.969 - 1.000)	0.976 (0.943 - 1.000)	0.924 (0.868 - 0.980)	0.003 (0.000 - 0.015)	0.376 (0.273 - 0.479)	0.155
-10	0.969 (0.936 - 1.000)	0.917 (0.865 - 0.969)	0.902 (0.846 - 0.958)	0.016 (0.000 - 0.040)	0.383 (0.292 - 0.474)	0.123
-9	0.955 (0.913 - 0.997)	0.894 (0.832 - 0.956)	0.882 (0.817 - 0.947)	0.017 (0.000 - 0.043)	0.478 (0.377 - 0.579)	0.184
-8	0.942 (0.890 - 0.994)	0.846 (0.766 - 0.926)	0.906 (0.841 - 0.971)	0.019 (0.000 - 0.049)	0.488 (0.377 - 0.599)	0.104
-7	0.924 (0.868 - 0.980)	0.793 (0.708 - 0.878)	0.881 (0.813 - 0.949)	0.030 (0.000 - 0.066)	0.534 (0.429 - 0.639)	0.187
-6	0.914 (0.859 - 0.969)	0.842 (0.771 - 0.913)	0.868 (0.802 - 0.934)	0.028 (0.000 - 0.060)	0.503 (0.405 - 0.601)	0.181
-5	0.909 (0.849 - 0.969)	0.841 (0.765 - 0.917)	0.815 (0.734 - 0.896)	0.025 (0.000 - 0.058)	0.624 (0.523 - 0.725)	0.180
-4	0.899 (0.832 - 0.966)	0.870 (0.795 - 0.945)	0.787 (0.696 - 0.878)	0.018 (0.000 - 0.048)	0.684 (0.580 - 0.788)	0.116
-3	0.914 (0.855 - 0.973)	0.721 (0.626 - 0.816)	0.924 (0.868 - 0.980)	0.037 (0.000 - 0.077)	0.451 (0.346 - 0.556)	0.247
-2	0.906 (0.849 - 0.963)	0.901 (0.843 - 0.959)	0.783 (0.703 - 0.863)	0.019 (0.000 - 0.046)	0.611 (0.516 - 0.706)	0.123
-1	0.909 (0.853 - 0.965)	0.881 (0.818 - 0.944)	0.780 (0.699 - 0.861)	0.023 (0.000 - 0.052)	0.620 (0.525 - 0.715)	0.098
Mean ± SD	**0.939 ± 0.035**	**0.875 ± 0.070**	**0.875 ± 0.064**	**0.019 ± 0.010**	**0.489 ± 0.125**	

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**LR**					
-13	0.959 (0.903 - 1.000)	0.938 (0.870 - 1.000)	0.955 (0.896 - 1.000)	0.005 (0.000 - 0.025)	0.400 (0.261 - 0.539)	0.001
-12	0.933 (0.880 - 0.986)	0.884 (0.816 - 0.952)	0.945 (0.897 - 0.993)	0.016 (0.000 - 0.043)	0.321 (0.222 - 0.420)	0.001
-11	0.884 (0.816 - 0.952)	0.800 (0.715 - 0.885)	0.965 (0.926 - 1.000)	0.026 (0.000 - 0.060)	0.253 (0.161 - 0.345)	0.001
-10	0.946 (0.904 - 0.988)	0.917 (0.865 - 0.969)	0.858 (0.792 - 0.924)	0.016 (0.000 - 0.040)	0.474 (0.380 - 0.568)	0.001
-9	0.889 (0.825 - 0.953)	0.787 (0.704 - 0.870)	0.940 (0.892 - 0.988)	0.032 (0.000 - 0.068)	0.345 (0.249 - 0.441)	0.142
-8	0.921 (0.861 - 0.981)	0.872 (0.798 - 0.946)	0.878 (0.805 - 0.951)	0.017 (0.000 - 0.046)	0.547 (0.437 - 0.657)	0.098
-7	0.895 (0.831 - 0.959)	0.874 (0.804 - 0.944)	0.866 (0.794 - 0.938)	0.019 (0.000 - 0.048)	0.539 (0.434 - 0.644)	0.058
-6	0.914 (0.859 - 0.969)	0.891 (0.830 - 0.952)	0.822 (0.747 - 0.897)	0.020 (0.000 - 0.047)	0.563 (0.466 - 0.660)	0.080
-5	0.850 (0.775 - 0.925)	0.761 (0.672 - 0.850)	0.881 (0.813 - 0.949)	0.035 (0.000 - 0.073)	0.541 (0.437 - 0.645)	0.122
-4	0.864 (0.787 - 0.941)	0.831 (0.747 - 0.915)	0.831 (0.747 - 0.915)	0.022 (0.000 - 0.055)	0.642 (0.535 - 0.749)	0.094
-3	0.865 (0.793 - 0.937)	0.791 (0.705 - 0.877)	0.838 (0.760 - 0.916)	0.031 (0.000 - 0.068)	0.616 (0.513 - 0.719)	0.092
-2	0.900 (0.841 - 0.959)	0.842 (0.771 - 0.913)	0.841 (0.770 - 0.912)	0.028 (0.000 - 0.060)	0.553 (0.456 - 0.650)	0.099
-1	0.895 (0.835 - 0.955)	0.832 (0.759 - 0.905)	0.867 (0.801 - 0.933)	0.029 (0.000 - 0.062)	0.512 (0.415 - 0.609)	0.140
Mean ± SD	**0.901 ± 0.033**	**0.848 ± 0.054**	**0.884 ± 0.050**	**0.023 ± 0.008**	**0.485 ± 0.119**	
Model	**RF**					
-13	0.977 (0.935 - 1.000)	0.917 (0.839 - 0.995)	0.967 (0.916 - 1.000)	0.006 (0.000 - 0.028)	0.333 (0.200 - 0.466)	0.245
-12	0.973 (0.939 - 1.000)	0.942 (0.893 - 0.991)	0.925 (0.869 - 0.981)	0.008 (0.000 - 0.027)	0.377 (0.275 - 0.479)	0.100
-11	0.980 (0.950 - 1.000)	0.953 (0.908 - 0.998)	0.910 (0.849 - 0.971)	0.007 (0.000 - 0.025)	0.421 (0.316 - 0.526)	0.140
-10	0.947 (0.905 - 0.989)	0.917 (0.865 - 0.969)	0.858 (0.792 - 0.924)	0.016 (0.000 - 0.040)	0.474 (0.380 - 0.568)	0.100
-9	0.935 (0.885 - 0.985)	0.851 (0.779 - 0.923)	0.877 (0.811 - 0.943)	0.024 (0.000 - 0.055)	0.500 (0.399 - 0.601)	0.160
-8	0.920 (0.860 - 0.980)	0.936 (0.882 - 0.990)	0.775 (0.682 - 0.868)	0.010 (0.000 - 0.032)	0.674 (0.570 - 0.778)	0.105
-7	0.891 (0.826 - 0.956)	0.724 (0.630 - 0.818)	0.897 (0.833 - 0.961)	0.039 (0.000 - 0.080)	0.519 (0.414 - 0.624)	0.235
-6	0.867 (0.801 - 0.933)	0.812 (0.736 - 0.888)	0.796 (0.717 - 0.875)	0.035 (0.000 - 0.071)	0.619 (0.524 - 0.714)	0.130
-5	0.872 (0.802 - 0.942)	0.864 (0.792 - 0.936)	0.710 (0.615 - 0.805)	0.025 (0.000 - 0.058)	0.717 (0.623 - 0.811)	0.105
-4	0.850 (0.770 - 0.930)	0.740 (0.642 - 0.838)	0.813 (0.726 - 0.900)	0.035 (0.000 - 0.076)	0.690 (0.587 - 0.793)	0.170
-3	0.853 (0.778 - 0.928)	0.721 (0.626 - 0.816)	0.822 (0.741 - 0.903)	0.042 (0.000 - 0.084)	0.659 (0.559 - 0.759)	0.185
-2	0.868 (0.802 - 0.934)	0.802 (0.724 - 0.880)	0.777 (0.696 - 0.858)	0.038 (0.001 - 0.075)	0.645 (0.552 - 0.738)	0.125
-1	0.858 (0.790 - 0.926)	0.871 (0.806 - 0.936)	0.694 (0.604 - 0.784)	0.028 (0.000 - 0.060)	0.697 (0.607 - 0.787)	0.095
Mean ± SD	**0.907 ± 0.050**	**0.850 ± 0.084**	**0.832 ± 0.083**	**0.024 ± 0.013**	**0.563 ± 0.133**	
Model	**SVM**					
-13	0.993 (0.969 - 1.000)	0.979 (0.938 - 1.000)	0.943 (0.877 - 1.000)	0.002 (0.000 - 0.015)	0.447 (0.306 - 0.588)	0.040
-12	0.985 (0.959 - 1.000)	0.942 (0.893 - 0.991)	0.972 (0.937 - 1.000)	0.008 (0.000 - 0.027)	0.182 (0.100 - 0.264)	0.105

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-11	0.995 (0.980 - 1.000)	0.988 (0.965 - 1.000)	0.959 (0.917 - 1.000)	0.002 (0.000 - 0.011)	0.243 (0.152 - 0.334)	0.089
-10	0.981 (0.955 - 1.000)	0.927 (0.878 - 0.976)	0.921 (0.870 - 0.972)	0.014 (0.000 - 0.036)	0.331 (0.243 - 0.419)	0.083
-9	0.960 (0.920 - 1.000)	0.926 (0.873 - 0.979)	0.867 (0.798 - 0.936)	0.012 (0.000 - 0.034)	0.500 (0.399 - 0.601)	0.073
-8	0.934 (0.879 - 0.989)	0.872 (0.798 - 0.946)	0.907 (0.843 - 0.971)	0.016 (0.000 - 0.044)	0.477 (0.366 - 0.588)	0.160
-7	0.933 (0.880 - 0.986)	0.920 (0.863 - 0.977)	0.827 (0.748 - 0.906)	0.013 (0.000 - 0.037)	0.590 (0.487 - 0.693)	0.085
-6	0.924 (0.872 - 0.976)	0.891 (0.830 - 0.952)	0.802 (0.724 - 0.880)	0.021 (0.000 - 0.049)	0.589 (0.493 - 0.685)	0.080
-5	0.897 (0.833 - 0.961)	0.886 (0.820 - 0.952)	0.773 (0.685 - 0.861)	0.019 (0.000 - 0.048)	0.659 (0.560 - 0.758)	0.069
-4	0.892 (0.823 - 0.961)	0.857 (0.779 - 0.935)	0.832 (0.748 - 0.916)	0.019 (0.000 - 0.049)	0.633 (0.525 - 0.741)	0.110
-3	0.905 (0.843 - 0.967)	0.814 (0.732 - 0.896)	0.880 (0.811 - 0.949)	0.026 (0.000 - 0.060)	0.536 (0.431 - 0.641)	0.158
-2	0.916 (0.862 - 0.970)	0.901 (0.843 - 0.959)	0.817 (0.742 - 0.892)	0.018 (0.000 - 0.044)	0.571 (0.474 - 0.668)	0.102
-1	0.914 (0.859 - 0.969)	0.891 (0.830 - 0.952)	0.805 (0.728 - 0.882)	0.020 (0.000 - 0.047)	0.589 (0.493 - 0.685)	0.102
Mean ± SD	**0.941 ± 0.038**	**0.907 ± 0.048**	**0.870 ± 0.066**	**0.015 ± 0.007**	**0.488 ± 0.150**	

Table 8.3.B. Performance measures for each model at each lead time before delirium when predicting delirium within 21h since ICU admission, using complete training data and missing values in testing data.

