DOCTORAL THESIS

Survival Analysis Techniques and Applications for Hospital Readmission Modelling.

Todd, James

Award date:
2022

Survival Analysis Techniques and Applications for Hospital Readmission Modelling

James Todd

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

June 2021

Bond Business School

Dr Steven Stern | Professor of Data Science

Dr Bruce Vanstone | Professor of Data Science

Dr Adrian Gepp | Associate Professor of Data Analytics

This research was supported by an Australian Government Research Training Program Scholarship and Healthcare Logic Pty Ltd via Australian Innovation Connections Grants (ICG00433 and ICG000945).

Abstract

Hospital readmissions lead to greater demand for healthcare resources, financial costs, and poorer patient outcomes. They are also often preventable with improved care and management. This has led to their use as a quality-of-care indicator and the development of healthcare policy linking readmission outcomes to hospital funding in the USA, England, Germany, and most recently in Australia. The negative consequences of avoidable readmissions for funding and patient welfare have also spurred development of computational models predicting readmission risk to enable hospitals to identify high-risk patients for interventions.

These prediction models have been overwhelmingly characterised by classification approaches focusing on predicting readmission status at a single fixed time after discharge, most commonly 30 days. Research developing and validating these models is ongoing, driven by the poor performance of models currently used and the need for customisation to specific regions, populations, and conditions. To improve the performance of these models, research has increasingly considered machine learning techniques and leveraged novel data sources. Despite the additional information provided by survival models compared with classification models, survival approaches have received little attention with respect to available machine learning techniques, practical applications, and appropriate performance measures.

This research identified available and relevant machine learning survival techniques, including decision trees, ensembles, and artificial neural networks. The value of previously unconsidered machine learning survival techniques was investigated for predicting 30-day unplanned readmissions. This investigation considered adult patients admitted to hospital through the emergency department of Gold Coast University Hospital (n = 46,659) and Robina Hospital (n = 23,976) in Queensland, Australia. The value of both statistical and machine learning survival models for novel applications supporting managerial decision-making were also investigated. The proposed applications leverage survival predictions to dynamically rank patients by risk, account for patient-specific risk profiles, and forecast future readmissions. The important aspects of model performance for such applications were determined to be discrimination and calibration of predictions over time. Time-dependent concordance and D-calibration were identified as appropriate metrics capturing these aspects of model performance.

For the more complex population (Robina Hospital), machine learning survival models improved on statistical survival models for both 30-day readmission and risk over time prediction. Even compared to the benchmark classification model, select machine learning models exhibited competitive discrimination and better calibration in predicting 30-day readmissions. These models should be considered when developing tools supporting readmission management under classification approaches as well as survival approaches. The proposed model applications under survival approaches were demonstrated to be feasible, with varying levels of discrimination but consistent calibration across both machine learning and survival models. These models should benefit hospitals managing readmissions through better intervention targeting, follow-up care customisation, and demand forecasting. This in turn should lead to reduced costs and better outcomes for patients. The area is also advanced more generally through the highlighting of available machine learning techniques, applications, and performance measures under survival approaches.

Declaration by Author

This thesis is submitted to Bond University in fulfilment of the requirements of the degree of Doctor of Philosophy by Research.

I declare that the research presented within this thesis is a product of my own original ideas and work and contains no material which has previously been submitted for a degree at this university or any other institution, except where due acknowledgement has been made.

James Todd

Date: 18/06/2021

Ethics Declaration

The research associated with this thesis received ethics approval from the Bond University Human Research Ethics Committee. Ethics application number JT00253.

Copyright Declaration

No published manuscripts were included for publication within this dissertation.

Acknowledgements

While I cannot thank every person who has helped me complete this journey, I can try to thank those who played the biggest roles.

To my supervisors Steve, Bruce, and Adrian, thank you for your guidance over the past three years. Beyond being a rewarding process, it was fun with our team. You have made it almost, but not quite, a regret to finish. Adrian and Bruce especially, you are obviously a big part of why I finished, but you are just as large a part of why I started. Thank you for having my back since well before PhD.

To my family, especially my mom and dad, thank you for your love, understanding, and support. I am the person I am today because of you. You have always encouraged and supported me in finding my own path and made it easy for me even if I didn't always make it easy for you.

Ronnie, I'm grateful that I've been able to share so many wonderful moments with you. Mitch, you've always been a good friend and helped keep me grounded. Tom and Jason, thank you for getting me out and occasionally even having fun. Milind, you kept me active and moving. Nikki, this document would look substantially worse without you. And to everyone I have not named but were undeniably part of the journey, thank you as well. I am grateful for everyone I have met during my time at Bond and how you have all shaped who I am.

Finally, I am grateful to Bond University and Healthcare Logic for their support that allowed me to dedicate myself to this project.

Table of Contents

List of Figures

List of Tables

Acronyms

Acronym	Expanded Form
AFT	Accelerated Failure Time
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AUC	Area Under the Receiver Operating Characteristic Curve
BART	Bayesian Additive Regression Trees
BIC	Bayesian Information Criterion
CABG	Coronary Artery Bypass Graft
CDF	Cumulative Distribution Function
CDU	Clinical Decision Unit
CHF	Cumulative Hazard Function
COPD	Chronic Obstructive Pulmonary Disease
CURE	Censoring Unbiased Regression Ensemble
CURT	Censoring Unbiased Regression Tree
CUT	Censoring Unbiased Transformation
DOB	Date of Birth
DRG	Diagnosis Related Groups
DRR	Dynamic Risk Ranking
ED	Emergency Department
EN	Elastic Net
ER	Expected Readmissions
ERP	Elevated Risk Period
ERPP	Elevated Risk Period Probability
ERT	Extremely Randomised Tree
GCUH	Gold Coast University Hospital
HCL	Healthcare Logic Pty Ltd
HL	Hosmer-Lemeshow
HOSPITAL	Risk index comprised of components for Haemoglobin, discharge from an Oncology service, Sodium level, Procedure during the index admission, Index Type of admission, number of Admissions in prior 12 months, and Length of stay.
HRRP	Hospital Readmissions Reduction Program

Acronym	Expanded Form
IBS	Integrated Brier Score
ICD	International Classification of Disease
ICU	Intensive Care Unit
IPCW	Inverse Probability of Censoring Weighted
ISAR	Identification of Seniors at Risk
KEN	Kernel Elastic Net
KM	Kaplan-Meier
LACE	Risk index comprised of components for Length of stay, Acuity of admission, Comorbidities, and Emergency Department visits.
LARS	Least Angle Regression
LASSO	Least Absolute Shrinking and Selection Operator
LIME	Local Interpretable Model-Agnostic Explanations
LOS	Length of Stay
LR	Logistic Regression
LT	Life Table
LUPI	Learning Using Privileged Information
MCMC	Markov chain Monte Carlo
MTP	Multiple Time Point
NA	Nelson-Aalen
NLP	Natural Language Processing
OSCAR	Octagonal Shrinkage and Clustering Algorithm for Regression
PLANN	Partial Logistic Artificial Neural Network
PLANN-ARD	Partial Logistic Artificial Neural Network - Automatic Relevance Determination
PLANN-CR	Partial Logistic Artificial Neural Network - Competing Risks
PLANN-CR-ARD	Partial Logistic Artificial Neural Network - Competing Risks - Automatic Relevance Determination
PPM	Power Performance Manager
RF	Random Forest
RH	Robina Hospital
RIST	Recursively Imputed Survival Trees
RNN	Recurrent Neural Network

Acronym	Expanded Form
ROC	Receiver Operating Characteristic
RQ1	Research Question 1
RQ2	Research Question 2
RSF	Random Survival Forest
SCAD	Smoothly Clipped Absolute Deviations
SD	Standard Deviation
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Over-sampling Technique
SSU	Short Stay Unit
SVCR	Support Vector Censored Regression
SVM	Support Vector Machine
SVR	Support Vector Regression
TRST	Triage Risk Screening Tool
UK	United Kingdom
USA	United States of America
VIP	Variable Indicative of Placement

1 Introduction

1.1 Background

A hospital readmission is the event in which a patient who is discharged from a hospital is readmitted again within a short period of time, with the exact period in the Australian setting depending on the readmission condition (Independent Hospital Pricing Authority, 2021b). Unplanned and early readmissions put patients at greater risk of adverse outcomes, burdens limited hospital resources, and imposes costs on the healthcare system (Artetxe, Beristain, & Graña, 2018). Readmissions may also indicate underlying issues in the quality of care being provided to patients before and after their discharge. While substantial differences in healthcare systems around the world makes reliable comparisons difficult, there is acknowledgement across many healthcare systems that patient readmissions are not rare. In the USA Medicare population, a study of over five million hospital admissions in 2008 and 2009 found 17.9% of these admissions resulted in a subsequent acute care encounter, defined as being either an emergency department visit or hospital readmission (Vashi et al., 2013). A study of over 62 million hospital admissions in the English National Health Service found average risk-adjusted 30-day readmission rates rose slightly from 6.56% to 6.64% between 2006 and 2016 (Friebel, Hauck, Aylin, & Steventon, 2018). Estimation of overall readmission rates in Australia is made difficult by a limited body of literature, but more narrowly focused studies have reported unplanned readmission rates of 32% for patients with atherothrombotic disease (Atkins, Geelhoed, Knuiman, & Briffa, 2014), up to 25% for older patients with acute hospital admissions (Scott, Shohag, & Ahmed, 2014), and 11.14% for older women (Shebeshi, Dolja-Gore, & Byles, 2020). Beyond the cost to patient welfare, the financial costs are enormous. The cost of unplanned readmissions in the USA was estimated to be US$17.4 billion in 2004 (Jencks, Williams, & Coleman, 2009).