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**BARTm**					
-13	0.989 (0.958 - 1.000)	0.953 (0.890 - 1.000)	0.968 (0.915 - 1.000)	0.003 (0.000 - 0.019)	0.349 (0.207 - 0.491)	0.127
-12	0.979 (0.944 - 1.000)	0.922 (0.856 - 0.988)	0.959 (0.910 - 1.000)	0.007 (0.000 - 0.027)	0.322 (0.208 - 0.436)	0.131
-11	0.972 (0.934 - 1.000)	0.917 (0.853 - 0.981)	0.953 (0.904 - 1.000)	0.009 (0.000 - 0.031)	0.327 (0.219 - 0.435)	0.148
-10	0.973 (0.936 - 1.000)	0.917 (0.853 - 0.981)	0.929 (0.870 - 0.988)	0.009 (0.000 - 0.031)	0.426 (0.312 - 0.540)	0.165
-9	0.965 (0.927 - 1.000)	0.910 (0.851 - 0.969)	0.907 (0.847 - 0.967)	0.013 (0.000 - 0.037)	0.438 (0.335 - 0.541)	0.155
-8	0.935 (0.888 - 0.982)	0.875 (0.811 - 0.939)	0.861 (0.795 - 0.927)	0.022 (0.000 - 0.050)	0.503 (0.407 - 0.599)	0.133
-7	0.920 (0.865 - 0.975)	0.904 (0.844 - 0.964)	0.781 (0.697 - 0.865)	0.017 (0.000 - 0.043)	0.634 (0.537 - 0.731)	0.099
-6	0.913 (0.855 - 0.971)	0.870 (0.801 - 0.939)	0.823 (0.745 - 0.901)	0.021 (0.000 - 0.050)	0.600 (0.500 - 0.700)	0.104
-5	0.891 (0.831 - 0.951)	0.904 (0.847 - 0.961)	0.719 (0.633 - 0.805)	0.020 (0.000 - 0.047)	0.665 (0.574 - 0.756)	0.071
-4	0.900 (0.840 - 0.960)	0.895 (0.833 - 0.957)	0.736 (0.647 - 0.825)	0.020 (0.000 - 0.048)	0.678 (0.584 - 0.772)	0.103
-3	0.910 (0.856 - 0.964)	0.830 (0.758 - 0.902)	0.841 (0.771 - 0.911)	0.031 (0.000 - 0.064)	0.546 (0.451 - 0.641)	0.161
-2	0.872 (0.807 - 0.937)	0.780 (0.699 - 0.861)	0.835 (0.762 - 0.908)	0.037 (0.000 - 0.074)	0.589 (0.493 - 0.685)	0.146
-1	0.873 (0.808 - 0.938)	0.830 (0.756 - 0.904)	0.756 (0.672 - 0.840)	0.032 (0.000 - 0.066)	0.665 (0.572 - 0.758)	0.095
Mean ± SD	**0.930 ± 0.041**	**0.885 ± 0.047**	**0.851 ± 0.087**	**0.023 ± 0.008**	**0.485 ± 0.119**	
Model	**C5.0**					
-13	0.988 (0.955 - 1.000)	0.953 (0.890 - 1.000)	0.977 (0.932 - 1.000)	0.003 (0.000 - 0.019)	0.281 (0.147 - 0.415)	0.273
-12	0.965 (0.920 - 1.000)	0.922 (0.856 - 0.988)	0.938 (0.879 - 0.997)	0.008 (0.000 - 0.030)	0.422 (0.301 - 0.543)	0.186
-11	0.944 (0.891 - 0.997)	0.792 (0.698 - 0.886)	0.981 (0.949 - 1.000)	0.022 (0.000 - 0.056)	0.186 (0.096 - 0.276)	0.358
-10	0.971 (0.932 - 1.000)	0.931 (0.872 - 0.990)	0.897 (0.827 - 0.967)	0.008 (0.000 - 0.029)	0.514 (0.399 - 0.629)	0.198
-9	0.933 (0.881 - 0.985)	0.809 (0.727 - 0.891)	0.932 (0.880 - 0.984)	0.026 (0.000 - 0.059)	0.390 (0.289 - 0.491)	0.277
-8	0.891 (0.831 - 0.951)	0.856 (0.789 - 0.923)	0.776 (0.696 - 0.856)	0.028 (0.000 - 0.060)	0.624 (0.531 - 0.717)	0.162
-7	0.876 (0.809 - 0.943)	0.723 (0.633 - 0.813)	0.878 (0.812 - 0.944)	0.042 (0.001 - 0.083)	0.547 (0.446 - 0.648)	0.254
-6	0.876 (0.809 - 0.943)	0.880 (0.814 - 0.946)	0.734 (0.644 - 0.824)	0.022 (0.000 - 0.052)	0.690 (0.595 - 0.785)	0.154
-5	0.863 (0.797 - 0.929)	0.635 (0.542 - 0.728)	0.920 (0.868 - 0.972)	0.058 (0.013 - 0.103)	0.445 (0.349 - 0.541)	0.294
-4	0.842 (0.769 - 0.915)	0.779 (0.696 - 0.862)	0.777 (0.693 - 0.861)	0.038 (0.000 - 0.076)	0.671 (0.577 - 0.765)	0.186
-3	0.836 (0.766 - 0.906)	0.774 (0.694 - 0.854)	0.807 (0.732 - 0.882)	0.043 (0.004 - 0.082)	0.611 (0.518 - 0.704)	0.189
-2	0.867 (0.800 - 0.934)	0.800 (0.722 - 0.878)	0.811 (0.734 - 0.888)	0.035 (0.000 - 0.071)	0.615 (0.520 - 0.710)	0.188
-1	0.819 (0.744 - 0.894)	0.830 (0.756 - 0.904)	0.671 (0.579 - 0.763)	0.036 (0.000 - 0.073)	0.729 (0.642 - 0.816)	0.099
Mean ± SD	**0.898 ± 0.056**	**0.822 ± 0.089**	**0.854 ± 0.098**	**0.030 ± 0.018**	**0.477 ± 0.144**	

Table 8.3.C. Performance measures for each model at each lead time before delirium when predicting delirium within 21h since ICU admission, using complete training data and median imputation in testing data.

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**AdaBoost**					
-13	0.860 (0.756 - 0.964)	0.640 (0.497 - 0.783)	0.900 (0.810 - 0.990)	0.056 (0.000 - 0.125)	0.515 (0.366 - 0.664)	0.330
-12	0.879 (0.799 - 0.959)	0.800 (0.702 - 0.898)	0.823 (0.729 - 0.917)	0.035 (0.000 - 0.080)	0.600 (0.480 - 0.720)	0.270
-11	0.883 (0.809 - 0.957)	0.830 (0.743 - 0.917)	0.831 (0.744 - 0.918)	0.031 (0.000 - 0.071)	0.562 (0.447 - 0.677)	0.250
-10	0.910 (0.844 - 0.976)	0.853 (0.771 - 0.935)	0.822 (0.734 - 0.910)	0.025 (0.000 - 0.061)	0.599 (0.486 - 0.712)	0.252
-9	0.881 (0.814 - 0.948)	0.865 (0.794 - 0.936)	0.742 (0.651 - 0.833)	0.028 (0.000 - 0.062)	0.656 (0.557 - 0.755)	0.203
-8	0.903 (0.846 - 0.960)	0.870 (0.805 - 0.935)	0.789 (0.711 - 0.867)	0.022 (0.000 - 0.050)	0.641 (0.549 - 0.733)	0.198
-7	0.907 (0.848 - 0.966)	0.862 (0.792 - 0.932)	0.801 (0.720 - 0.882)	0.024 (0.000 - 0.055)	0.623 (0.525 - 0.721)	0.242
-6	0.936 (0.886 - 0.986)	0.894 (0.831 - 0.957)	0.858 (0.787 - 0.929)	0.019 (0.000 - 0.047)	0.503 (0.401 - 0.605)	0.248
-5	0.960 (0.922 - 0.998)	0.888 (0.827 - 0.949)	0.923 (0.872 - 0.974)	0.016 (0.000 - 0.040)	0.397 (0.303 - 0.491)	0.282
-4	0.975 (0.944 - 1.000)	0.944 (0.898 - 0.990)	0.936 (0.887 - 0.985)	0.006 (0.000 - 0.022)	0.393 (0.295 - 0.491)	0.274
-3	0.970 (0.938 - 1.000)	0.889 (0.829 - 0.949)	0.967 (0.933 - 1.000)	0.012 (0.000 - 0.033)	0.264 (0.180 - 0.348)	0.307
-2	0.984 (0.959 - 1.000)	0.938 (0.891 - 0.985)	0.948 (0.904 - 0.992)	0.006 (0.000 - 0.021)	0.375 (0.280 - 0.470)	0.213
-1	0.994 (0.979 - 1.000)	0.977 (0.948 - 1.000)	0.981 (0.954 - 1.000)	0.001 (0.000 - 0.007)	0.236 (0.153 - 0.319)	0.273
Mean ± SD	**0.926 ± 0.865**	**0.871 ± 0.022**	**0.490 ± 0.046**	**0.083 ± 0.076**	**0.015 ± 0.143**	
Model	**BARTm**					
-13	0.986 (0.951 - 1.000)	0.953 (0.890 - 1.000)	0.987 (0.953 - 1.000)	0.003 (0.000 - 0.019)	0.180 (0.065 - 0.295)	0.227
-12	0.980 (0.946 - 1.000)	0.891 (0.815 - 0.967)	0.990 (0.966 - 1.000)	0.010 (0.000 - 0.034)	0.109 (0.033 - 0.185)	0.264
-11	0.974 (0.937 - 1.000)	0.931 (0.872 - 0.990)	0.918 (0.855 - 0.981)	0.008 (0.000 - 0.029)	0.455 (0.340 - 0.570)	0.111
-10	0.980 (0.948 - 1.000)	0.958 (0.912 - 1.000)	0.945 (0.892 - 0.998)	0.005 (0.000 - 0.021)	0.355 (0.244 - 0.466)	0.980
-9	0.960 (0.919 - 1.000)	0.944 (0.896 - 0.992)	0.887 (0.821 - 0.953)	0.008 (0.000 - 0.027)	0.475 (0.371 - 0.579)	0.131
-8	0.932 (0.884 - 0.980)	0.865 (0.799 - 0.931)	0.832 (0.760 - 0.904)	0.025 (0.000 - 0.055)	0.552 (0.456 - 0.648)	0.113
-7	0.920 (0.865 - 0.975)	0.840 (0.766 - 0.914)	0.853 (0.781 - 0.925)	0.023 (0.000 - 0.053)	0.556 (0.456 - 0.656)	0.136
-6	0.901 (0.840 - 0.962)	0.761 (0.674 - 0.848)	0.907 (0.848 - 0.966)	0.035 (0.000 - 0.073)	0.474 (0.372 - 0.576)	0.185
-5	0.910 (0.855 - 0.965)	0.846 (0.777 - 0.915)	0.827 (0.754 - 0.900)	0.028 (0.000 - 0.060)	0.567 (0.472 - 0.662)	0.134
-4	0.884 (0.820 - 0.948)	0.716 (0.625 - 0.807)	0.898 (0.837 - 0.959)	0.042 (0.002 - 0.082)	0.504 (0.403 - 0.605)	0.230
-3	0.898 (0.840 - 0.956)	0.858 (0.792 - 0.924)	0.805 (0.730 - 0.880)	0.027 (0.000 - 0.058)	0.588 (0.494 - 0.682)	0.115
-2	0.882 (0.819 - 0.945)	0.730 (0.643 - 0.817)	0.882 (0.819 - 0.945)	0.043 (0.003 - 0.083)	0.523 (0.425 - 0.621)	0.184
-1	0.870 (0.804 - 0.936)	0.840 (0.768 - 0.912)	0.739 (0.653 - 0.825)	0.031 (0.000 - 0.065)	0.678 (0.586 - 0.770)	0.093
Mean ± SD	**0.929 ± 0.856**	**0.882 ± 0.022**	**0.463 ± 0.042**	**0.081 ± 0.072**	**0.014 ± 0.161**	

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**C5.0**					
-13	0.997 (0.981 - 1.000)	0.999 (0.990 - 1.000)	0.961 (0.903 - 1.000)	0.000 (0.000 - 0.000)	0.386 (0.240 - 0.532)	0.169
-12	0.968 (0.925 - 1.000)	0.875 (0.794 - 0.956)	0.984 (0.953 - 1.000)	0.012 (0.000 - 0.039)	0.164 (0.073 - 0.255)	0.340
-11	0.949 (0.898 - 1.000)	0.889 (0.816 - 0.962)	0.943 (0.889 - 0.997)	0.012 (0.000 - 0.037)	0.379 (0.267 - 0.491)	0.222
-10	0.959 (0.913 - 1.000)	0.903 (0.835 - 0.971)	0.935 (0.878 - 0.992)	0.011 (0.000 - 0.035)	0.409 (0.295 - 0.523)	0.243
-9	0.941 (0.892 - 0.990)	0.899 (0.836 - 0.962)	0.886 (0.820 - 0.952)	0.015 (0.000 - 0.040)	0.490 (0.386 - 0.594)	0.214
-8	0.920 (0.868 - 0.972)	0.837 (0.766 - 0.908)	0.903 (0.846 - 0.960)	0.028 (0.000 - 0.060)	0.424 (0.329 - 0.519)	0.257
-7	0.874 (0.807 - 0.941)	0.809 (0.730 - 0.888)	0.804 (0.724 - 0.884)	0.032 (0.000 - 0.068)	0.635 (0.538 - 0.732)	0.185
-6	0.880 (0.814 - 0.946)	0.783 (0.699 - 0.867)	0.840 (0.765 - 0.915)	0.034 (0.000 - 0.071)	0.600 (0.500 - 0.700)	0.208
-5	0.863 (0.797 - 0.929)	0.635 (0.542 - 0.728)	0.920 (0.868 - 0.972)	0.058 (0.013 - 0.103)	0.445 (0.349 - 0.541)	0.294
-4	0.815 (0.737 - 0.893)	0.705 (0.613 - 0.797)	0.824 (0.747 - 0.901)	0.048 (0.005 - 0.091)	0.640 (0.543 - 0.737)	0.278
-3	0.858 (0.792 - 0.924)	0.736 (0.652 - 0.820)	0.847 (0.778 - 0.916)	0.047 (0.007 - 0.087)	0.567 (0.473 - 0.661)	0.237
-2	0.873 (0.808 - 0.938)	0.750 (0.665 - 0.835)	0.835 (0.762 - 0.908)	0.042 (0.003 - 0.081)	0.599 (0.503 - 0.695)	0.207
-1	0.849 (0.779 - 0.919)	0.670 (0.578 - 0.762)	0.874 (0.809 - 0.939)	0.053 (0.009 - 0.097)	0.559 (0.462 - 0.656)	0.291
Mean ± SD	**0.904 ± 0.807**	**0.889 ± 0.030**	**0.484 ± 0.055**	**0.106 ± 0.057**	**0.019 ± 0.135**	
Model	**GBM**					
-13	0.997 (0.981 - 1.000)	0.999 (0.990 - 1.000)	0.946 (0.878 - 1.000)	0.000 (0.000 - 0.000)	0.462 (0.313 - 0.611)	0.119
-12	0.976 (0.939 - 1.000)	0.891 (0.815 - 0.967)	0.983 (0.951 - 1.000)	0.010 (0.000 - 0.034)	0.174 (0.081 - 0.267)	0.280
-11	0.968 (0.927 - 1.000)	0.917 (0.853 - 0.981)	0.927 (0.867 - 0.987)	0.009 (0.000 - 0.031)	0.431 (0.317 - 0.545)	0.158
-10	0.975 (0.939 - 1.000)	0.889 (0.816 - 0.962)	0.962 (0.918 - 1.000)	0.012 (0.000 - 0.037)	0.289 (0.184 - 0.394)	0.262
-9	0.962 (0.922 - 1.000)	0.854 (0.781 - 0.927)	0.926 (0.872 - 0.980)	0.020 (0.000 - 0.049)	0.397 (0.295 - 0.499)	0.221
-8	0.939 (0.893 - 0.985)	0.827 (0.754 - 0.900)	0.906 (0.850 - 0.962)	0.029 (0.000 - 0.061)	0.419 (0.324 - 0.514)	0.108
-7	0.926 (0.873 - 0.979)	0.851 (0.779 - 0.923)	0.847 (0.774 - 0.920)	0.024 (0.000 - 0.055)	0.563 (0.463 - 0.663)	0.158
-6	0.902 (0.841 - 0.963)	0.837 (0.762 - 0.912)	0.805 (0.724 - 0.886)	0.027 (0.000 - 0.060)	0.632 (0.533 - 0.731)	0.138
-5	0.910 (0.855 - 0.965)	0.875 (0.811 - 0.939)	0.799 (0.722 - 0.876)	0.024 (0.000 - 0.053)	0.596 (0.502 - 0.690)	0.141
-4	0.905 (0.846 - 0.964)	0.800 (0.720 - 0.880)	0.889 (0.826 - 0.952)	0.031 (0.000 - 0.066)	0.497 (0.396 - 0.598)	0.182
-3	0.891 (0.832 - 0.950)	0.849 (0.781 - 0.917)	0.792 (0.715 - 0.869)	0.029 (0.000 - 0.061)	0.607 (0.514 - 0.700)	0.129
-2	0.891 (0.830 - 0.952)	0.810 (0.733 - 0.887)	0.814 (0.738 - 0.890)	0.033 (0.000 - 0.068)	0.609 (0.513 - 0.705)	0.158
-1	0.865 (0.798 - 0.932)	0.780 (0.699 - 0.861)	0.787 (0.707 - 0.867)	0.040 (0.002 - 0.078)	0.649 (0.555 - 0.743)	0.109
Mean ± SD	**0.931 ± 0.860**	**0.876 ± 0.022**	**0.487 ± 0.041**	**0.057 ± 0.071**	**0.011 ± 0.144**	
Model	**LR**					
-13	0.957 (0.896 - 1.000)	0.907 (0.820 - 0.994)	0.975 (0.928 - 1.000)	0.006 (0.000 - 0.029)	0.304 (0.167 - 0.441)	0.001
-12	0.911 (0.841 - 0.981)	0.844 (0.755 - 0.933)	0.959 (0.910 - 1.000)	0.015 (0.000 - 0.045)	0.341 (0.225 - 0.457)	0.001

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
-11	0.897 (0.827 - 0.967)	0.833 (0.747 - 0.919)	0.920 (0.857 - 0.983)	0.019 (0.000 - 0.051)	0.478 (0.363 - 0.593)	0.001
-10	0.944 (0.891 - 0.997)	0.944 (0.891 - 0.997)	0.887 (0.814 - 0.960)	0.007 (0.000 - 0.026)	0.534 (0.419 - 0.649)	0.059
-9	0.919 (0.862 - 0.976)	0.854 (0.781 - 0.927)	0.907 (0.847 - 0.967)	0.021 (0.000 - 0.051)	0.453 (0.350 - 0.556)	0.085
-8	0.888 (0.827 - 0.949)	0.827 (0.754 - 0.900)	0.884 (0.822 - 0.946)	0.030 (0.000 - 0.063)	0.472 (0.376 - 0.568)	0.087
-7	0.870 (0.802 - 0.938)	0.830 (0.754 - 0.906)	0.824 (0.747 - 0.901)	0.028 (0.000 - 0.061)	0.602 (0.503 - 0.701)	0.065
-6	0.854 (0.782 - 0.926)	0.804 (0.723 - 0.885)	0.843 (0.769 - 0.917)	0.031 (0.000 - 0.066)	0.589 (0.488 - 0.690)	0.067
-5	0.855 (0.787 - 0.923)	0.808 (0.732 - 0.884)	0.794 (0.716 - 0.872)	0.036 (0.000 - 0.072)	0.620 (0.527 - 0.713)	0.046
-4	0.853 (0.782 - 0.924)	0.747 (0.660 - 0.834)	0.848 (0.776 - 0.920)	0.040 (0.001 - 0.079)	0.592 (0.493 - 0.691)	0.144
-3	0.857 (0.790 - 0.924)	0.783 (0.705 - 0.861)	0.862 (0.796 - 0.928)	0.038 (0.002 - 0.074)	0.526 (0.431 - 0.621)	0.142
-2	0.815 (0.739 - 0.891)	0.720 (0.632 - 0.808)	0.826 (0.752 - 0.900)	0.048 (0.006 - 0.090)	0.621 (0.526 - 0.716)	0.101
-1	0.820 (0.745 - 0.895)	0.700 (0.610 - 0.790)	0.846 (0.775 - 0.917)	0.050 (0.007 - 0.093)	0.598 (0.502 - 0.694)	0.134
Mean ± SD	**0.880 ± 0.815**	**0.875 ± 0.028**	**0.518 ± 0.044**	**0.068 ± 0.054**	**0.014 ± 0.104**	
Model	**RF**					
-13	0.993 (0.968 - 1.000)	0.999 (0.990 - 1.000)	0.950 (0.885 - 1.000)	0.000 (0.000 - 0.000)	0.442 (0.294 - 0.590)	0.145
-12	0.966 (0.922 - 1.000)	0.875 (0.794 - 0.956)	0.980 (0.946 - 1.000)	0.012 (0.000 - 0.039)	0.200 (0.102 - 0.298)	0.270
-11	0.956 (0.909 - 1.000)	0.917 (0.853 - 0.981)	0.875 (0.799 - 0.951)	0.010 (0.000 - 0.033)	0.566 (0.452 - 0.680)	0.105
-10	0.966 (0.924 - 1.000)	0.944 (0.891 - 0.997)	0.901 (0.832 - 0.970)	0.006 (0.000 - 0.024)	0.500 (0.385 - 0.615)	0.170
-9	0.945 (0.898 - 0.992)	0.876 (0.808 - 0.944)	0.865 (0.794 - 0.936)	0.018 (0.000 - 0.046)	0.538 (0.434 - 0.642)	0.175
-8	0.915 (0.861 - 0.969)	0.798 (0.721 - 0.875)	0.908 (0.852 - 0.964)	0.034 (0.000 - 0.069)	0.424 (0.329 - 0.519)	0.245
-7	0.882 (0.817 - 0.947)	0.883 (0.818 - 0.948)	0.738 (0.649 - 0.827)	0.022 (0.000 - 0.052)	0.680 (0.586 - 0.774)	0.110
-6	0.881 (0.815 - 0.947)	0.815 (0.736 - 0.894)	0.770 (0.684 - 0.856)	0.032 (0.000 - 0.068)	0.675 (0.579 - 0.771)	0.130
-5	0.859 (0.792 - 0.926)	0.837 (0.766 - 0.908)	0.752 (0.669 - 0.835)	0.033 (0.000 - 0.067)	0.655 (0.564 - 0.746)	0.105
-4	0.879 (0.813 - 0.945)	0.821 (0.744 - 0.898)	0.819 (0.742 - 0.896)	0.030 (0.000 - 0.064)	0.612 (0.514 - 0.710)	0.170
-3	0.861 (0.795 - 0.927)	0.802 (0.726 - 0.878)	0.780 (0.701 - 0.859)	0.039 (0.002 - 0.076)	0.634 (0.542 - 0.726)	0.125
-2	0.880 (0.816 - 0.944)	0.810 (0.733 - 0.887)	0.801 (0.723 - 0.879)	0.034 (0.000 - 0.070)	0.625 (0.530 - 0.720)	0.155
-1	0.853 (0.784 - 0.922)	0.700 (0.610 - 0.790)	0.840 (0.768 - 0.912)	0.050 (0.007 - 0.093)	0.607 (0.511 - 0.703)	0.195
Mean ± SD	**0.910 ± 0.852**	**0.845 ± 0.025**	**0.551 ± 0.049**	**0.076 ± 0.077**	**0.015 ± 0.134**	
Model	**SVM**					
-13	0.996 (0.977 - 1.000)	0.999 (0.990 - 1.000)	0.949 (0.883 - 1.000)	0.000 (0.000 - 0.000)	0.449 (0.300 - 0.598)	0.041
-12	0.979 (0.944 - 1.000)	0.953 (0.901 - 1.000)	0.945 (0.889 - 1.000)	0.005 (0.000 - 0.022)	0.384 (0.265 - 0.503)	0.042
-11	0.969 (0.929 - 1.000)	0.944 (0.891 - 0.997)	0.924 (0.863 - 0.985)	0.006 (0.000 - 0.024)	0.433 (0.319 - 0.547)	0.062
-10	0.981 (0.949 - 1.000)	0.917 (0.853 - 0.981)	0.967 (0.926 - 1.000)	0.009 (0.000 - 0.031)	0.258 (0.157 - 0.359)	0.208
-9	0.960 (0.919 - 1.000)	0.910 (0.851 - 0.969)	0.926 (0.872 - 0.980)	0.013 (0.000 - 0.037)	0.382 (0.281 - 0.483)	0.155

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-8	0.941 (0.896 - 0.986)	0.894 (0.835 - 0.953)	0.876 (0.813 - 0.939)	0.019 (0.000 - 0.045)	0.469 (0.373 - 0.565)	0.102
-7	0.931 (0.880 - 0.982)	0.904 (0.844 - 0.964)	0.789 (0.707 - 0.871)	0.017 (0.000 - 0.043)	0.626 (0.528 - 0.724)	0.078
-6	0.905 (0.845 - 0.965)	0.804 (0.723 - 0.885)	0.886 (0.821 - 0.951)	0.029 (0.000 - 0.063)	0.510 (0.408 - 0.612)	0.135
-5	0.895 (0.836 - 0.954)	0.846 (0.777 - 0.915)	0.851 (0.783 - 0.919)	0.027 (0.000 - 0.058)	0.529 (0.433 - 0.625)	0.118
-4	0.901 (0.841 - 0.961)	0.853 (0.782 - 0.924)	0.833 (0.758 - 0.908)	0.024 (0.000 - 0.055)	0.582 (0.483 - 0.681)	0.132
-3	0.897 (0.839 - 0.955)	0.830 (0.758 - 0.902)	0.832 (0.761 - 0.903)	0.031 (0.000 - 0.064)	0.560 (0.466 - 0.654)	0.104
-2	0.877 (0.813 - 0.941)	0.770 (0.688 - 0.852)	0.812 (0.735 - 0.889)	0.040 (0.002 - 0.078)	0.623 (0.528 - 0.718)	0.112
-1	0.873 (0.808 - 0.938)	0.740 (0.654 - 0.826)	0.848 (0.778 - 0.918)	0.043 (0.003 - 0.083)	0.582 (0.485 - 0.679)	0.145
Mean ± SD	**0.931 ± 0.874**	**0.880 ± 0.020**	**0.491 ± 0.043**	**0.075 ± 0.058**	**0.014 ± 0.109**	

Table 8.3.D. Performance measures for each model at each lead time before delirium when predicting delirium within 21h since ICU admission, using complete training data and 0 imputation in testing data.