The Hospital Readmissions Reductions Program (HRRP) introduced in 2012 in the USA is the most prominent example of healthcare policy targeting readmissions, under which hospital risk-adjusted readmission rates for certain conditions are tied to funding (Centers for Medicare and Medicaid Services, 2020). Healthcare policies targeting readmissions have similarly been implemented in Germany, Denmark[1], and England (Kristensen, Bech,

[1] Readmission rates are reported but not tied to financial incentives in Denmark.

& Quentin, 2015). Most recently, Australia's Independent Hospital Pricing Authority (2021a) has developed a pricing model adjusting funding for admission episodes based on readmission outcome, condition and complexity. Such policy aims to incentivise hospitals to improve quality of care, communication and management of high-risk patients to reduce readmissions. A core principle of such policy is that financial penalties for high readmission rates should reflect hospital performance rather than the risk level of the serviced population. This has necessitated the use of risk adjustment models relating patient-specific factors to risk of readmission or, in England, local clinical review.

The usage of financial penalties has been criticised in some cases (Kristensen et al., 2015). These critiques have related to how well the risk-adjustment models that are used in policy appropriately control for patient-specific risk (Zheng, Hanchate, & Shwartz, 2019), and whether readmission rates are a valid metric of quality (Fischer et al., 2014). There is, however, agreement that many readmissions are avoidable (van Walraven, Bennett, Jennings, Austin, & Forster, 2011). In Australia, avoidable hospital readmissions have been defined as those in which readmission occurs in a specific time frame, are related to the original admission, and could have been avoided through better clinical management and discharge planning (Australian Commission on Safety and Quality in Healthcare, 2019). A range of strategies can be employed by institutions to reduce avoidable readmissions to improve patient outcomes and lessen costs (Warchol, Monestime, Mayer, & Chien, 2019). Research has found robust interventions to be effective, though resource requirements make it important to identify high-risk patients for intervention targeting (Kripalani, Theobald, Anctil, & Vasilevskis, 2014). To assist institutions in the identification of high-risk patients for prioritisation of interventions and follow-up care, much research has focused on risk prediction models relating patient-specific factors to readmission risk. These models are distinguished from the risk adjustment models used more broadly in healthcare policy in that they are restricted to data available at the time of discharge.

The need for risk assessment tools for both performance measurement and to assist institutions in risk stratification and efficient resource allocation has led to the growth of an active and continuing research area. Better tools for understanding readmission risk would benefit stakeholders at the system level where risk adjustment underlies performance measurement, the institutional level where risk-stratification can improve resource allocation and direct preventative interventions, and the patient level where

better outcomes could be achieved. In particular, this work focuses on the readmissions of patients who were initially admitted to hospital after presenting to the Emergency Department.

1.2 Research Questions

Readmission research has proposed many models to attempt to quantify the risk of readmission given a patient's available information, though these models have tended to be characterised by unimpressive performance. At a high level, two approaches have been taken to problem formulation. The most common has been to consider the problem of predicting whether patients are readmitted by a single fixed time point, generally 30 days. This formulation has allowed for the application of classification models such as logistic regression and matches policy definitions. For example, the HRRP in the USA determines financial penalties using 30-day readmission rates (Centers for Medicare and Medicaid Services, 2020). This classification formulation has set the standard for how readmission risk models are evaluated. The less common approach is to consider readmissions as a survival problem, where the time until readmission is of interest and some patients are right-censored, meaning that their time until readmission is only known to be larger than their follow-up time. This approach is more complicated than treating readmissions as a classification problem, but survival models provide risk predictions over time rather than only at a fixed point. Despite providing risk over time, research using survival models have largely assessed these models based on their ability to predict risk at a fixed point as in classification approaches.

As part of the effort to develop improved readmission models, machine learning techniques have increasingly been employed. Machine learning has been instrumental for improving models in other areas of healthcare, and may better account for non-standard data and highly complex relationships than classical statistical techniques. While early results have been encouraging, further research is needed to establish the degree to which they may be able to improve on standard techniques. However, this increased machine learning consideration has overwhelmingly been as part of research taking a classification approach, with almost no research considering machine learning survival techniques. A systematic literature review of published research, described in Chapter 2, identified only one type of machine learning survival technique used in prior research and comparisons

with statistical techniques were not made (Hao et al., 2015; Padhukasahasram, Reddy, Li, & Lanfear, 2015).

The limited consideration of machine learning survival techniques and lack of assessment specific to survival model applications serve as the motivation for the two research questions.

RQ1: Can machine learning survival techniques improve upon statistical survival techniques when predicting 30-day hospital readmissions?

Machine learning techniques have been increasingly considered for classification approaches to readmission modelling. These have the potential to outperform statistical techniques on certain problems because of their capacity to capture highly non-linear and complex relationships. This applies to both survival and classification approaches. Despite this, consideration of machine learning techniques has been much rarer for survival approaches. This research aims to use a wide range of machine learning survival techniques to assess performance in relation to statistical techniques. Statistical techniques will include the gold standard in survival and classification approaches of Cox regression and logistic regression respectively. The inclusion of logistic regression in comparisons is driven by the fact that prior research involving predictive models, whether survival or classification, has assessed them exclusively in terms of fixed-point prediction. Given the classification basis for model evaluation, logistic regression is included as the most common classification technique.

RQ2: How well can various survival modelling techniques capture aspects of hospital readmission risk over time relevant to managerial decision-making?

Both survival and classification approaches have been considered in the literature, but the dominance of the classification approach has led to the framing of model applications from both approaches to be based on classification outputs. That is, if models are intended for practical use, it is in assessing the risk of readmission by some fixed time. This ignores the additional information provided by survival models where risk predictions are not restricted to a single time. This research aims to assess survival models based on novel applications not possible for classification models. These would complement existing model applications and provide additional tools for institutions to manage readmissions more effectively. Proposed applications include dynamic risk ranking, identifying duration of elevated readmission risk, and readmission forecasting. As measures

previously employed in readmission literature are insufficient for the novel applications proposed, this research also identifies appropriate measures of model performance.

In addressing these research questions, the cohort considered will be hospital admissions of adult patients presenting to the Emergency Department. This setting is characterised by a need for dynamic decision-making as well as diversity in the reasons for presentation. Such characteristics make tools improving risk assessment and management valuable.

Several secondary contributions from addressing the research questions will also be achieved. First, this research will add to the existing literature in Australia. Australia has been the setting for few readmission studies despite the recent development of healthcare policy aimed at reducing them. Secondly, findings will provide an empirical comparison of a wider range of survival analysis techniques from both statistical and machine learning fields than any previous work. Many distinct machine learning survival techniques have been proposed but there have been few comparisons between types of techniques, meaning there is little evidence to indicate clear consensus regarding which work best in which contexts.

1.3 Dissertation Structure

The remainder of this dissertation is structured as follows. In Chapter 2, a systematic review of hospital readmission modelling literature and discussion of research gaps is provided. This systematic review and the identified gaps motivate the second component of the literature review, which considers available statistical survival techniques as well as the extensions of major types of machine learning techniques to survival analysis. This includes consideration of decision trees, ensembles, support vector machines, and artificial neural networks.

After the literature review, Chapter 3 builds on the readmission research gaps and available survival techniques identified to develop the two research questions of this paper. These research questions are discussed with respect to these gaps and their contributions to both practice and literature. Guiding principles for decision-making within the project are also outlined.

Having set out the research questions, Chapter 4 describes the data used in addressing them, as well as relevant processing steps including data splitting and feature recoding.

Chapter 5 then details the project's methodology. This includes specifying which of the techniques identified in Chapter 2 are relevant to the research questions, describing the model selection process, and establishing the performance measures used in each research question.

Chapter 6 provides a more thorough description of each of the techniques found to be relevant in Chapter 5. This includes details of their implementation for the project and the way the hyperparameters were varied when selecting the final model of each technique for each research question.

Having described the data, basis for evaluation, and techniques, Chapter 7 presents the results of the project for both research questions. The results are discussed and contributions for each question are detailed.

Finally, Chapter 8 briefly highlights the conclusions of the project and their implications for practice and research. Suggestions for future research are then outlined and key limitations acknowledged.

2 Literature Review

2.1 Hospital Readmission Research

In this literature review, a systematic approach was taken to find studies of hospital readmissions and prediction. The goals of this review were to identify and describe the major categories of research as well as to highlight the trends and major points of similarity or difference between studies. While this work considers hospital admissions of adults from the Emergency Department, this review aimed to more broadly assess the literature and employed methodologies in the area of hospital readmissions more broadly. Accordingly, the search terms used did not restrict results to studies involving admissions from the Emergency Department or to adult patients.

The Scopus and Web of Science databases were both searched using the same search terms, returning all results as of the 9th of December 2019. No restrictions were placed on the publication date range for articles, the language they were written in, or other aspects of the results. Instead, decisions regarding exclusions were relegated to a manual review of abstracts and titles to ensure that key articles were not excluded by too restrictive conditions.