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**AdaBoost**					
-13	0.994 (0.971 - 1.000)	0.977 (0.932 - 1.000)	0.980 (0.938 - 1.000)	0.001 (0.000 - 0.010)	0.025 (0.000 - 0.072)	0.273
-12	0.984 (0.953 - 1.000)	0.938 (0.879 - 0.997)	0.944 (0.888 - 1.000)	0.006 (0.000 - 0.025)	0.394 (0.274 - 0.514)	0.215
-11	0.969 (0.929 - 1.000)	0.889 (0.816 - 0.962)	0.962 (0.918 - 1.000)	0.012 (0.000 - 0.037)	0.289 (0.184 - 0.394)	0.307
-10	0.975 (0.939 - 1.000)	0.931 (0.872 - 0.990)	0.953 (0.904 - 1.000)	0.008 (0.000 - 0.029)	0.323 (0.215 - 0.431)	0.299
-9	0.959 (0.918 - 1.000)	0.888 (0.822 - 0.954)	0.921 (0.865 - 0.977)	0.016 (0.000 - 0.042)	0.402 (0.300 - 0.504)	0.282
-8	0.934 (0.886 - 0.982)	0.894 (0.835 - 0.953)	0.850 (0.781 - 0.919)	0.019 (0.000 - 0.045)	0.516 (0.420 - 0.612)	0.248
-7	0.906 (0.847 - 0.965)	0.872 (0.804 - 0.940)	0.796 (0.715 - 0.877)	0.022 (0.000 - 0.052)	0.626 (0.528 - 0.724)	0.229
-6	0.902 (0.841 - 0.963)	0.870 (0.801 - 0.939)	0.789 (0.706 - 0.872)	0.022 (0.000 - 0.052)	0.641 (0.543 - 0.739)	0.198
-5	0.880 (0.818 - 0.942)	0.750 (0.667 - 0.833)	0.854 (0.786 - 0.922)	0.044 (0.005 - 0.083)	0.554 (0.458 - 0.650)	0.280
-4	0.909 (0.851 - 0.967)	0.853 (0.782 - 0.924)	0.820 (0.743 - 0.897)	0.025 (0.000 - 0.056)	0.601 (0.503 - 0.699)	0.252
-3	0.883 (0.822 - 0.944)	0.840 (0.770 - 0.910)	0.829 (0.757 - 0.901)	0.030 (0.000 - 0.062)	0.562 (0.468 - 0.656)	0.250
-2	0.879 (0.815 - 0.943)	0.830 (0.756 - 0.904)	0.789 (0.709 - 0.869)	0.031 (0.000 - 0.065)	0.633 (0.539 - 0.727)	0.250
-1	0.861 (0.793 - 0.929)	0.650 (0.557 - 0.743)	0.898 (0.839 - 0.957)	0.054 (0.010 - 0.098)	0.515 (0.417 - 0.613)	0.330
Mean ± SD	**0.926 ± 0.046**	**0.860 ± 0.084**	**0.876 ± 0.070**	**0.022 ± 0.015**	**0.468 ± 0.178**	
Model	**BARTm**					
-13	0.991 (0.963 - 1.000)	0.977 (0.932 - 1.000)	0.959 (0.900 - 1.000)	0.002 (0.000 - 0.015)	0.400 (0.254 - 0.546)	0.094
-12	0.979 (0.944 - 1.000)	0.938 (0.879 - 0.997)	0.942 (0.885 - 0.999)	0.006 (0.000 - 0.025)	0.400 (0.280 - 0.520)	0.095
-11	0.978 (0.944 - 1.000)	0.931 (0.872 - 0.990)	0.946 (0.894 - 0.998)	0.008 (0.000 - 0.029)	0.356 (0.245 - 0.467)	0.125
-10	0.984 (0.955 - 1.000)	0.931 (0.872 - 0.990)	0.964 (0.921 - 1.000)	0.007 (0.000 - 0.026)	0.272 (0.169 - 0.375)	0.190
-9	0.963 (0.924 - 1.000)	0.933 (0.881 - 0.985)	0.919 (0.862 - 0.976)	0.010 (0.000 - 0.031)	0.399 (0.297 - 0.501)	0.166
-8	0.923 (0.872 - 0.974)	0.865 (0.799 - 0.931)	0.843 (0.773 - 0.913)	0.025 (0.000 - 0.055)	0.536 (0.440 - 0.632)	0.123
-7	0.923 (0.869 - 0.977)	0.840 (0.766 - 0.914)	0.860 (0.790 - 0.930)	0.025 (0.000 - 0.057)	0.543 (0.442 - 0.644)	0.135
-6	0.903 (0.843 - 0.963)	0.870 (0.801 - 0.939)	0.808 (0.728 - 0.888)	0.021 (0.000 - 0.050)	0.619 (0.520 - 0.718)	0.100
-5	0.913 (0.859 - 0.967)	0.865 (0.799 - 0.931)	0.817 (0.743 - 0.891)	0.025 (0.000 - 0.055)	0.575 (0.480 - 0.670)	0.128
-4	0.891 (0.828 - 0.954)	0.800 (0.720 - 0.880)	0.860 (0.790 - 0.930)	0.032 (0.000 - 0.067)	0.556 (0.456 - 0.656)	0.176
-3	0.904 (0.848 - 0.960)	0.858 (0.792 - 0.924)	0.795 (0.718 - 0.872)	0.027 (0.000 - 0.058)	0.601 (0.508 - 0.694)	0.113
-2	0.866 (0.799 - 0.933)	0.770 (0.688 - 0.852)	0.838 (0.766 - 0.910)	0.039 (0.001 - 0.077)	0.588 (0.492 - 0.684)	0.136
-1	0.860 (0.792 - 0.928)	0.810 (0.733 - 0.887)	0.750 (0.665 - 0.835)	0.036 (0.000 - 0.073)	0.676 (0.584 - 0.768)	0.089
Mean ± SD	**0.929 ± 0.045**	**0.876 ± 0.062**	**0.869 ± 0.070**	**0.020 ± 0.012**	**0.502 ± 0.122**	

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**C5.0**					
-13	0.965 (0.910 - 1.000)	0.891 (0.798 - 0.984)	0.971 (0.921 - 1.000)	0.010 (0.000 - 0.040)	0.260 (0.129 - 0.391)	0.277
-12	0.943 (0.886 - 1.000)	0.792 (0.693 - 0.891)	0.981 (0.948 - 1.000)	0.022 (0.000 - 0.058)	0.186 (0.091 - 0.281)	0.358
-11	0.970 (0.931 - 1.000)	0.931 (0.872 - 0.990)	0.897 (0.827 - 0.967)	0.008 (0.000 - 0.029)	0.514 (0.399 - 0.629)	0.198
-10	0.932 (0.874 - 0.990)	0.809 (0.718 - 0.900)	0.930 (0.871 - 0.989)	0.026 (0.000 - 0.063)	0.395 (0.282 - 0.508)	0.277
-9	0.893 (0.829 - 0.957)	0.779 (0.693 - 0.865)	0.852 (0.778 - 0.926)	0.039 (0.000 - 0.079)	0.547 (0.444 - 0.650)	0.195
-8	0.876 (0.813 - 0.939)	0.723 (0.637 - 0.809)	0.876 (0.813 - 0.939)	0.042 (0.003 - 0.081)	0.550 (0.454 - 0.646)	0.254
-7	0.873 (0.806 - 0.940)	0.870 (0.802 - 0.938)	0.731 (0.641 - 0.821)	0.024 (0.000 - 0.055)	0.695 (0.602 - 0.788)	0.154
-6	0.864 (0.794 - 0.934)	0.933 (0.882 - 0.984)	0.626 (0.527 - 0.725)	0.017 (0.000 - 0.043)	0.720 (0.628 - 0.812)	0.094
-5	0.840 (0.770 - 0.910)	0.779 (0.699 - 0.859)	0.774 (0.694 - 0.854)	0.038 (0.001 - 0.075)	0.674 (0.584 - 0.764)	0.186
-4	0.837 (0.763 - 0.911)	0.783 (0.700 - 0.866)	0.804 (0.724 - 0.884)	0.041 (0.001 - 0.081)	0.612 (0.514 - 0.710)	0.189
-3	0.869 (0.805 - 0.933)	0.800 (0.724 - 0.876)	0.808 (0.733 - 0.883)	0.035 (0.000 - 0.070)	0.619 (0.527 - 0.711)	0.188
-2	0.821 (0.746 - 0.896)	0.830 (0.756 - 0.904)	0.668 (0.576 - 0.760)	0.036 (0.000 - 0.073)	0.731 (0.644 - 0.818)	0.099
-1	0.853 (0.784 - 0.922)	0.680 (0.589 - 0.771)	0.874 (0.809 - 0.939)	0.051 (0.008 - 0.094)	0.556 (0.459 - 0.653)	0.291
Mean ± SD	**0.887 ± 0.050**	**0.815 ± 0.075**	**0.830 ± 0.109**	**0.030 ± 0.013**	**0.543 ± 0.170**	
Model	**GBM**					
-13	0.997 (0.981 - 1.000)	0.977 (0.932 - 1.000)	0.966 (0.912 - 1.000)	0.002 (0.000 - 0.015)	0.354 (0.211 - 0.497)	0.119
-12	0.976 (0.939 - 1.000)	0.891 (0.815 - 0.967)	0.981 (0.948 - 1.000)	0.010 (0.000 - 0.034)	0.186 (0.091 - 0.281)	0.280
-11	0.968 (0.927 - 1.000)	0.917 (0.853 - 0.981)	0.927 (0.867 - 0.987)	0.009 (0.000 - 0.031)	0.431 (0.317 - 0.545)	0.159
-10	0.975 (0.939 - 1.000)	0.903 (0.835 - 0.971)	0.945 (0.892 - 0.998)	0.011 (0.000 - 0.035)	0.369 (0.258 - 0.480)	0.252
-9	0.961 (0.921 - 1.000)	0.854 (0.781 - 0.927)	0.926 (0.872 - 0.980)	0.020 (0.000 - 0.049)	0.397 (0.295 - 0.499)	0.221
-8	0.939 (0.893 - 0.985)	0.837 (0.766 - 0.908)	0.903 (0.846 - 0.960)	0.028 (0.000 - 0.060)	0.424 (0.329 - 0.519)	0.108
-7	0.926 (0.873 - 0.979)	0.851 (0.779 - 0.923)	0.847 (0.774 - 0.920)	0.024 (0.000 - 0.055)	0.563 (0.463 - 0.663)	0.158
-6	0.902 (0.841 - 0.963)	0.848 (0.775 - 0.921)	0.804 (0.723 - 0.885)	0.025 (0.000 - 0.057)	0.630 (0.531 - 0.729)	0.142
-5	0.909 (0.854 - 0.964)	0.875 (0.811 - 0.939)	0.793 (0.715 - 0.871)	0.024 (0.000 - 0.053)	0.603 (0.509 - 0.697)	0.141
-4	0.904 (0.845 - 0.963)	0.800 (0.720 - 0.880)	0.888 (0.825 - 0.951)	0.031 (0.000 - 0.066)	0.500 (0.399 - 0.601)	0.182
-3	0.889 (0.829 - 0.949)	0.849 (0.781 - 0.917)	0.786 (0.708 - 0.864)	0.030 (0.000 - 0.062)	0.614 (0.521 - 0.707)	0.129
-2	0.894 (0.834 - 0.954)	0.830 (0.756 - 0.904)	0.811 (0.734 - 0.888)	0.030 (0.000 - 0.063)	0.607 (0.511 - 0.703)	0.158
-1	0.866 (0.799 - 0.933)	0.720 (0.632 - 0.808)	0.858 (0.790 - 0.926)	0.046 (0.005 - 0.087)	0.571 (0.474 - 0.668)	0.109
Mean ± SD	**0.931 ± 0.041**	**0.858 ± 0.061**	**0.880 ± 0.068**	**0.022 ± 0.012**	**0.481 ± 0.134**	
Model	**LR**					
-13	0.920 (0.839 - 1.000)	0.837 (0.727 - 0.947)	0.975 (0.928 - 1.000)	0.010 (0.000 - 0.040)	0.321 (0.181 - 0.461)	0.001
-12	0.905 (0.833 - 0.977)	0.812 (0.716 - 0.908)	0.973 (0.933 - 1.000)	0.018 (0.000 - 0.051)	0.268 (0.159 - 0.377)	0.982

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-11	0.898 (0.828 - 0.968)	0.847 (0.764 - 0.930)	0.911 (0.845 - 0.977)	0.017 (0.000 - 0.047)	0.500 (0.385 - 0.615)	0.001
-10	0.920 (0.857 - 0.983)	0.917 (0.853 - 0.981)	0.900 (0.831 - 0.969)	0.010 (0.000 - 0.033)	0.511 (0.396 - 0.626)	0.091
-9	0.914 (0.856 - 0.972)	0.865 (0.794 - 0.936)	0.898 (0.835 - 0.961)	0.019 (0.000 - 0.047)	0.473 (0.369 - 0.577)	0.085
-8	0.877 (0.814 - 0.940)	0.808 (0.732 - 0.884)	0.887 (0.826 - 0.948)	0.033 (0.000 - 0.067)	0.472 (0.376 - 0.568)	0.087
-7	0.868 (0.800 - 0.936)	0.819 (0.741 - 0.897)	0.824 (0.747 - 0.901)	0.030 (0.000 - 0.064)	0.605 (0.506 - 0.704)	0.065
-6	0.849 (0.776 - 0.922)	0.815 (0.736 - 0.894)	0.839 (0.764 - 0.914)	0.029 (0.000 - 0.063)	0.592 (0.492 - 0.692)	0.067
-5	0.848 (0.779 - 0.917)	0.817 (0.743 - 0.891)	0.787 (0.708 - 0.866)	0.035 (0.000 - 0.070)	0.626 (0.533 - 0.719)	0.046
-4	0.847 (0.775 - 0.919)	0.747 (0.660 - 0.834)	0.842 (0.769 - 0.915)	0.040 (0.001 - 0.079)	0.601 (0.503 - 0.699)	0.144
-3	0.850 (0.782 - 0.918)	0.783 (0.705 - 0.861)	0.855 (0.788 - 0.922)	0.039 (0.002 - 0.076)	0.539 (0.444 - 0.634)	0.142
-2	0.810 (0.733 - 0.887)	0.720 (0.632 - 0.808)	0.820 (0.745 - 0.895)	0.048 (0.006 - 0.090)	0.629 (0.534 - 0.724)	0.101
-1	0.823 (0.748 - 0.898)	0.710 (0.621 - 0.799)	0.840 (0.768 - 0.912)	0.048 (0.006 - 0.090)	0.603 (0.507 - 0.699)	0.134
Mean ± SD	**0.871 ± 0.037**	**0.807 ± 0.057**	**0.873 ± 0.057**	**0.029 ± 0.013**	**0.518 ± 0.115**	
Model	**RF**					
-13	0.993 (0.968 - 1.000)	0.999 (0.990 - 1.000)	0.943 (0.874 - 1.000)	0.000 (0.000 - 0.000)	0.476 (0.327 - 0.625)	0.145
-12	0.964 (0.918 - 1.000)	0.875 (0.794 - 0.956)	0.978 (0.942 - 1.000)	0.012 (0.000 - 0.039)	0.211 (0.111 - 0.311)	0.285
-11	0.955 (0.907 - 1.000)	0.833 (0.747 - 0.919)	0.953 (0.904 - 1.000)	0.018 (0.000 - 0.049)	0.348 (0.238 - 0.458)	0.230
-10	0.966 (0.924 - 1.000)	0.958 (0.912 - 1.000)	0.884 (0.810 - 0.958)	0.005 (0.000 - 0.021)	0.537 (0.422 - 0.652)	0.155
-9	0.944 (0.896 - 0.992)	0.876 (0.808 - 0.944)	0.856 (0.783 - 0.929)	0.019 (0.000 - 0.047)	0.554 (0.451 - 0.657)	0.175
-8	0.913 (0.859 - 0.967)	0.808 (0.732 - 0.884)	0.902 (0.845 - 0.959)	0.032 (0.000 - 0.066)	0.436 (0.341 - 0.531)	0.230
-7	0.881 (0.816 - 0.946)	0.894 (0.832 - 0.956)	0.731 (0.641 - 0.821)	0.020 (0.000 - 0.048)	0.683 (0.589 - 0.777)	0.110
-6	0.879 (0.812 - 0.946)	0.685 (0.590 - 0.780)	0.903 (0.843 - 0.963)	0.045 (0.003 - 0.087)	0.512 (0.410 - 0.614)	0.235
-5	0.854 (0.786 - 0.922)	0.837 (0.766 - 0.908)	0.736 (0.651 - 0.821)	0.034 (0.000 - 0.069)	0.669 (0.579 - 0.759)	0.105
-4	0.875 (0.808 - 0.942)	0.821 (0.744 - 0.898)	0.808 (0.729 - 0.887)	0.030 (0.000 - 0.064)	0.625 (0.528 - 0.722)	0.170
-3	0.857 (0.790 - 0.924)	0.802 (0.726 - 0.878)	0.765 (0.684 - 0.846)	0.039 (0.002 - 0.076)	0.649 (0.558 - 0.740)	0.125
-2	0.880 (0.816 - 0.944)	0.820 (0.745 - 0.895)	0.790 (0.710 - 0.870)	0.033 (0.000 - 0.068)	0.634 (0.540 - 0.728)	0.155
-1	0.852 (0.782 - 0.922)	0.710 (0.621 - 0.799)	0.830 (0.756 - 0.904)	0.049 (0.007 - 0.091)	0.618 (0.523 - 0.713)	0.195
Mean ± SD	**0.909 ± 0.049**	**0.840 ± 0.086**	**0.852 ± 0.083**	**0.026 ± 0.015**	**0.535 ± 0.139**	
Model	**SVM**					
-13	0.993 (0.968 - 1.000)	0.999 (0.990 - 1.000)	0.940 (0.869 - 1.000)	0.000 (0.000 - 0.000)	0.488 (0.339 - 0.637)	0.041
-12	0.977 (0.940 - 1.000)	0.953 (0.901 - 1.000)	0.933 (0.872 - 0.994)	0.005 (0.000 - 0.022)	0.430 (0.309 - 0.551)	0.042
-11	0.966 (0.924 - 1.000)	0.944 (0.891 - 0.997)	0.913 (0.848 - 0.978)	0.006 (0.000 - 0.024)	0.469 (0.354 - 0.584)	0.062
-10	0.978 (0.944 - 1.000)	0.917 (0.853 - 0.981)	0.952 (0.903 - 1.000)	0.009 (0.000 - 0.031)	0.333 (0.224 - 0.442)	0.208
-9	0.957 (0.915 - 0.999)	0.910 (0.851 - 0.969)	0.914 (0.856 - 0.972)	0.013 (0.000 - 0.037)	0.417 (0.315 - 0.519)	0.150

Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-8	0.936 (0.889 - 0.983)	0.894 (0.835 - 0.953)	0.865 (0.799 - 0.931)	0.019 (0.000 - 0.045)	0.489 (0.393 - 0.585)	0.102
-7	0.924 (0.870 - 0.978)	0.904 (0.844 - 0.964)	0.781 (0.697 - 0.865)	0.017 (0.000 - 0.043)	0.634 (0.537 - 0.731)	0.078
-6	0.898 (0.836 - 0.960)	0.804 (0.723 - 0.885)	0.876 (0.809 - 0.943)	0.029 (0.000 - 0.063)	0.532 (0.430 - 0.634)	0.135
-5	0.890 (0.830 - 0.950)	0.846 (0.777 - 0.915)	0.841 (0.771 - 0.911)	0.028 (0.000 - 0.060)	0.546 (0.450 - 0.642)	0.118
-4	0.895 (0.833 - 0.957)	0.853 (0.782 - 0.924)	0.823 (0.746 - 0.900)	0.024 (0.000 - 0.055)	0.597 (0.498 - 0.696)	0.132
-3	0.891 (0.832 - 0.950)	0.830 (0.758 - 0.902)	0.822 (0.749 - 0.895)	0.032 (0.000 - 0.066)	0.575 (0.481 - 0.669)	0.104
-2	0.874 (0.809 - 0.939)	0.770 (0.688 - 0.852)	0.798 (0.719 - 0.877)	0.041 (0.002 - 0.080)	0.640 (0.546 - 0.734)	0.112
-1	0.869 (0.803 - 0.935)	0.740 (0.654 - 0.826)	0.836 (0.763 - 0.909)	0.044 (0.004 - 0.084)	0.600 (0.504 - 0.696)	0.145
Mean ± SD	**0.927 ± 0.043**	**0.874 ± 0.075**	**0.869 ± 0.057**	**0.021 ± 0.014**	**0.519 ± 0.092**	

Table 8.3.E. Performance measures for each model at each lead time before delirium when predicting delirium within 21h since ICU admission, using complete training data and missForest imputation in testing data.

Lead Time	**AUC (95% CI)**	**Sens (95% CI)**	**Spec (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**Cut-off**
Model	**AdaBoost**					
-13	0.994 (0.971 - 1.000)	0.977 (0.932 - 1.000)	0.981 (0.940 - 1.000)	0.001 (0.000 - 0.010)	0.236 (0.109 - 0.363)	0.273
-12	0.984 (0.953 - 1.000)	0.938 (0.879 - 0.997)	0.946 (0.891 - 1.000)	0.006 (0.000 - 0.025)	0.381 (0.262 - 0.500)	0.213
-11	0.969 (0.929 - 1.000)	0.889 (0.816 - 0.962)	0.965 (0.923 - 1.000)	0.012 (0.000 - 0.037)	0.273 (0.170 - 0.376)	0.307
-10	0.975 (0.939 - 1.000)	0.944 (0.891 - 0.997)	0.935 (0.878 - 0.992)	0.006 (0.000 - 0.024)	0.398 (0.285 - 0.511)	0.274
-9	0.960 (0.919 - 1.000)	0.888 (0.822 - 0.954)	0.923 (0.868 - 0.978)	0.016 (0.000 - 0.042)	0.397 (0.295 - 0.499)	0.282
-8	0.936 (0.889 - 0.983)	0.894 (0.835 - 0.953)	0.856 (0.789 - 0.923)	0.019 (0.000 - 0.045)	0.505 (0.409 - 0.601)	0.248
-7	0.907 (0.848 - 0.966)	0.862 (0.792 - 0.932)	0.801 (0.720 - 0.882)	0.024 (0.000 - 0.055)	0.623 (0.525 - 0.721)	0.242
-6	0.903 (0.843 - 0.963)	0.870 (0.801 - 0.939)	0.789 (0.706 - 0.872)	0.022 (0.000 - 0.052)	0.641 (0.543 - 0.739)	0.198
-5	0.881 (0.819 - 0.943)	0.865 (0.799 - 0.931)	0.742 (0.658 - 0.826)	0.028 (0.000 - 0.060)	0.656 (0.565 - 0.747)	0.203
-4	0.910 (0.852 - 0.968)	0.853 (0.782 - 0.924)	0.822 (0.745 - 0.899)	0.025 (0.000 - 0.056)	0.599 (0.500 - 0.698)	0.252
-3	0.883 (0.822 - 0.944)	0.830 (0.758 - 0.902)	0.831 (0.760 - 0.902)	0.031 (0.000 - 0.064)	0.562 (0.468 - 0.656)	0.250
-2	0.879 (0.815 - 0.943)	0.800 (0.722 - 0.878)	0.823 (0.748 - 0.898)	0.035 (0.000 - 0.071)	0.600 (0.504 - 0.696)	0.270
-1	0.860 (0.792 - 0.928)	0.640 (0.546 - 0.734)	0.900 (0.841 - 0.959)	0.056 (0.011 - 0.101)	0.515 (0.417 - 0.613)	0.330
Mean ± SD	**0.926 ± 0.865**	**0.870 ± 0.022**	**0.491 ± 0.046**	**0.083 ± 0.076**	**0.015 ± 0.141**	
Model	**BARTm**					
-13	0.990 (0.960 - 1.000)	0.953 (0.890 - 1.000)	0.981 (0.940 - 1.000)	0.003 (0.000 - 0.019)	0.241 (0.113 - 0.369)	0.168
-12	0.976 (0.939 - 1.000)	0.906 (0.835 - 0.977)	0.980 (0.946 - 1.000)	0.009 (0.000 - 0.032)	0.194 (0.097 - 0.291)	0.153
-11	0.977 (0.942 - 1.000)	0.903 (0.835 - 0.971)	0.952 (0.903 - 1.000)	0.011 (0.000 - 0.035)	0.337 (0.228 - 0.446)	0.138
-10	0.982 (0.951 - 1.000)	0.931 (0.872 - 0.990)	0.964 (0.921 - 1.000)	0.007 (0.000 - 0.026)	0.272 (0.169 - 0.375)	0.187
-9	0.965 (0.927 - 1.000)	0.933 (0.881 - 0.985)	0.898 (0.835 - 0.961)	0.010 (0.000 - 0.031)	0.454 (0.351 - 0.557)	0.143
-8	0.943 (0.898 - 0.988)	0.885 (0.824 - 0.946)	0.879 (0.816 - 0.942)	0.020 (0.000 - 0.047)	0.465 (0.369 - 0.561)	0.138
-7	0.913 (0.856 - 0.970)	0.830 (0.754 - 0.906)	0.847 (0.774 - 0.920)	0.027 (0.000 - 0.060)	0.569 (0.469 - 0.669)	0.127
-6	0.909 (0.850 - 0.968)	0.859 (0.788 - 0.930)	0.811 (0.731 - 0.891)	0.023 (0.000 - 0.054)	0.618 (0.519 - 0.717)	0.101
-5	0.898 (0.840 - 0.956)	0.827 (0.754 - 0.900)	0.794 (0.716 - 0.872)	0.033 (0.000 - 0.067)	0.614 (0.520 - 0.708)	0.110
-4	0.885 (0.821 - 0.949)	0.768 (0.683 - 0.853)	0.870 (0.802 - 0.938)	0.036 (0.000 - 0.073)	0.547 (0.447 - 0.647)	0.187
-3	0.895 (0.837 - 0.953)	0.802 (0.726 - 0.878)	0.867 (0.802 - 0.932)	0.035 (0.000 - 0.070)	0.511 (0.416 - 0.606)	0.895
-2	0.874 (0.809 - 0.939)	0.770 (0.688 - 0.852)	0.863 (0.796 - 0.930)	0.038 (0.001 - 0.075)	0.547 (0.449 - 0.645)	0.159
-1	0.870 (0.804 - 0.936)	0.680 (0.589 - 0.771)	0.894 (0.834 - 0.954)	0.050 (0.007 - 0.093)	0.514 (0.416 - 0.612)	0.204
Mean ± SD	**0.929 ± 0.850**	**0.892 ± 0.023**	**0.453 ± 0.044**	**0.080 ± 0.061**	**0.015 ± 0.144**	