The search term for both databases was:

"hospital readmi pred*" OR "pred* hospital readmi*" OR "hospital readmi* model*" OR "model* hospital readmi*" OR "models for hospital readmission"*

This returned 116 citations from Scopus and 112 citations from Web of Science. After removing duplicates, these consisted of 141 unique citations. The title and abstract (where available) of each citation were then manually reviewed for relevance by the author and two supervisors. Papers were excluded if they were erroneously returned by the databases (for instance, entire conference proceedings or medical handbooks) or did not directly pertain to readmission risk in terms of risk factors, models, or reviews of relevant literature (for instance, studies into why patient appointments are not attended or other facets of patient outcomes). Studies were excluded only if all three reviewers agreed on their exclusion. This process excluded 13 studies from consideration. A further six were excluded as they could not be obtained in full or in abstract form. In general, the author opted to include almost all studies to ensure comprehensive coverage of the research area

given that the number of studies under consideration was not prohibitively large. The final 122 studies included consisted of mostly peer-reviewed English journal articles, with a small number of studies that were abstracts, poster presentations, or in another language (3). Two citations returned by the search were conference papers which had since been published as journal articles that were also returned by the search. The journal article versions were reviewed instead of the earlier conference papers, further reducing the set of studies to 120. Finally, nine articles were identified through reference lists that were felt to be pertinent to the topic and were added to the review, bringing the final number of studies to 129.

To summarise this process:

- 116 citations from Scopus and 112 citations were obtained from Web of Science, resulting in 141 unique citations.
- A manual review excluded 13 studies based on whether they were actual studies and relevant, resulting in 128 citations.
- Inability to obtain articles or their abstracts excluded another 6 studies, resulting in 122 citations.
- Four citations represented two conference papers and their later journal article versions, and so the two conference papers were not included, resulting in 120 citations.
- Nine papers were identified from other sources, resulting in 129 citations for the final review.

2.1.1 Research Categorisation

Several overlapping categories of research were identified from this review and are briefly described here. The first category consists of those studies which focus on the various risk factors for hospital readmission (68). The second revolves around the development and validation of those predictive models which have gained some widespread acceptance (19). Third is the research focusing on the continuing development of new models predicting readmission risk (41). Related to this is the fourth category, investigating the potential of machine learning methods to add value through more complex modelling leveraging larger datasets or integration of non-traditional data (29). The categorisation of each citation was done subjectively by the author. 11 papers

did not fall into at least one category, with these involving reviews (Ardura-Garcia, Stolbrink, Zaidi, Cooper, & Blakey, 2018; Artetxe et al., 2018; Kansagara et al., 2011; Mehta et al., 2017; Weinreich et al., 2016; Yoo et al., 2015), a guide for clinicians (Jelinek & Yunyongying, 2016), descriptions of larger datasets (Shafer, 2019), pre-processing techniques (Duggal, Shukla, Chandra, Shukla, & Khatri, 2016a), and a statistical paper using readmissions as a case study (Neumann, Holstein, Chatellier, & Lepage, 2004).

2.1.1.1 Category 1 – Readmission Risk Factors

This category of research is most general and accordingly most prolific. These studies have typically revolved around identifying, understanding, and leveraging the different risk factors for hospital readmission of various patient groups. With risk factors identified, they demonstrate what variables should be considered by more extensive measures aiming to quantify each patient's readmission risk as well as highlighting what information is important to collect. Further, a better understanding of how these factors influence readmission risk enables proactive measures addressing the risk sources. For example, Shyu, Chen, and Lee (2002) analysed how the needs of caregivers to discharged elderly patients in Taiwan influenced the likelihood of readmission. They found that the ability of elders to take care of themselves as well as the needs of the caregivers both influenced readmission probability. By looking at these factors individually, recommendations were made to as to how the risk factors could be better managed, specifically through provision of support groups and matching support services to caregiver needs. From a preventative perspective, such practice-focused recommendations highlight the distinction between explanatory research considering risk factors rather than predictive research developing computational models, with the value of the latter typically being in patient prioritisation. Many risk factors identified are commonly found to be important across study cohorts, with examples including previous admissions, hospital acquired complications, length of stay, whether the admission was considered acute, and comorbidities. Other risk factors were novel or specific to certain cohorts. These included exposure to certain medication classes for older adults (Pavon, Zhao, McConnell, & Hastings, 2014) and frailty for patients admitted to hospital with chronic obstructive pulmonary disease disorder (Bernabeu-Mora et al., 2017).

There has also been increased interest in non-standard information sources, such as textual data from electronic health records (Xiao, Ma, Dieng, Blei, & Wang, 2018) and hierarchical disease classifications (Jovanovic, Radovanovic, Vukicevic, Van Poucke, &

Delibasic, 2016). These have historically been difficult to leverage in readmission prediction because of the challenges they pose to traditional modelling techniques. The ability of machine learning techniques to better account for such non-standard information is linked to their increased usage, discussed further below.

2.1.1.2 Category 2 – Existing Predictive Models

Predictive models for hospital readmissions attempt to relate a range of risk factors to the risk of readmission, either through a score to stratify risk (van Walraven et al., 2010), or by directly predicting the probability of readmission. Several such models have gained some degree of acceptance and are used in practice. They are used to identify high-risk patients for prioritising preventative interventions and to measure readmission rate performance after adjusting for the risk of presenting patients, particularly given the increasing use of readmission rates as a quality-of-care metric. Prominent examples include the LACE index (van Walraven et al., 2010) and the HOSPITAL score (Donzé, Aujesky, Williams, & Schnipper, 2013), both of which provide risk stratification based on a straightforward equation using a small number of variables commonly recorded for patient admissions. The small number of key variables required facilitates use across institutions which may have differences in data collection practices. The simple equations used allow for quick calculation of the index as well as increased transparency regarding the contributing risk factors. These aspects are highlighted as they reflect the intention for these models to be used across institutions, potentially at the cost of performance. This intended application and consequent small number of variables in a simple construct has influenced the techniques considered when developing models to be used across institutions, with logistic regression being most prominent. Literature considering these types of models typically pertains to their development or validation, with the latter being particularly important for establishing how much value they add and whether they can be applied in new contexts. For example, Cotter, Bhalla, Wallis, and Biram (2012) investigated the performance of the LACE index in an older UK population, finding that it did not generalise well to this population. This validation process may also aim to determine which of several applicable tools exhibits the best performance. An example of this is by Deschodt et al. (2012), who compared the Identification of Seniors at Risk (ISAR), Triage Risk Screening Tool (TRST), and the Variable Indicative of Placement (VIP) tools for predicting readmission for elderly patients. Both value and generalisability

to new contexts of these existing models are linked to the motivation for the third category of research focusing on the continuing development of new models.

2.1.1.3 Category 3 – Model Development

While certain models have achieved some acceptance as standard measures, the development of new models has been an active research area. This has been driven by the need for better performing and context-specific (e.g., geographic, population, or disease group) ones. An example of this was provided by Low et al. (2017), who validated the LACE index in an older population of patients in Singapore and found that it had poor ability to discriminate between high- and low-risk patients. This is an example of a model exhibiting poor performance when applied in a new geographic region and specific population group, indicating that better tools for predicting risk are needed. Poor differentiation of patients who were and were not readmitted based on the LACE index has also been noted by H. Wang et al. (2014) while Duncan and Huynh (2018) demonstrated that predictions can be improved by including additional risk factors to the LACE index. Improving on the existing models comes partly from better customising to the regions they will be applied (e.g., re-deriving the LACE index for the UK setting) as well as, depending on intended usage, relaxing the need for them to be simple and easily constructed. These avenues for improvement are often pursued in tandem, where context-specific models are being developed with more complex structures and using a wider range of information. More complex modelling allows for better capturing of potentially highly complex relationships between risk factors and readmission outcomes that might not be adequately accounted for by simpler approaches. In particular, there has been an increasing interest in machine learning over statistical techniques, with Morgan et al. (2019) finding favourable performance of a machine learning-based system when compared to the modified LACE, HOSPITAL, Maxim/RightCare scores. Using more information than only some of the most commonly recorded features of hospital stays means that a greater proportion of relevant risk factors can be accounted for. More specifically, this means that more variables can be considered, including those which may be context specific or otherwise not universally recorded by all institutions. This additional information has in some cases been relatively simple to incorporate when generally recorded and in a standard format, but some research has also focused on leveraging new sources of information or those in non-standard formats.

2.1.1.4 Category 4 – Machine Learning Models

The consideration of machine learning makes up a sub-category in the model development literature. Machine learning models have increasingly been successfully applied to healthcare problems in recent years, and this has helped them to gain popularity for predicting hospital readmissions. Their motivation for use over statistical techniques lies in their lack of parametric assumptions regarding the nature of the underlying data, and their greater capacity to capture highly non-linear and complex relationships without explicit specification. If the parametric assumptions of a statistical model are violated or it is believed that the relationships in the data are complex, machine learning models may offer improved performance. In most cases, this improved performance requires larger datasets for training and comes at the cost of reduced interpretability. That is, it can be challenging to attribute a patient's risk level to a specific aspect of their situation, with the interrelation of risk factors in the model being difficult to decompose. Thus, such models are typically intended for use in contexts where risk prediction is the primary goal rather than understanding meaningful relationships between risk factors and readmission risk. The limited interpretability of such techniques has become less relevant, however, through the increased availability of methods for extracting rules and relationships from machine learning models (Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016).

A variety of machine learning techniques have been applied in readmission modelling, including decision trees, support vector machines and neural networks (Alajmani & Elazhary, 2019; Turgeman & May, 2016; J. Zhang, Yoon, Khasawneh, Srihari, & Poranki, 2013). Though differences between study contexts make comparisons across studies difficult, an issue discussed in Section 2.1.2.4, these more complex techniques have often found success with improved performance compared to more traditional techniques (Futoma, Morris, & Lucas, 2015; Kalagara, Eltorai, Durand, Mason DePasse, & Daniels, 2019; Reddy & Delen, 2018). Beyond just representing an alternative approach to modelling the relationship between predictors and readmission outcomes, some machine learning methods also have great flexibility in the types of information used as inputs. This has complemented those studies looking to leverage new sources of information for predicting readmissions.