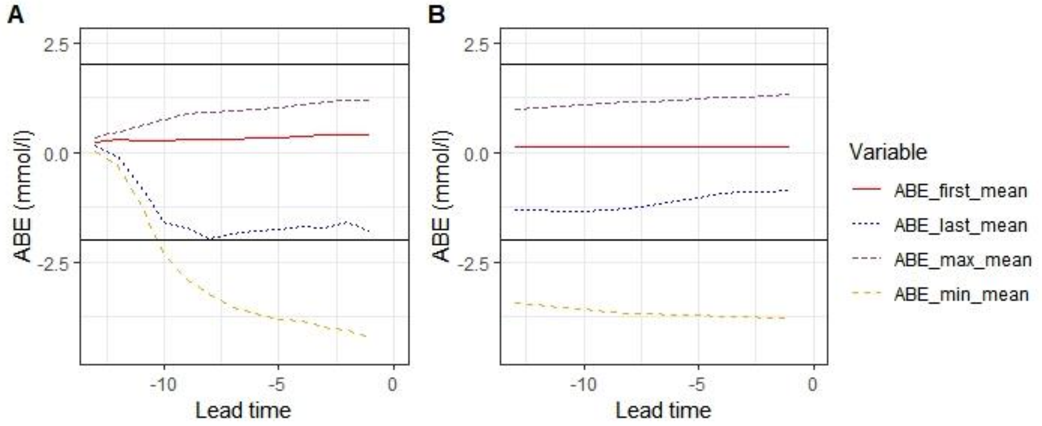
Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
Model	**C5.0**					
-13	0.988 (0.955 - 1.000)	0.953 (0.890 - 1.000)	0.977 (0.932 - 1.000)	0.003 (0.000 - 0.019)	0.281 (0.147 - 0.415)	0.273
-12	0.968 (0.925 - 1.000)	0.875 (0.794 - 0.956)	0.984 (0.953 - 1.000)	0.012 (0.000 - 0.039)	0.164 (0.073 - 0.255)	0.340
-11	0.949 (0.898 - 1.000)	0.889 (0.816 - 0.962)	0.943 (0.889 - 0.997)	0.012 (0.000 - 0.037)	0.379 (0.267 - 0.491)	0.222
-10	0.959 (0.913 - 1.000)	0.903 (0.835 - 0.971)	0.933 (0.875 - 0.991)	0.011 (0.000 - 0.035)	0.414 (0.300 - 0.528)	0.242
-9	0.941 (0.892 - 0.990)	0.899 (0.836 - 0.962)	0.886 (0.820 - 0.952)	0.015 (0.000 - 0.040)	0.490 (0.386 - 0.594)	0.214
-8	0.917 (0.864 - 0.970)	0.837 (0.766 - 0.908)	0.903 (0.846 - 0.960)	0.028 (0.000 - 0.060)	0.424 (0.329 - 0.519)	0.257
-7	0.874 (0.807 - 0.941)	0.809 (0.730 - 0.888)	0.804 (0.724 - 0.884)	0.032 (0.000 - 0.068)	0.635 (0.538 - 0.732)	0.185
-6	0.880 (0.814 - 0.946)	0.783 (0.699 - 0.867)	0.840 (0.765 - 0.915)	0.034 (0.000 - 0.071)	0.600 (0.500 - 0.700)	0.208
-5	0.863 (0.797 - 0.929)	0.635 (0.542 - 0.728)	0.920 (0.868 - 0.972)	0.058 (0.013 - 0.103)	0.445 (0.349 - 0.541)	0.294
-4	0.815 (0.737 - 0.893)	0.705 (0.613 - 0.797)	0.824 (0.747 - 0.901)	0.048 (0.005 - 0.091)	0.640 (0.543 - 0.737)	0.278
-3	0.858 (0.792 - 0.924)	0.736 (0.652 - 0.820)	0.847 (0.778 - 0.916)	0.047 (0.007 - 0.087)	0.567 (0.473 - 0.661)	0.237
-2	0.873 (0.808 - 0.938)	0.750 (0.665 - 0.835)	0.835 (0.762 - 0.908)	0.042 (0.003 - 0.081)	0.599 (0.503 - 0.695)	0.207
-1	0.849 (0.779 - 0.919)	0.670 (0.578 - 0.762)	0.874 (0.809 - 0.939)	0.053 (0.009 - 0.097)	0.559 (0.462 - 0.656)	0.291
Mean ± SD	**0.903 ± 0.803**	**0.890 ± 0.030**	**0.477 ± 0.054**	**0.099 ± 0.059**	**0.018 ± 0.144**	
Model	**GBM**					
-13	0.997 (0.981 - 1.000)	0.999 (0.990 - 1.000)	0.946 (0.878 - 1.000)	0.000 (0.000 - 0.000)	0.462 (0.313 - 0.611)	0.119
-12	0.976 (0.939 - 1.000)	0.891 (0.815 - 0.967)	0.983 (0.951 - 1.000)	0.010 (0.000 - 0.034)	0.174 (0.081 - 0.267)	0.280
-11	0.968 (0.927 - 1.000)	0.917 (0.853 - 0.981)	0.927 (0.867 - 0.987)	0.009 (0.000 - 0.031)	0.431 (0.317 - 0.545)	0.158
-10	0.975 (0.939 - 1.000)	0.889 (0.816 - 0.962)	0.961 (0.916 - 1.000)	0.012 (0.000 - 0.037)	0.297 (0.191 - 0.403)	0.270
-9	0.962 (0.922 - 1.000)	0.854 (0.781 - 0.927)	0.924 (0.869 - 0.979)	0.020 (0.000 - 0.049)	0.402 (0.300 - 0.504)	0.221
-8	0.939 (0.893 - 0.985)	0.837 (0.766 - 0.908)	0.906 (0.850 - 0.962)	0.028 (0.000 - 0.060)	0.416 (0.321 - 0.511)	0.108
-7	0.926 (0.873 - 0.979)	0.851 (0.779 - 0.923)	0.847 (0.774 - 0.920)	0.024 (0.000 - 0.055)	0.563 (0.463 - 0.663)	0.158
-6	0.902 (0.841 - 0.963)	0.837 (0.762 - 0.912)	0.805 (0.724 - 0.886)	0.027 (0.000 - 0.060)	0.632 (0.533 - 0.731)	0.138
-5	0.910 (0.855 - 0.965)	0.875 (0.811 - 0.939)	0.799 (0.722 - 0.876)	0.024 (0.000 - 0.053)	0.596 (0.502 - 0.690)	0.141
-4	0.905 (0.846 - 0.964)	0.800 (0.720 - 0.880)	0.889 (0.826 - 0.952)	0.031 (0.000 - 0.066)	0.497 (0.396 - 0.598)	0.182
-3	0.891 (0.832 - 0.950)	0.849 (0.781 - 0.917)	0.792 (0.715 - 0.869)	0.029 (0.000 - 0.061)	0.607 (0.514 - 0.700)	0.129
-2	0.891 (0.830 - 0.952)	0.810 (0.733 - 0.887)	0.814 (0.738 - 0.890)	0.033 (0.000 - 0.068)	0.609 (0.513 - 0.705)	0.158
-1	0.865 (0.798 - 0.932)	0.780 (0.699 - 0.861)	0.787 (0.707 - 0.867)	0.040 (0.002 - 0.078)	0.649 (0.555 - 0.743)	0.109
Mean ± SD	**0.931 ± 0.861**	**0.875 ± 0.022**	**0.487 ± 0.041**	**0.057 ± 0.071**	**0.011 ± 0.143**	
Model	**LR**					
-13	0.957 (0.896 - 1.000)	0.907 (0.820 - 0.994)	0.975 (0.928 - 1.000)	0.006 (0.000 - 0.029)	0.304 (0.167 - 0.441)	0.001
-12	0.911 (0.841 - 0.981)	0.828 (0.736 - 0.920)	0.981 (0.948 - 1.000)	0.016 (0.000 - 0.047)	0.197 (0.100 - 0.294)	0.982

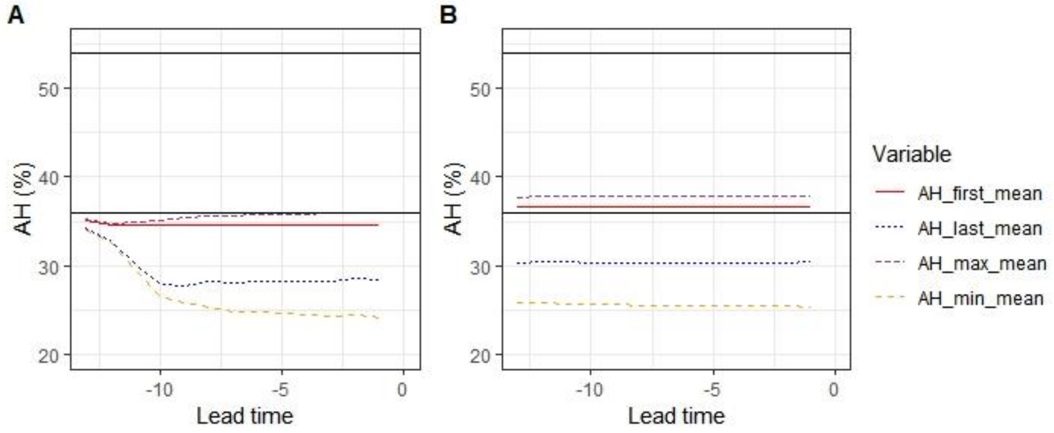
Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-11	0.903 (0.835 - 0.971)	0.847 (0.764 - 0.930)	0.920 (0.857 - 0.983)	0.017 (0.000 - 0.047)	0.474 (0.359 - 0.589)	0.001
-10	0.943 (0.889 - 0.997)	0.931 (0.872 - 0.990)	0.895 (0.824 - 0.966)	0.008 (0.000 - 0.029)	0.518 (0.403 - 0.633)	0.091
-9	0.920 (0.864 - 0.976)	0.865 (0.794 - 0.936)	0.905 (0.844 - 0.966)	0.019 (0.000 - 0.047)	0.454 (0.351 - 0.557)	0.085
-8	0.879 (0.816 - 0.942)	0.817 (0.743 - 0.891)	0.884 (0.822 - 0.946)	0.032 (0.000 - 0.066)	0.475 (0.379 - 0.571)	0.087
-7	0.868 (0.800 - 0.936)	0.830 (0.754 - 0.906)	0.824 (0.747 - 0.901)	0.028 (0.000 - 0.061)	0.602 (0.503 - 0.701)	0.065
-6	0.853 (0.781 - 0.925)	0.804 (0.723 - 0.885)	0.845 (0.771 - 0.919)	0.031 (0.000 - 0.066)	0.587 (0.486 - 0.688)	0.067
-5	0.843 (0.773 - 0.913)	0.798 (0.721 - 0.875)	0.796 (0.719 - 0.873)	0.038 (0.001 - 0.075)	0.621 (0.528 - 0.714)	0.046
-4	0.857 (0.787 - 0.927)	0.758 (0.672 - 0.844)	0.851 (0.779 - 0.923)	0.038 (0.000 - 0.076)	0.584 (0.485 - 0.683)	0.144
-3	0.860 (0.794 - 0.926)	0.783 (0.705 - 0.861)	0.865 (0.800 - 0.930)	0.038 (0.002 - 0.074)	0.520 (0.425 - 0.615)	0.142
-2	0.815 (0.739 - 0.891)	0.720 (0.632 - 0.808)	0.827 (0.753 - 0.901)	0.048 (0.006 - 0.090)	0.619 (0.524 - 0.714)	0.101
-1	0.823 (0.748 - 0.898)	0.700 (0.610 - 0.790)	0.846 (0.775 - 0.917)	0.050 (0.007 - 0.093)	0.598 (0.502 - 0.694)	0.134
Mean ± SD	**0.879 ± 0.814**	**0.878 ± 0.028**	**0.504 ± 0.044**	**0.066 ± 0.056**	**0.014 ± 0.128**	
Model	**RF**					
-13	0.993 (0.968 - 1.000)	0.999 (0.990 - 1.000)	0.949 (0.883 - 1.000)	0.000 (0.000 - 0.000)	0.449 (0.300 - 0.598)	0.145
-12	0.966 (0.922 - 1.000)	0.875 (0.794 - 0.956)	0.978 (0.942 - 1.000)	0.012 (0.000 - 0.039)	0.211 (0.111 - 0.311)	0.265
-11	0.956 (0.909 - 1.000)	0.917 (0.853 - 0.981)	0.875 (0.799 - 0.951)	0.010 (0.000 - 0.033)	0.566 (0.452 - 0.680)	0.105
-10	0.966 (0.924 - 1.000)	0.944 (0.891 - 0.997)	0.900 (0.831 - 0.969)	0.006 (0.000 - 0.024)	0.504 (0.389 - 0.619)	0.170
-9	0.945 (0.898 - 0.992)	0.876 (0.808 - 0.944)	0.865 (0.794 - 0.936)	0.018 (0.000 - 0.046)	0.538 (0.434 - 0.642)	0.175
-8	0.915 (0.861 - 0.969)	0.798 (0.721 - 0.875)	0.908 (0.852 - 0.964)	0.034 (0.000 - 0.069)	0.424 (0.329 - 0.519)	0.245
-7	0.882 (0.817 - 0.947)	0.883 (0.818 - 0.948)	0.740 (0.651 - 0.829)	0.022 (0.000 - 0.052)	0.678 (0.584 - 0.772)	0.110
-6	0.880 (0.814 - 0.946)	0.815 (0.736 - 0.894)	0.768 (0.682 - 0.854)	0.032 (0.000 - 0.068)	0.677 (0.581 - 0.773)	0.130
-5	0.859 (0.792 - 0.926)	0.837 (0.766 - 0.908)	0.752 (0.669 - 0.835)	0.033 (0.000 - 0.067)	0.655 (0.564 - 0.746)	0.105
-4	0.878 (0.812 - 0.944)	0.821 (0.744 - 0.898)	0.819 (0.742 - 0.896)	0.030 (0.000 - 0.064)	0.612 (0.514 - 0.710)	0.170
-3	0.861 (0.795 - 0.927)	0.802 (0.726 - 0.878)	0.780 (0.701 - 0.859)	0.039 (0.002 - 0.076)	0.634 (0.542 - 0.726)	0.125
-2	0.880 (0.816 - 0.944)	0.810 (0.733 - 0.887)	0.801 (0.723 - 0.879)	0.034 (0.000 - 0.070)	0.625 (0.530 - 0.720)	0.155
-1	0.853 (0.784 - 0.922)	0.700 (0.610 - 0.790)	0.840 (0.768 - 0.912)	0.050 (0.007 - 0.093)	0.607 (0.511 - 0.703)	0.195
Mean ± SD	**0.910 ± 0.852**	**0.844 ± 0.025**	**0.552 ± 0.049**	**0.076 ± 0.076**	**0.015 ± 0.131**	
Model	**SVM**					
-13	0.996 (0.977 - 1.000)	0.999 (0.990 - 1.000)	0.949 (0.883 - 1.000)	0.000 (0.000 - 0.000)	0.449 (0.300 - 0.598)	0.041
-12	0.979 (0.944 - 1.000)	0.953 (0.901 - 1.000)	0.944 (0.888 - 1.000)	0.005 (0.000 - 0.022)	0.390 (0.271 - 0.509)	0.042
-11	0.970 (0.931 - 1.000)	0.944 (0.891 - 0.997)	0.924 (0.863 - 0.985)	0.006 (0.000 - 0.024)	0.433 (0.319 - 0.547)	0.062
-10	0.981 (0.949 - 1.000)	0.917 (0.853 - 0.981)	0.964 (0.921 - 1.000)	0.009 (0.000 - 0.031)	0.275 (0.172 - 0.378)	0.208
-9	0.960 (0.919 - 1.000)	0.910 (0.851 - 0.969)	0.926 (0.872 - 0.980)	0.013 (0.000 - 0.037)	0.382 (0.281 - 0.483)	0.154