2.1.1.5 Other Research

While the categories of research mentioned (readmission risk factors, existing predictive models, model development, and machine learning) include nearly all returned studies in

the area of readmission modelling, some did not fall neatly into any of them. These studies included reviews of literature (Artetxe et al., 2018; Kansagara et al., 2011; Mehta et al., 2017; Weinreich et al., 2016; Yoo et al., 2015), those exploring and describing patterns in large datasets (Barnes et al., 2017; Garg, Sarvepalli, Goyal, Kandlakunta, & Sanaka, 2018; Shafer, 2019), one describing the effect of data pre-processing techniques on readmission prediction (Duggal et al., 2016a), and one describing a statistical regularisation method with readmissions as an example (Neumann et al., 2004). As these studies do not fall into a natural categorisation and do not represent major research trends, they are not discussed in any detail in their own subsection. They are, however, mentioned where relevant in the following discussion sections regarding the characteristics and themes of the research area.

2.1.2 Research Themes

The categorisation used for distinguishing the major classes of research in this area is useful for understanding the varied goals of researchers as well as to identify where research is continuing, particularly with respect to more complex modelling techniques and incorporating new information. With this important context established, the characteristics and themes of research can be explored in greater depth. In particular, several aspects of research in the area are common across these categorisations, though they may manifest slightly differently. The four themes of research identified through the review are problem formulation, techniques employed, performance measurement, and study heterogeneity. These themes are discussed in the following four subsections, with reference to specific categories where appropriate. It should be noted that while explanatory research identifying risk factors is a large contributor to this area, the emphasis in the following discussions will be on predictive research. General discussion relating to overall trends and gaps in the literature are deferred to Section 2.1.3.

2.1.2.1 Theme A – Problem Formulation

Readmission studies have formulated the problem in two ways for practical purposes, with either formulation having implications for the applicable modelling techniques.

The first and most popular approach is to treat readmissions as a classification problem with a limited number of possible outcomes by considering patient status at a fixed time point. For example, a study might dichotomise patients into those who were and were not readmitted within 30 days. This makes the problem tractable for the most common techniques in healthcare research which require binary outcomes. Further, it may be

reasonable to ignore readmissions occurring after some length of time if they are believed to be unrelated to the original admission, also known as the index admission. The example given above reflects the most popular approach to formulating readmission both in the use of the 30-day cut-off and the dichotomous outcomes. Studies are not wholly homogeneous in these respects, however, as will be detailed further in Section 2.1.2.4.

The popularity of the classification formulation, particularly for predictive research, is likely due to two key factors. First, classification is a common and relatively straightforward task. This means it is familiar to many medical practitioners and it also has a wide range of applicable statistical and machine learning techniques. Secondly, considering readmission using a fixed time point matches real-world policy measures. For example, the HRRP in the USA focuses on risk adjusted readmission rates for six conditions or procedures. and reduces payments to hospitals found to have excess readmissions (Centers for Medicare and Medicaid Services, 2020).

While well-established and matching policy initiatives, the classification formulation for readmission does have several limitations. First, there is a lack of flexibility in the predictions made by a classification model based on a fixed time point. These predictions are useful for measurement of readmission rates and can be used for identifying the highest risk patients as at discharge, but they cannot be adjusted to estimate risk for a different time point without re-estimating the entire model. Beyond this, they do not allow for estimating the risk faced by a discharged patient once gone for any length of time. For example, if a patient was discharged three days prior, the probability of readmission in the remaining 27 days is not naturally calculable. Secondly, the use of a fixed cut-off time introduces a boundary problem (Yu et al., 2015) on account of the way timing of readmission is largely ignored. The boundary problem relates to potential exaggeration of differences between similar cases and understatement of differences between dissimilar cases. This is best illustrated through an example: a patient readmitted in 30 days has a very similar outcome to a patient readmitted in 31 days, but the two patients would have opposite class memberships. Further, a patient readmitted in one day may be very different from a patient readmitted in 30 days, but the two patients would have the same class membership. Thirdly, related to the boundary problem, the classification formulation takes a one-size-fits-all approach with its use of a single fixed cut-off time. It is assumed, at least implicitly, that patients face an elevated risk of readmission shortly after their discharge because of the potential for complications stemming from the index

admission. After this time, readmissions may occur but are not necessarily due to factors immediately related to the index admission. The limitation of the classification formulation here is that the same period is used for all patients, ignoring the potential variation in the length of time for which different patients face elevated risk related to their index admission. As it is the risk related to index admission that is of interest, particularly for maximising patient welfare through preventative measures, this risk may be under- or over-estimated for patients whose elevated risk periods are not approximately equal to the cut-off period. To illustrate this, Figure 1 shows the daily risk of readmission for a real patient for whom the elevated risk period may not correspond to a generic fixed time-period.

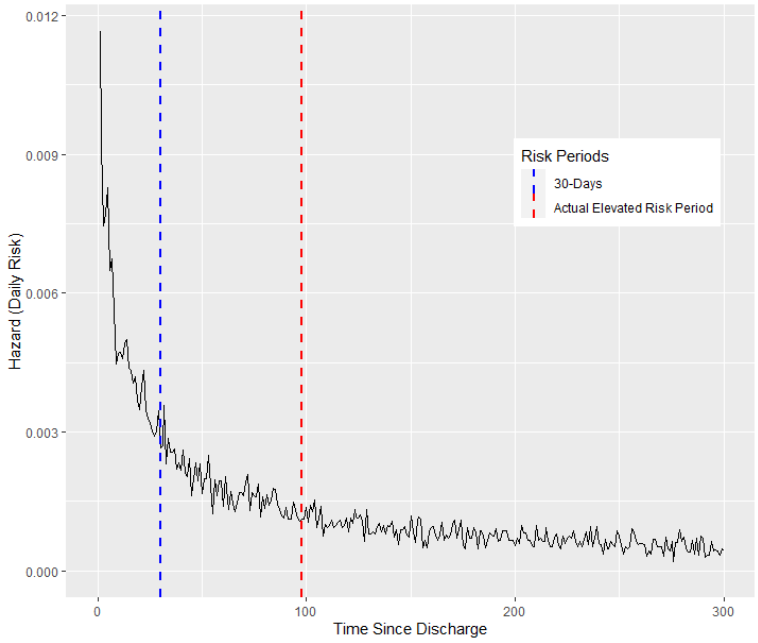

Figure 1. Evolution of Readmission Risk Over Time

The second and rarer approach is to treat readmissions as a time-to-event or survival problem, such as by Grzyb et al. (2017). The field of survival analysis is concerned with modelling data in which the time until an event occurs is of interest, but the event is not always observed before the end of the study. This approach estimates the risk of event occurrence as a function of time rather than being limited to a single point. The differences in research under a survival approach as compared to a classification approach are relatively minor for explanatory studies investigating potential risk factors. For predictive modelling studies, however, survival approaches are very rare as they do not conform to the most common practical application of said models. They may, however,

be more useful for patient management through the additional information they provide. The differences in techniques and measures of performance depending on the problem formulation is discussed in Section 2.1.2.2 and Section 2.1.2.3 respectively.

The lack of popularity of survival formulations is largely driven by the same factors that make classification formulations popular. First, survival techniques, at least for predictive modelling, are harder to interpret and there are fewer well-established machine learning survival techniques. Secondly, they do not correspond to the current real-world policy-measures being employed which assess readmission rates by a fixed time point. As survival techniques aim to estimate risk across all time points, they can be expected to have inferior performance for this measure compared to classification models which focus on estimating risk only at this time point (Yu et al., 2015). This makes them less appropriate for application in contexts requiring only a single and standardised estimate of risk, such as in performance measurement under broadly applicable policy.

Survival formulations can, however, address the limitations of classification formulations by estimating risk across time. They allow for estimating risk at any time point and can produce risk estimates which account for patients who have been readmission-free for some period. By considering risk across time and not requiring specification of a fixed cut-off, they avoid the boundary problem as well. Finally, by modelling the hazard function many survival models allow for inspection of risk across time, potentially allowing for identification of the length of time a patient faces elevated risk post-discharge. These comparative advantages may make survival models more applicable to within-hospital use than classification techniques.

Overall, survival approaches are sometimes seen in explanatory research, but classification formulations are more common. Of the 103 studies included in this review which involved predictive models in some form, only eight considered survival approaches either through choice of models considered or methods of assessing risk factor significance. The popularity of classification approaches was also noted in a previous review (Artetxe et al., 2018).

2.1.2.2 Theme B – Techniques Employed

In describing the techniques employed in research around readmission modelling a distinction is made between the explanatory research investigating risk factors compared to predictive research developing computational models. Accordingly, techniques used

for each are discussed separately. As classification formulations of the readmission problem are most common, tools employed for explanatory and predictive studies are described on this basis. The tools and methods under survival formulations are then described as deviations from this standard.

Explanatory research

Explanatory research aims to establish whether potential risk factors are significantly associated with readmissions and the nature of the relationship. In general, this research has adopted three broad methodologies, based on establishing univariate significance, multivariate significance, or model improvement.

When establishing univariate significance of potential risk factors, admissions are broken into two groups according to whether a subsequent readmission was observed or not by the specified time frame. An appropriate univariate test statistic is then used to determine whether the difference in values for a potential risk factor between the two groups is statistically significant. For example, Millien, Townsend, Goldberg, and Fuhrman (2017) considered a range of potential readmission risk factors for patients who had undergone surgery for perforated appendicitis. For continuous risk factors, t-tests were used to establish the significance of differences between readmitted and non-readmitted groups, whereas for categorical risk factors chi-squared tests were used. Other valid tests for differences of categorical risk factors include the Wilcoxon signed rank test and Mann-Whitney U-test, as used by Wong et al. (2008). For survival analysis formulations, univariate significance can be established for a binary risk factor by considering two groups of patients based on the risk factor (dos Santos et al., 2015; Dupuis-Lozeron, Soccal, Janssens, Similowski, & Adler, 2018; Koulouridis, Price, Madias, & Jaber, 2015). Kaplan-Meier survival curves are then constructed for each group and a log-rank test can be used to determine whether the survival curves are significantly different. Alternatively, the significance of the risk factor in a univariate Cox regression model can be assessed, allowing for testing of non-binary risk factors (Cheng & Silverberg, 2019).

In establishing univariate significance, variables are considered individually without allowing for the effects of other risk factors. Studies investigating multivariate significance aim to establish whether potential risk factors significantly influence readmission risk after controlling for other variables. To do this, predictive models are used. More specifically, logistic regression and, for the less common survival

formulation, Cox regression models are constructed. Unlike machine learning methods, these statistical models allow for straightforward assessment of the statistical significance of included predictors. For a classification formulation, a logistic regression model is constructed using a collection of relevant control predictors and the potential risk factor under consideration. For a potential risk factor to be significant in this multivariate context, it must add information not already captured by the other predictors. A survival formulation uses a similar process, but with a Cox regression rather than logistic regression model (Cheng & Silverberg, 2019).

Lastly, a few studies eschew directly assessing variable significance and instead shift the focus to assessing whether the potential risk factors result in improved model performance. In these cases, model performance with and without the potential risk factors under investigation are assessed. Improvements for the model including the potential risk factors directly link to the potential value of these factors in future development of predictive models. This has two implications for methodology which are distinct from investigating univariate or multivariate significance. First, it is now the improvement in common model performance measures which is of interest. These measures include the area under the receiver operating characteristic curve (AUC) and accuracy. Performance measures for predictive models are discussed in Section 2.1.2.3 and thus not described further here. Secondly, as statistical significance of the variables within the model is no longer the focus, machine learning methods can be applied as they might be in predictive modelling research. Two such examples were identified using machine learning methods in conjunction with risk factor investigations. The first is by Chandra et al. (2019) who aimed to both develop a predictive model as well as determine the risk factors for readmissions of patients discharged to skilled nursing facilities. The authors found the most appropriate approach to determining variables to include in a model (i.e., the relevant risk factors) was to evaluate model performance when including different combinations and groups of variables. They did not, however, assess the significance of differences in performance measures under different scenarios. The second example is by Kalagara et al. (2019), who compared models using all available patient factors compared to those only known prior to hospital discharge using a machine learning model and a logistic regression model. Through this approach, they demonstrated that factors only available post-discharge did influence readmission risk

and their inclusion improved model performance. No studies were identified that assessed how well a potential risk factor improved a survival model.

While the above paragraphs have discussed the various approaches taken to determining whether potential factors do inform the risk of readmission, this is not to imply that they are distinct from one another. In many studies, a combination of these methodologies is applied. Univariate significance and multivariate significance may both be assessed (Fathi et al., 2017; Weinland, Braun, Mühle, Kornhuber, & Lenz, 2017) or univariate significance may be used to determine which variables require controlling for when establishing multivariate significance (Almussallam et al., 2016; Mosquera, Vohra, Fitzgerald, & Zervos, 2016). Even in studies primarily interested in multivariate significance, model performance may also be assessed to provide an indication of the actual value of the predictors being considered beyond the binary result as to whether they are statistically significant predictors of readmission risk (Tabata et al., 2014). Further, while classification and survival formulations are mentioned separately, in some instances aspects of both are used, such as by Cheng and Silverberg (2019) who looked at readmission within specified time frames (a classification formulation) as well as using survival tools.

Overall, explanatory research tends to be characterised by the univariate and multivariate significance methodologies in classification problem formulations. Using model performance as the primary determinant of risk factor relevance is much rarer, as are survival formulations under any methodology. Finally, while there is some variation in how potential risk factors are assessed, this was not associated with chronological trends.

Predictive Research

The focus now shifts to the techniques employed by predictive research. Historically, this research has used statistical techniques for both classification and survival formulations. By far the most common approach has been to use logistic regression, a generalised linear model for binary problems. Of the 51 studies involving the development of models for benchmarking (such as the LACE index), institution-specific model development, or machine learning models, 30 used logistic regression in some capacity. Cox regression was used only four times. Ignoring the difference in problem formulation, both are statistical techniques making certain assumptions regarding the nature of the data being modelled and both require specification of the relationship between outcomes and

predictors. In recent years, the emphasis on statistical models has lessened, at least in classification formulations, with greater application of machine learning models. While logistic regression has remained the most used model and as the benchmark for readmission prediction, there has been increasing interest in the application of machine learning models, driven by their encouraging performance in other areas of healthcare. Such models offer advantages on data where relationships between outcomes and predictors are highly complex and non-linear, but they also require a greater amount of data to train. The data requirements of these models for larger datasets have become less restrictive in the modern environment which is increasingly characterised by big data, with greater volume and variety of data available. Machine learning models also allow for the incorporation of new information not previously used in statistical models, such as textual information (Xiao et al., 2018). This can improve performance, though textual and other non-standard data forms can also pose new challenges in terms of data quality, manual processing, and the need for input from domain experts. Of the 51 studies mentioned, the main types of machine learning techniques used, in order of prevalence, are: artificial neural networks (ANNs) (Alajmani & Elazhary, 2019; Graña, Lopez-Guede, Irazusta, Labayen, & Besga, 2019; Jiang, Chin, Qu, & Tsui, 2018; Wolff, Grana, Ríos, & Yarza, 2019; J. Zhang, Lam, & Poranki, 2013), support vector machines (SVMs) (Baechle, Agarwal, Behara, Zhu, & Ieee, 2017; Turgeman & May, 2016; J. Zhang, Yoon, et al., 2013), random forests (RFs) (Deschepper, Eeckloo, Vogelaers, & Waegeman, 2019; Futoma et al., 2015; Hammoudeh, Al-Naymat, Ghannam, & Obied, 2018), and decision trees (Pham et al., 2019; H. Wang et al., 2018). Other model types noted but less frequently seen included Naïve Bayes (Almardini & Raś, 2017), gradient boosting machines (Chandra et al., 2019), K-Nearest Neighbours (Graña et al., 2019), and ensembles of models (Pham et al., 2019). The models have typically been compared to logistic regression (e.g., Ottenbacher et al. (2001)), with encouraging initial results. Comparisons between machine learning techniques have also been made. For example Chopra, Sinha, Jaroli, Shukla, and Maheshwari (2017) looked at the performance of a recurrent neural network against an SVM, RF, and feed-forward ANN. Similarly, Futoma et al. (2015) compared a range of machine learning techniques and a penalised logistic regression, finding that a deep ANN provided superior discrimination across key patient groups associated with readmission penalties. Such findings and comparisons are difficult to generalise across studies, however, because of the use of different datasets, patient groups, and conditions. This is discussed further in Section 2.1.2.4. In general, however,

it was observed that relatively simple logistic regression models were outperformed by more complex ones. This may indicate that the dynamics of hospital readmission are complex enough to warrant such models or that, in at least some cases, the benchmark model's implementation was too simplistic.

One consequence of the increased use of machine learning methods over statistical is the imbalanced nature of the readmission problem. This imbalance refers to the fact that, in the context of binary classification, one outcome is much more frequent than the other. In this case, non-readmissions tend to be substantially more frequent than readmissions. This has influenced the metrics used to assess performance (discussed further in Section 2.1.2.3), as high-level metrics like accuracy mask class-specific performance. Various approaches try to address this problem, including oversampling the minority class (Duggal et al., 2016a; Vukićević, Radovanović, Kovačević, Štiglic, & Obradovic, 2015; J. Zhang, Lam, et al., 2013; J. Zhang, Yoon, et al., 2013), undersampling the majority class (Almardini & Raś, 2017; Zhao & Yoo, 2017; K. Zhu et al., 2015), propensity score matching (Koulouridis et al., 2015) and Synthetic Minority Over-sampling Technique (SMOTE) (Hammoudeh et al., 2018; Jiang et al., 2018; Kalagara et al., 2019; Reddy & Delen, 2018; Sundararaman, Valady Ramanathan, & Thati, 2018; Wolff et al., 2019). Such approaches have been infrequently used in this domain, but with increased recent interest. All instances of SMOTE being used identified through this review have been since 2018, and other methods were primarily identified in studies from the past 6 years. Of these, several studies reported improved performance after the application of these methods (Duggal et al., 2016a; J. Zhang, Lam, et al., 2013).