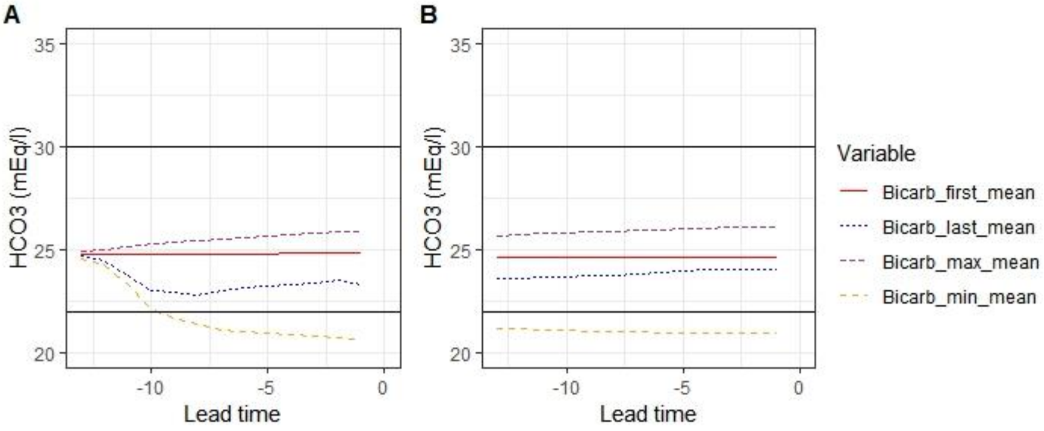
Lead Time	AUC (95% CI)	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	Cut-off
-8	0.941 (0.896 - 0.986)	0.894 (0.835 - 0.953)	0.877 (0.814 - 0.940)	0.019 (0.000 - 0.045)	0.466 (0.370 - 0.562)	0.102
-7	0.931 (0.880 - 0.982)	0.904 (0.844 - 0.964)	0.790 (0.708 - 0.872)	0.017 (0.000 - 0.043)	0.624 (0.526 - 0.722)	0.078
-6	0.905 (0.845 - 0.965)	0.804 (0.723 - 0.885)	0.886 (0.821 - 0.951)	0.029 (0.000 - 0.063)	0.510 (0.408 - 0.612)	0.135
-5	0.896 (0.837 - 0.955)	0.846 (0.777 - 0.915)	0.851 (0.783 - 0.919)	0.027 (0.000 - 0.058)	0.529 (0.433 - 0.625)	0.118
-4	0.901 (0.841 - 0.961)	0.853 (0.782 - 0.924)	0.833 (0.758 - 0.908)	0.024 (0.000 - 0.055)	0.582 (0.483 - 0.681)	0.132
-3	0.897 (0.839 - 0.955)	0.830 (0.758 - 0.902)	0.832 (0.761 - 0.903)	0.031 (0.000 - 0.064)	0.560 (0.466 - 0.654)	0.104
-2	0.877 (0.813 - 0.941)	0.770 (0.688 - 0.852)	0.812 (0.735 - 0.889)	0.040 (0.002 - 0.078)	0.623 (0.528 - 0.718)	0.112
-1	0.873 (0.808 - 0.938)	0.740 (0.654 - 0.826)	0.848 (0.778 - 0.918)	0.043 (0.003 - 0.083)	0.582 (0.485 - 0.679)	0.145
Mean ± SD	**0.931 ± 0.874**	**0.880 ± 0.020**	**0.493 ± 0.043**	**0.075 ± 0.057**	**0.014 ± 0.105**	

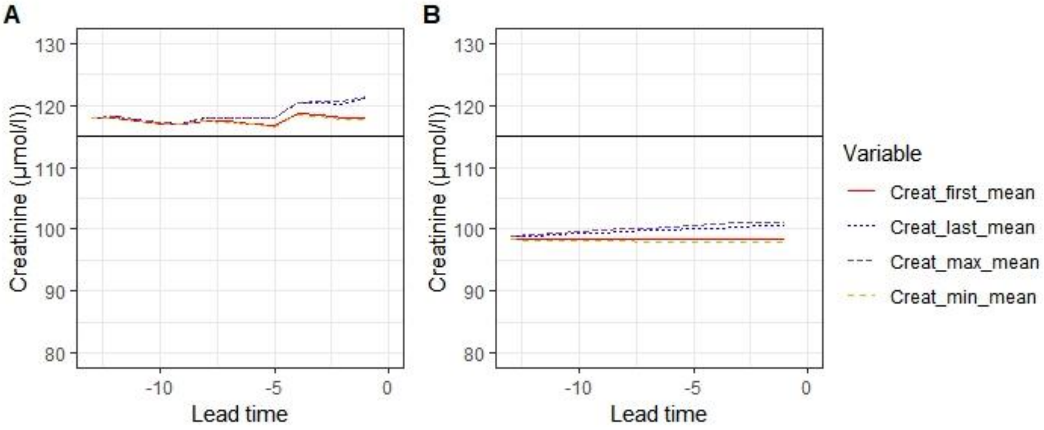
Appendix 8.4: How Laboratory Values Change as the Lead Times Change for Delirium vs non-Delirium Patients

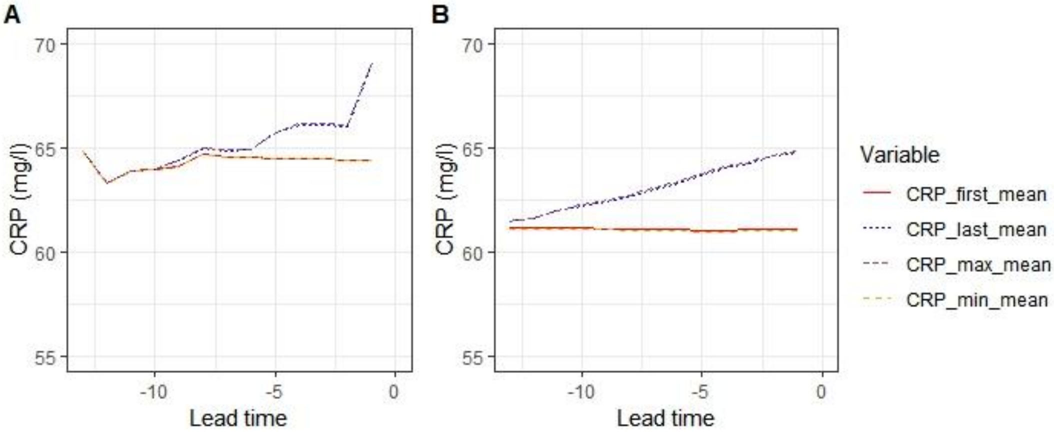
The changing of laboratory variable summary values as the lead time changes for patients with delirium (A) and for patients without delirium (B). The horizontal lines indicate the normal range for each respective laboratory value. Abbreviations: ABE, arterial base excess; AH, arterial haematocrit; CRP, C-Reactive Protein; HCO3, bicarbonate; H+, Hydrogen ion.