While the trend towards machine learning research is clear for classification formulations, it is less apparent for survival formulations. Those studies using a survival formulation with the Cox regression model have largely focused on prognostic insights rather than actual model application. Though the motivation for machine learning techniques under classification approaches is also relevant for survival approaches, such research has not seen the same attention. Only two studies were identified in this review where machine learning techniques for survival analysis were used, and in both cases these were random survival forests (Hao et al., 2015; Padhukasahasram et al., 2015). Padhukasahasram et al. (2015) demonstrated that using both clinical and behavioural variables resulted in models with significantly higher discrimination than models using either category of variables alone but did not make comparisons between random survival forests and variations of

the Cox model. Hao et al. (2015) used a random survival forest to develop a 30-day readmission risk assessment tool and did not consider alternative model types. While survival curves were presented, the 30-day readmission outcome was the primary focus and basis for model assessment. The assessment of survival models based on binary outcomes was also noted for statistical techniques. For example, Yu et al. (2015) evaluated Cox regression models but only in the context of 30-day readmission prediction.

One final comment to make regarding the techniques relates to the variable selection methods employed. The variable selection aspect of methodologies is important for predictive as well as explanatory research with multivariate risk factors. Variable selection is the process of deciding what variables should be included in a model. Including unimportant variables can be expected to increase the degree to which a model fits random error (known as overfitting) while excluding important variables can prevent a model from capturing real relationships (known as underfitting). In the context of explanatory research with a multivariate approach, excluding an important variable may also result in the incorrect conclusion that the variables under consideration add information not already captured. Variable selection methods have differed greatly between studies. In explanatory research, it is often based on domain knowledge, consideration of literature, univariate significance tests, or multivariate significance. For predictive research, stepwise procedures, multivariate significance and regularisation tools (such as LASSO by Rana et al. (2014)) are more common and some machine learning techniques (such as decision trees and random forests) automatically perform variable selection. Additionally, in some cases it is not reported for both explanatory and predictive research.

Overall, statistical techniques for readmission prediction have been the traditional approach in the literature, both for classification and survival formulations. More recently, machine learning techniques have been investigated for classification formulations with encouraging results, but this trend is not as apparent for survival formulations. There has also been increasing consideration of methods like SMOTE that aim to address the issue of class imbalance.

2.1.2.3 Theme C – Performance Measurement

Having described the two formulations of the readmission problem as well as the techniques employed for both explanatory and predictive research, the measures used to

assess performance are now discussed. For explanatory research, the techniques and performance measures are intertwined – the univariate or multivariate significance is the measure of interest captured by appropriate statistical tests. Accordingly, performance measures in the context of predictive research are the focus here. The following discussion relates to explanatory research only in that some studies have assessed the value of potential risk factors through assessing model performance with and without their inclusion.

The measures of performance used in the literature relate to classification formulations. Only two instances were identified by this review where model evaluation was based on a survival formulation. Padhukasahasram et al. (2015) evaluated survival models using a variation of the area under the Receiver Operating Characteristic (ROC) curve for censored data but did not explore the application of these models outside of a prioritisation tool for use only at the time of discharge. Grzyb et al. (2017) similarly used this variation, known as Harrel's concordance index (Harrell, Lee, & Mark, 1996), which was developed in the context of Cox regression and relies on the assumption of time-invariant risk rankings. Where survival models were used in other studies, performance assessment was still based only on their ability to predict readmission by a single fixed time point.

A variety of performance measures have been used for assessing classification models. These have included accuracy, sensitivity, specificity, precision, recall, the F-score, the Brier score, and calibration curves. While the exact measures employed vary across studies, two elements of model performance have been generally accepted as important and are reported in most cases, namely discrimination and calibration.

Model performance has been primarily measured in terms of discrimination and, to a lesser degree, calibration. Discrimination refers to the ability of a model to differentiate between positive and negative instances in its outputs. That is, it assesses the circumstances where the model assigns higher scores or probabilities to positive instances than to negative instances. The vast majority of studies involving model development or validation use the ROC curve to report discrimination, which represents the combinations of specificity and sensitivity a model can achieve by varying the decision threshold at which an observation is classified as a positive or negative instance. Studies have less often simply reported the specificity and sensitivity of a model at a given decision threshold (Pham et al., 2019), but these measures are less informative. The ROC curve is

most often summarised into a single number by calculating the area under the ROC curve, referred to as the AUC or c-statistic. The AUC can be interpreted as the probability that a model assigns a higher score to a randomly selected positive instance than to a randomly selected negative instance. An AUC of 0.5 would indicate a model cannot differentiate between classes while an AUC of 1 would indicate a model was perfectly able to so.

Model calibration is less often assessed, but it is an important measure for evaluating how well a given model fits the relevant data and is the degree to which predicted probabilities of readmission match with the actual chance of readmission. This is typically assessed by considering groups of patients from low to high predicted levels of risk and comparing observed and expected readmission rates in each group, as in Benuzillo et al. (2018) and Rubin et al. (2016). This measure is particularly important for models which are intended for use in comparing institution performance, such as those used in policy implementations. Such models are trained on multiple institutions, and then are expected to output the readmission risk for patients based on the 'average' institution. The models provide an equitable basis for assessing institutional performance only if there is close agreement between predicted and actual risk. Well calibrated probabilistic models are also a requirement for institution-specific models used for patient decision-making based on absolute rather than relative risk. Decisions based on the relative risk of a patient require only that models correctly order the risk level of patients, whereas decisions based on absolute risk of a patient require that predicted probabilities are accurate for considering patients in isolation.

Most models developed tended to achieve AUC values between 0.6 and 0.75. Performance cannot be reliably compared between studies, however, for a variety of reasons discussed further in Section 2.1.2.4. Despite this, it can be reaffirmed that the performance of models developed thus far is often not satisfactory and performance is highly variable depending on study – an observation consistent with previous reviews (Artetxe et al., 2018; Kansagara et al., 2011). Even models used as benchmarks are often reported to have poor performance. For example, the LACE index (van Walraven et al., 2010) developed in 2010 has been treated as a gold standard in several studies (Yu et al., 2015) but has been shown to perform poorly in some populations (Cotter et al., 2012; Low et al., 2017) and no better than clinicians in others (Miller, Nguyen, Vangala, & Dowling, 2018). It is also worth noting that higher performance was observed for more narrowly defined populations, which is discussed further in the following subsection.

2.1.2.4 Theme D – Study Heterogeneity

A major driver of readmission prediction research has been study heterogeneity, as it has precluded most models from widespread application and led to the variable model performance mentioned previously. Several major attributes for study differences were identified, namely study region, population, data, and timeframe. These are each briefly discussed below.

Region

Holding constant other aspects of study design, readmission dynamics may differ between regions. As such, findings in one region may not hold in others. These regional differences relate to patient-level factors, such as underlying prevalence and susceptibility to conditions, as well as system-level factors, such as quality of care, accessibility of health services, data availability, and climate. By far the most common region considered in the reviewed literature was the USA, with this being the setting for over 50% of the identified predictive and explanatory studies. There is a smaller body of literature considering other regions such as China (Jiang et al., 2018; Wong et al., 2010; Wong et al., 2008; Yang et al., 2017), Australia (Parker & Hadzi-Pavlovic, 1995; Rana et al., 2014), and the UK (Billings, Dixon, Mijanovich, & Wennberg, 2006).

Population

Differences in the patient population under consideration, even in the same region, also pose obstacles to the generalisability of findings. Much of the research in this area considers populations defined in terms of demographics, conditions, or index admission characteristics. Demographic definitions of patient groups typically relate to age, which can act as proxy for factors such as frailty, need for assistance, and decision-making capacity. Many studies focus on elderly patients, who make up a growing proportion of the population and tend to be characterised by higher rates of readmission. These studies typically restrict their populations to those over the age of 65 (Krompass, Esteban, Tresp, Sedlmayr, & Ganslandt, 2015; Low et al., 2017; Morrissey, McElnay, Scott, & McConnell, 2003; Yoo et al., 2015) though with other cut-offs including 60 (Pavon et al., 2014), 70 (Graña et al., 2019), and 75 (Deschodt et al., 2012). Similarly, several studies have limited the focus to paediatric patients (Ardura-Garcia et al., 2018; Radovanović, Delibašić, Jovanović, Vukićević, & Suknović, 2019; Radovanović, Vukićević, Kovačević, Štiglic, & Obradovic, 2015; Vukićević et al., 2015), where important

dynamics may differ from adult patients. For example, the personal agency of adult patients is higher than in paediatric patients where a guardian may exercise greater control.

Specification of patient groups based on certain conditions reflects two factors. Firstly, considering a single condition type leads to a more homogeneous sample of patients which is easier to model. Studies with more narrowly defined patient groups were also found to be characterised by improved model performance in this review. This type of research typically aims to improve predictions or understanding of risk factors for only this condition. For example, several studies have focused on diabetic patients (Duggal, Shukla, Chandra, Shukla, & Khatri, 2016b; Zhao & Yoo, 2017). Secondly, the introduction of healthcare policy focusing on readmission rates has incentivised the targeting of certain conditions. For example, the HRRP in the USA targets six conditions with financial penalties for excess readmissions (Centers for Medicare and Medicaid Services, 2020). These conditions thus receive research interest both because of their policy focus as well as their high contributions to readmissions. Studies aiming to model readmission risk have often focused on these conditions, such as heart failure (Cheung & Dahl, 2018; Lagoe, Noetscher, & Murphy, 2001; Perez et al., 2017) and chronic obstructive pulmonary disease (COPD) (Bernabeu-Mora et al., 2017; Steer, Gibson, & Bourke, 2012; Troyano et al., 2018). Other studies have also opted to consider a range of conditions rather than limiting focus to only one (Cholleti et al., 2012; Fathi et al., 2017; Futoma et al., 2015).

In addition to demographic and condition groupings, cohorts have also been defined in terms of certain characteristics of the index admission. This may relate to the avenue through which the patient was first admitted, such as an emergency inpatient admission (Howell, Coory, Martin, & Duckett, 2009), or events associated with that first admission, such as a form of surgery (Pack et al., 2016; Raines, Ponce, Reed, Richman, & Hawn, 2015; Sabourin & Funk, 1999). Regarding patients admitted to hospital via the emergency department, lower readmission rates for elderly populations have been reported relative to other avenues of admission (Artetxe et al., 2018; Cui et al., 2015). Such findings supplement the practical knowledge of differing patient dynamics depending on whether the initial admission was direct to the hospital and potentially planned. Finally, a study cohort may be defined through a combination of demographic, condition, and index admission characteristic factors.

While so many delineating factors could result in many distinct study cohorts, it should be noted that there some cohort definitions are relatively common across studies. These include elderly or paediatric patients, those with COPD, heart failure, diabetes, or undergoing coronary artery bypass graft (CABG) surgery. Many studies have also opted to consider admissions more generally without specification of a distinct subset of patients.

Data

Data differences between studies relate to practical availability as well as the intended contribution of the research. The practical availability of data is due to differences in data collection and recording practices between institutions and regions. As mentioned previously, several models developed for multi-institutional use consist of only a small number of commonly recorded variables. As more variables are included in predictive models, differences in data collection between institutions become more relevant. Thus, predictive models may not be implementable in institutions other than the ones in which they were developed.

Holding constant data recording practices, the intended contribution of the research also plays a role in the data used. Research may aim to demonstrate that data not commonly collected should be captured as part of standard practices to better enable readmission risk assessment. For example, Troyano et al. (2018) looked at the use of an 'electronic nose' to measure compounds exhaled by patients, finding that it did help to identify COPD patients at risk of readmission. Alternatively, commonly collected but rarely considered data may be the focus to prompt its consideration in future model development research. An example of this is the consideration of textual information from electronic health records by Xiao et al. (2018), using an RNN with natural language processing (NLP) tools in addition to using more typical administrative information. These health records contain textual patient information, which has precluded its usage in statistical models but is now potentially valuable with certain machine learning models (Xiao et al., 2018) and application of NLP tools (Sundararaman et al., 2018).

The intended usage of the model also determines what data are relevant for predicting readmissions, particularly with respect to the time at which data are available. If the model is intended to better assess performance with respect to readmission rates or individual outcomes then most data can be used, even those which may only be available weeks after

a patient is discharged, such as certain laboratory results. This type of data would not be usable in the case that the model was intended to inform proactive management of patients. In this case, only data available at the time of discharge should be used to ensure model predictions are available early enough to be useful. Some studies have taken this further and limited consideration to data available shortly after admission (Almardini & Raś, 2017; Benuzillo et al., 2018). The overwhelmingly most common approach was to use data available as at discharge; these are primarily administrative data relating to prior hospital utilisation and basic demographic information such as age and sex. While data categorisations have been suggested (Kansagara et al., 2011), data reporting in the area has not conformed to a detailed and consistent standard.

There has also been great variety in the quantity of data available for analysis for both explanatory and predictive research. This is shown in Table 1. Of the 129 articles included, there were 12 articles where sample size was not relevant (e.g., reviews) or was not reported (Almardini & Raś, 2017; Lagoe et al., 2001). For seven of the remaining 117 articles, the number of observations is estimated because of only partial information being reported, such as the number of patients but not the number of admissions. The differences in cohort size evident in Table 1 tends to be linked to the specificity of the study, both in terms of patient cohort and data under consideration. Those studies with broadly defined cohorts and considering traditional information sources tend to be characterised by larger sample sizes. For example, H. Wang et al. (2016) considered a retrospective general patient population in the context of disease severity and post-discharge outpatient visits, using a sample of 55,532 admissions. Conversely, considering specific patient cohorts and rarely collected or rarely used data can result in smaller samples. Bae, Dey, and Low (2016) conducted a prospective study considering only post-surgery cancer patients and collected behavioural information via Fitbits using a sample of 25 patients.

Table 1. Number of Observations in Reviewed Studies

Range of Observations Included	Frequency	Relative Frequency
0 – 100	14	12.07%
101 – 1,000	36	31.03%
1,001 – 10,000	19	16.38%
10,001 – 100,000	32	27.59%
100,001 – 1,000,000	13	11.21%
1,000,001 – 5,000,000	2	1.72%

Outcome Definition

The last major dimension along which studies differ for both explanatory and predictive research lies in the way outcomes are defined. There are two relevant facets of outcome definitions – the event of interest and timing. The variation associated with outcome definition can be described as deviations from whether a patient was readmitted within 30 days of discharge. This reflects a very general definition of the event of interest, being any readmission, and a choice as to the relevant time frame. While considering any readmission as an event is simple to record, some studies have opted to be more specific or broad in what constitutes an event to more closely link outcomes to the goal of the research. More specific definitions typically address a specific type of readmission, with examples including unplanned readmissions (Thirlwell et al., 2016), readmission related to the original condition (Borer, Kokkirala, O'Sullivan, & Silverman, 2011; Rana et al., 2014; Tan, Jacob, Quek, & Omar, 2006), readmission to the emergency department (Cunha Ferré et al., 2019), and avoidable readmission (Neumann et al., 2004). Broader definitions may define outcomes more generally as adverse health outcomes after discharge. Almardini and Raś (2017) considered readmission and death and da Silva et al. (2016) considered complications as well.

Having 30 days as the default timeframe reflects both its use in healthcare policy and the assumption that readmission shortly after discharge is likely linked to the index admission. While the most common choice, many authors have deviated from the use of 30 days or considered other time frames in addition to it (Jiang et al., 2018; Krompass et al., 2015). Of the studies identified from the systematic search, readmission cut-offs ranged from as early as 72 hours (Cunha Ferré et al., 2019) to as late as three years (Tabata et al., 2014) post-discharge. These are fairly extreme cases, however, with more common time frames being around 60 days (Anderson & Steinberg, 1985; Ruiz, García, Aguirre, & Aguirre, 2008), 90 days (Koulouridis et al., 2015; Pederson et al., 2016; Tyson, Patton, Salevitz, Chen, & Castle, 2014), or one year (Cui et al., 2015; Morrissey et al., 2003; Parker & Hadzi-Pavlovic, 1995). As with more narrowly defined patient cohorts, shorter timeframes were observed to be associated with improved performance.

2.1.3 Discussion of Hospital Readmission Literature

Hospital readmission research remains an active area with respect to both explanatory and predictive focuses. With the key aspects of the reviewed literature already described, more general comments are made here to highlight trends and show where further

research is needed. These relate to the main drivers of continued research in the area and the ramifications of the literature's emphasis on classification rather than survival approaches.

The primary drivers of the continued interest into readmission risk have been the limited generalisability of studies in the area and poor performance of predictive models. With respect to the former, the heterogeneity of studies is a limitation faced by this research but it is not due to poor study design. Instead, it is a feature of the diverse applications and contexts of these readmission models. This is reflected in the major categories of differences mentioned previously, which relate to where and how the predictive models are intended to be applied. A universal model generalising across regions, populations and data availability is not yet a realistic goal. Even holding constant these differences, the second driver of continuing research is the room for improvement on those models which have been put forward. Any improvements in risk modelling for patient management or performance assessment can lead to more equitable resource allocations, and better clinical management and patient outcomes. To improve performance, authors have increasingly considered machine learning techniques as well as incorporating new variables, some of which may not have been possible to include using statistical techniques. There have been generally encouraging results linked to machine learning techniques in comparison to logistic regression, though broad comparisons are difficult to make across studies.

The increasing application of machine learning techniques is the most substantial trend noted in this review. This is largely because of their additional complexity in relationships modelled, lack of assumptions about the underlying data and ability to incorporate previously unused data sources. New information and methods for relating information to readmission risk have been the focus of research aiming to improve on past performance. A secondary and supplementary trend noted was the usage of sampling techniques in recognition of the difficulties faced with imbalanced classification problems. While these sampling techniques are not yet common, it is encouraging that since a dearth of studies using them was noted prior to 2018 (Artetxe et al., 2018) they have increased in frequency of application. This is exemplified by SMOTE, which was used in six studies in a two-year period after no use prior to 2018. These sampling techniques addressing imbalance should become more prevalent in the literature in future, as some recent studies found they improved performance.

Evident also is the emphasis on classification formulations for the readmission problem. Classification formulations align well with performance measurement goals, particularly as used in healthcare policy, while survival formulations provide less standardised but more informative risk predictions. Despite the apparent applicability of survival models to the problem, they are rarely seen in readmission modelling. This is particularly evident when considering the types of models being employed. Even when statistical models were used, this was almost exclusively logistic regression rather than Cox regression. The increased uptake of machine learning methods in recent years has similarly been limited to classification methods with very few exceptions (Hao et al., 2015; Padhukasahasram et al., 2015). This is despite the availability of various machine learning survival techniques. The classification focus has also guided the standard metrics used for performance measures. The AUC, calibration and accuracy metrics provide straightforward measures of classification model performance but are not as relevant or interpretable for survival models unless applied as if they were classification models. In this review, only one instance of a performance measure specific to survival data was identified, this being a modified version of AUC for censored data (Padhukasahasram et al., 2015). Model applications reflecting the information provided by survival models rather than classification models were not suggested in this case, however. The single instance identified of a model application specific to survival models was in generating a daily ranking of 30-day readmission risk for discharged patients by Hao et al. (2015), but models were assessed in this study only based on the standard 30-day readmission risk as at the time of discharge. No studies were identified which considered survival-specific model applications and also evaluated models on the basis of more than a single time point.