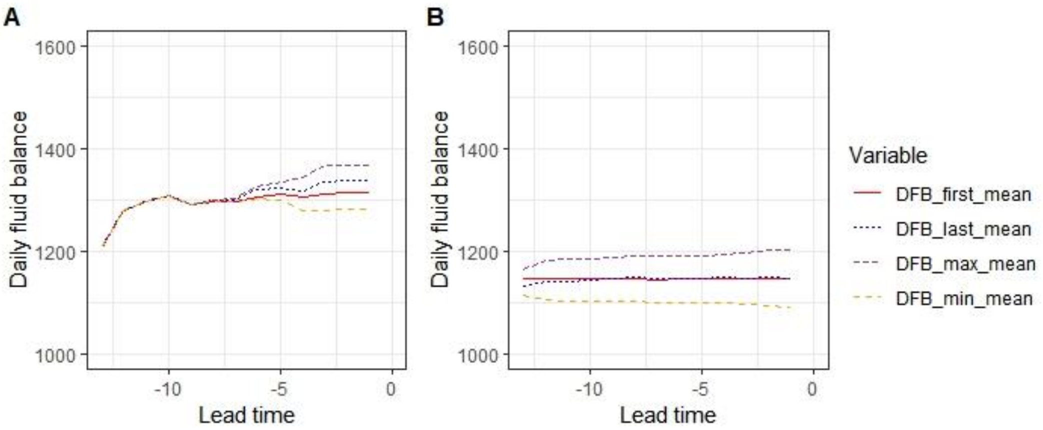


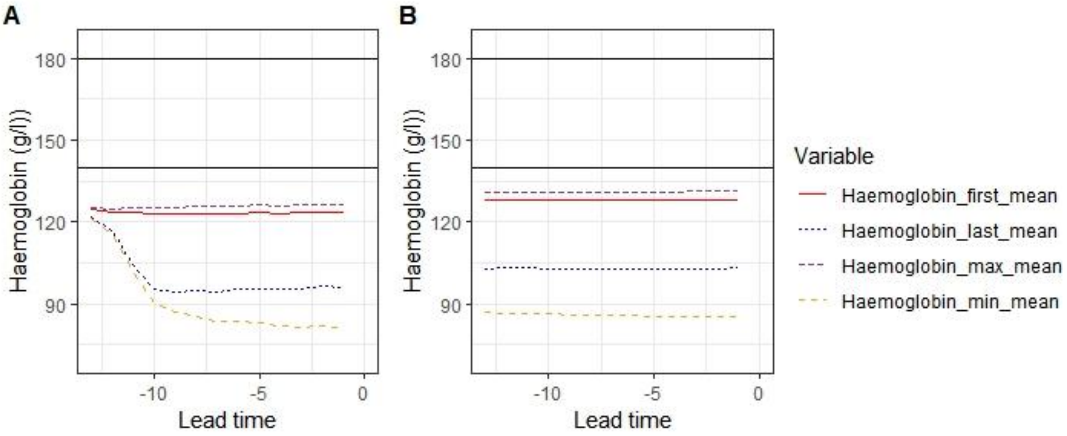


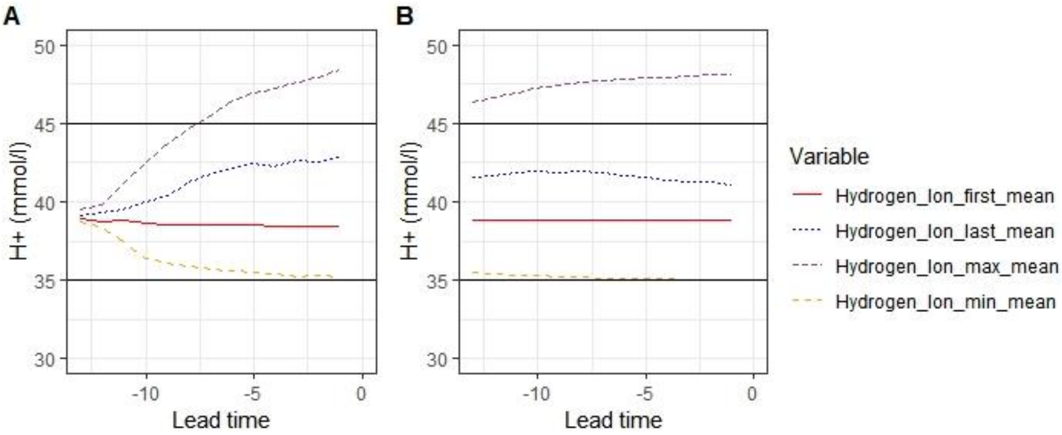

A
B
HCO3 (mEq/l)
35
30
25
20
-10
-5
0
Lead time
Variable
Bicarb_first_mean
Bicarb_last_mean
Bicarb_max_mean
Bicarb_min_mean

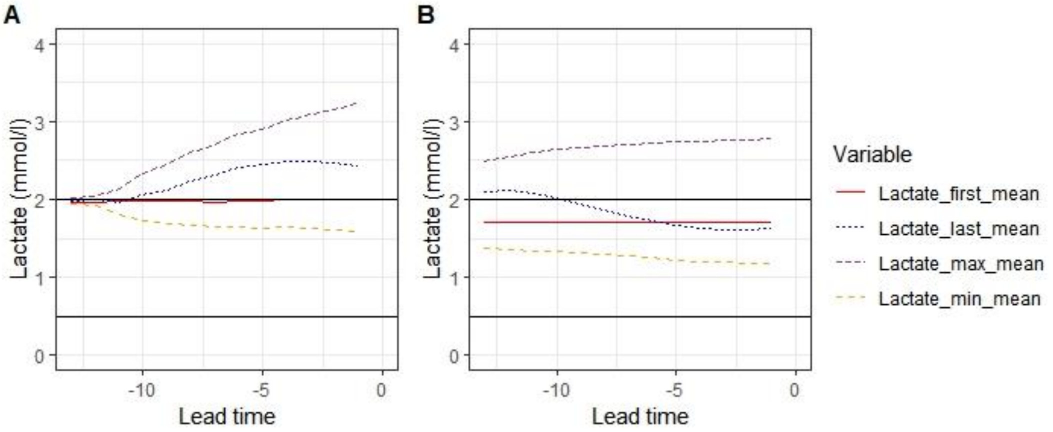

A
B
Creatinine (µmol/l))
130
120
110
100
90
80
-10
-5
0
Lead time
Variable
Creat_first_mean
Creat_last_mean
Creat_max_mean
Creat_min_mean

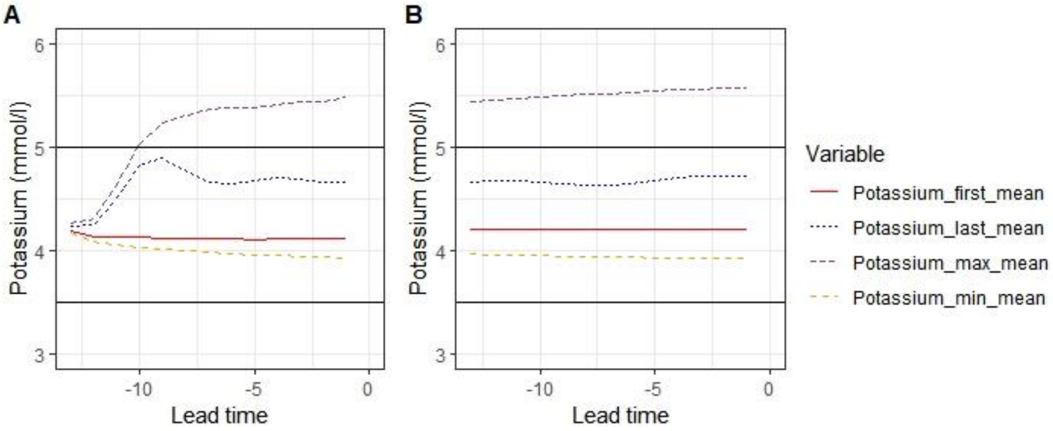

A
B
CRP (mg/l)
70
65
60
55
-10
-5
0
Lead time
Variable
CRP_first_mean
CRP_last_mean
CRP_max_mean
CRP_min_mean

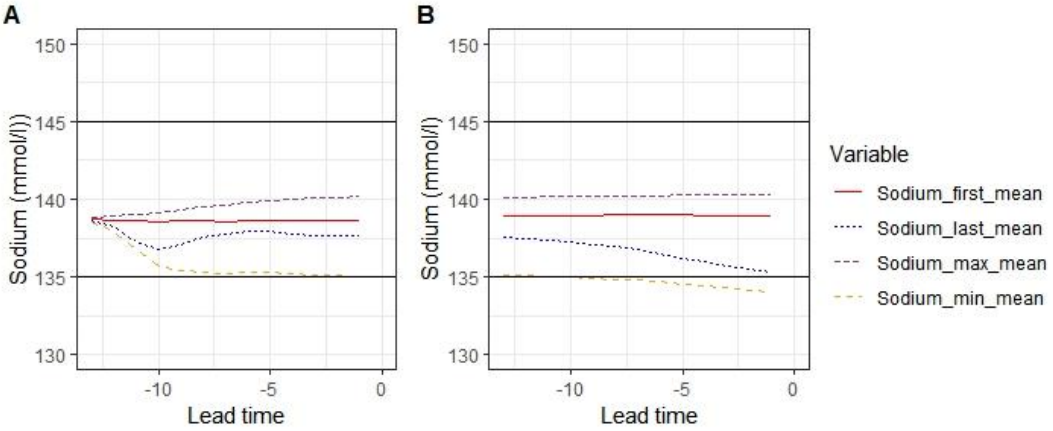

A
B
Daily fluid balance
1600
1400
1200
1000
-10
-5
0
Lead time
Variable
DFB_first_mean
DFB_last_mean
DFB_max_mean
DFB_min_mean

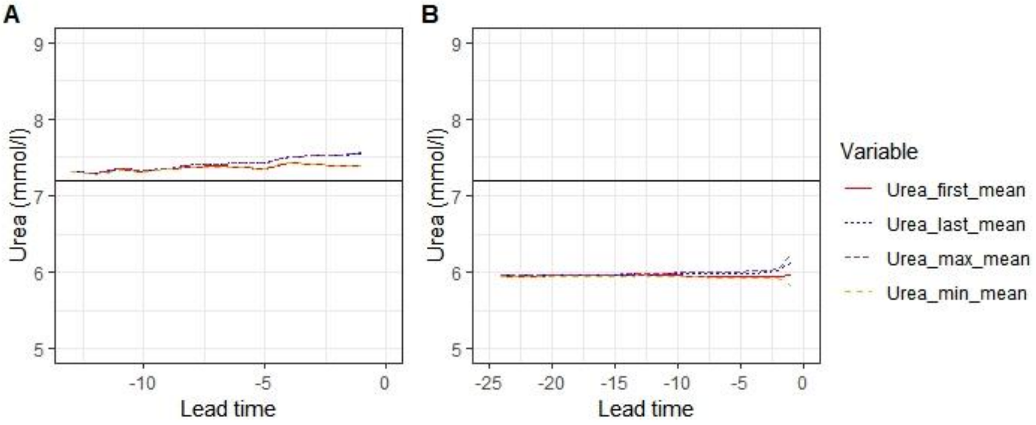

A
B
Haemoglobin (g/l))
180
150
120
90
-10
-5
0
Lead time
Variable
Haemoglobin_first_mean
Haemoglobin_last_mean
Haemoglobin_max_mean
Haemoglobin_min_mean

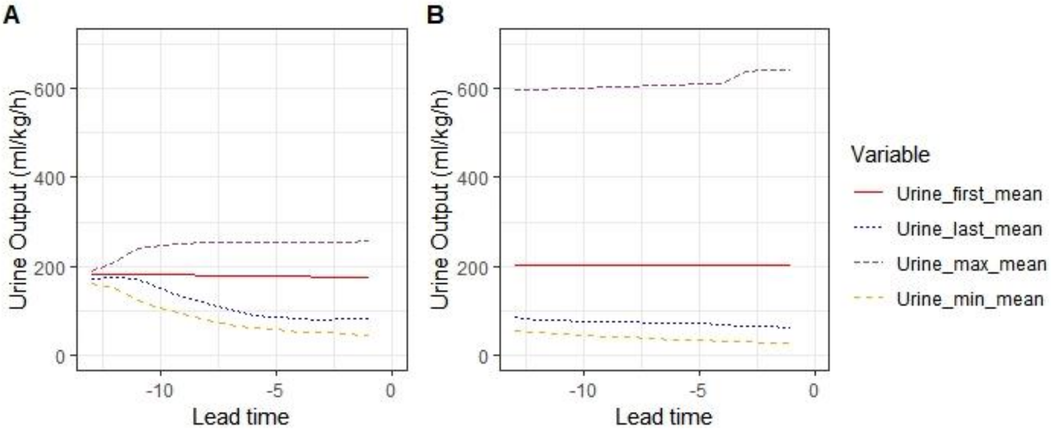

A
B
H+ (mmol/l)
50
45
40
35
30
-10
-5
0
Lead time
Variable
Hydrogen_Ion_first_mean
Hydrogen_Ion_last_mean
Hydrogen_Ion_max_mean
Hydrogen_Ion_min_mean


A
B
Lactate (mmol/l)
0
1
2
3
4
-10
-5
0
Lead time
Variable
Lactate_first_mean
Lactate_last_mean
Lactate_max_mean
Lactate_min_mean


A
B
Potassium (mmol/l)
3
4
5
6
-10
-5
0
Lead time
Variable
Potassium_first_mean
Potassium_last_mean
Potassium_max_mean
Potassium_min_mean


A
B
Sodium (mmol/l))
Sodium (mmol/l)
130
135
140
145
150
-10
-5
0
Lead time
Variable
Sodium_first_mean
Sodium_last_mean
Sodium_max_mean
Sodium_min_mean


A
B
Urea (mmol/l)
9
8
7
6
5
-10
-5
0
-25
-20
-15
Lead time
Variable
Urea_first_mean
Urea_last_mean
Urea_max_mean
Urea_min_mean


A
B
Urine Output (ml/kg/h)
600
400
200
0
-10
-5
0
Lead time
Variable
Urine_first_mean
Urine_last_mean
Urine_max_mean
Urine_min_mean

Appendix 8.5: Performance of Static Models Predicting Delirium in ICU

Here the performance measures of the models predicting delirium in ICU in a static manner are presented. Because the prediction was made in a static manner, the time of delirium occurrence was not taken into account. In addition, minimum, maximum, first and last laboratory values were included for all patients based on the total ICU stay.

Model	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Cut-off**
AB	0.883	0.893	0.726	0.022	0.664	0.204
BARTm	0.905	0.806	0.852	0.034	0.541	0.151
C5.0	0.886	0.903	0.745	0.020	0.645	0.153
GBM	0.897	0.825	0.833	0.032	0.566	0.762
LR	0.834	0.689	0.846	0.054	0.590	0.123
RF	0.863	0.874	0.705	0.027	0.685	0.105
SVM	0.905	0.835	0.825	0.030	0.574	0.105