Having highlighted the lack of consideration of survival formulations for the problem of readmissions, both in the consideration of machine learning techniques and potential model applications, survival modelling techniques are the focus of the remainder of this review. This includes statistical, decision tree, ensemble, support vector machine, and artificial neural network modelling techniques for survival data. Each of these areas is discussed separately as there was little to no overlap identified between them in the literature, though there were similarities in ways the machine learning methods have been modified to apply to a survival analysis context.

2.2 Survival Analysis Techniques

Before reviewing statistical and machine learning survival analysis techniques, survival analysis problems are briefly described. Survival or time-to-event problems are those in which the time until an event occurs is of interest. Common examples include time until death in clinical trials or time until failure in machinery. It is assumed that, given enough time, the event will occur for all subjects under consideration. For such problems, the goal is to model the risk of event occurrence as a function of time. The direct application of classification or regression techniques is made difficult by the need to consider both risk and time and by the presence of censored data. Censored data are data in which only partial information is known about the exact event time. In this review, the case of right-censored data is the focus, as is the case in most of the reviewed research. For observations which are right-censored, only a lower bound is known for event times. For example, a patient may be known to be event-free for five years at the end of a study, but it is unknown how long they remain event-free afterwards. This form of censoring may be due to the conclusion of a study, equipment being replaced before failure, or losing patients to follow-up. Simply ignoring censored observations results in biased models and ignores the partial information available. Survival analysis techniques, both machine learning and statistical, are those able to account for censoring and model event risk over time. The following subsections provide a review of survival techniques developed in both statistical and machine learning research areas.

While notation will be introduced as required in the following subsections, common elements are described here. Let t_i be the event time of interest, c_i be right-censoring time, and $y_i = \min(t_i, c_i)$ be the last time at which observation i is observed. Further, let $\delta_i = 1(t_i \leq c_i)$ be an indicator variable with value 1 if the event occurred and 0 if the observation was censored. Covariates for the i-th observation are denoted by $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, which is a p-dimensional vector. Conceptually, we can contrast the full data $[t_i, x_i; i = 1, \dots, N]$ usually available for regression modelling with the observed data $[y_i, \delta_i, x_i; i = 1, \dots, N]$ available for survival problems, where N is the number of observations available. Finally, the probability density of event occurrence at time t is denoted $f(t)$, the event rate at time t conditional on no prior event is denoted $h(t)$ and is known as the hazard rate, the probability of no event by time t is denoted $S(t)$, and the probability of an event by time t is denoted $F(t)$.

2.2.1 Statistical Learning Techniques

In traditional statistical approaches to modelling, different techniques are employed depending on the assumptions made about the data being modelled. In time-to-event problems, this has led to the development of a range of techniques for different assumptions about the distributional form of the underlying data and the relationship of covariates with survival times. For this review, techniques are categorised as non-parametric, semi-parametric, or parametric based on the type of assumptions made. Techniques relying on no assumptions are non-parametric methods. These techniques estimate survival functions by considering the empirically observed survival and event times in the at-risk population. Semi-parametric survival techniques avoid specifying a distribution for survival times but do assume a specific form for the relationship of covariates with survival time. Lastly, parametric survival techniques assume both a specific form for the relationship of covariates with survival times and that survival times follow a specified distribution.

Non-parametric and parametric techniques are described first, followed by semi-parametric techniques. This is done to reflect the greater amount of literature relevant to the development of semi-parametric methods, which pose an interesting challenge in allowing for covariate effects without making distributional assumptions. While also important, parts of the survival analysis literature, non-parametric and parametric methods are more straightforward in estimation and relevant extensions and thus are discussed more briefly.

2.2.1.1 Non-Parametric Techniques

If nothing is known about the relationships within, and distribution of, the data being used for time-to-event analysis, then non-parametric techniques are generally the most appropriate starting point. Semi-parametric and parametric models may give poor results on account of incorrect model specification, whereas non-parametric approaches are based only on the empirically observed data. These approaches include the Kaplan-Meier (KM) estimator (Kaplan & Meier, 1958), Nelson-Aalen (NA) estimator (Aalen, 1978; Nelson, 1972), and the actuarial life table (LT) approach (Cutler & Ederer, 1958). The KM estimator is the most widely used (Colosimo, Ferreira, Oliveira, & Sousa, 2002) and estimates the survival function by computing the probability of survival up to a given time as the product of the survival probabilities over a set of previous time intervals. Each interval of time is associated with a probability calculated by considering the number of

people at risk at the start of the interval, and the number of people who experienced the event by the end of the interval. Observations are removed from consideration after censoring, but contribute to the population at risk prior to the censoring event. Formally, the Kaplan-Meier survival function is given as:

$$S(t) = \prod_{j;t_{(j)} \leq t} \left(\frac{r_{(j)} - d_{(j)}}{r_{(j)}} \right)$$

where $t_{(j)}$ represents the j-th ordered event time, $r_{(j)}$ denotes the number of observations which were at risk immediately before $t_{(j)}$, and $d_{(j)}$ denotes the number which experienced the event at $t_{(j)}$.

As the KM estimator uses the product of survival probabilities over a set of intervals, it is often referred to as a product-limit estimator. The NA estimator is an alternative estimator which is also widely used and takes a similar approach but considers the accumulation of hazard over the intervals rather than survival probabilities. With the NA estimator, cumulative hazard or risk is estimated as:

$$\widehat{\Lambda}(t) = \sum_{j;t_{(j)} \leq t} \frac{d_{(j)}}{r_{(j)}}$$

And the corresponding estimate of the survival function is then given as:

$$S(t) = \exp[-\widehat{\Lambda}(t)]$$

Both the KM and NA estimator give very similar results with sufficiently large data sets (Colosimo et al., 2002) and result in step functions for survival probability over time. The less-used life table approach is primarily used in actuarial applications rather than in the survival analysis literature but is more appropriate in the case that interval censoring is present (Malov & O'Brien, 2018).

Non-parametric techniques are very good at avoiding a biased model resulting from an incorrect specification but can suffer from low interpretability because they do not explicitly incorporate any explanatory variables. When a small number of groups is of interest survival functions can be estimated for each group separately. For example, differences in survival functions between male and female patients may be of interest. However, this approach is insufficient in many scenarios where there is a range of potentially important explanatory variables or where variables are continuous. The

limited ability of these techniques to account for covariates makes their use relatively limited outside of large homogeneous datasets or as benchmarks. If interpretability is important or patient information is known to affect survival, then either semi-parametric or parametric techniques may be more appropriate.

2.2.1.2 Parametric Techniques

Parametric techniques in a survival context are closely related to methods employed in typical regression problems. In these methods, a distribution for the data is assumed as well as the relationship between covariates and the dependent variable. The dependent variable may be event time, but it is also frequently taken to be the natural logarithm of event times, with the latter type of model typically being referred to as an "Accelerated Failure Time" (AFT) model (Liu, 2012). For the case where it is assumed that covariates share a linear relationship with the dependent variable, the regression equation is identical to that seen for uncensored data problems:

$$t = \alpha + \beta' x + \epsilon$$

Where α is the intercept, β is the estimated coefficient vector and ϵ is the residual term. While the usual regression equation is used, parameter estimation must account for the fact that some outcomes are censored. This is most often done through maximum likelihood procedures which account for the partial information of censored observations given an assumed distribution. Below, a typical likelihood function is shown in the case that there is no censoring:

$$L(\alpha, \beta) = \prod_{i=1}^{n} f(y_i | x_i; \alpha, \beta)$$

The product is over all follow-up times, which are equivalent to event times in the absence of censoring. By maximising this product with respect to α and β, the likelihood of observing these values is maximised. In the case of right censored data, the dependent variable is known only to be greater than some censoring time and so the likelihood is modified accordingly.

$$L(\beta) = \prod_{i=1}^{n} f(y_i | x_i; \alpha, \beta)^{\delta_i} \big(S(y_i | x_i; \alpha, \beta)\big)^{1-\delta_i}$$

In this formulation, the usual $f(y_i|x_i;\alpha,\beta)$ term is used for cases where the event was observed, while $S(y_i|x_i;\alpha,\beta) = 1 - F(y_i|x_i;\alpha,\beta)$, the probability the observation would have survived at least as long as it did, is used for censored observations. Once again, parameter estimates are obtained by maximising this function with respect to α and β (and any other parameters in the specified distribution of ϵ).

This likelihood formulation addresses the problem of parameter estimation in the presence of right censoring but necessitates distributional assumptions. This contrasts with other estimation procedures which could be employed for uncensored data without explicit distributional assumptions, such as least squares estimates.

Common distributions employed for survival analysis problems include the Exponential, Weibull, Gamma, Lognormal, Log-logistic, Makeham, and Gompertz distributions. For a deeper review of these types of models, see Liu (2012).

2.2.1.3 Semi-Parametric Techniques

In scenarios where a range of covariates are believed to influence the timing and likelihood of events, but a data distribution is unknown, then neither non-parametric nor parametric techniques are ideal. In these scenarios, semi-parametric techniques may be appropriate. Semi-parametric techniques allow for parameter estimation without requiring distributional assumptions beyond specifying the nature of the relationship between covariates and the dependent variable. The lack of distributional assumptions makes model estimation, both for hazard rate and event time models, more difficult than in the parametric case as likelihood-based procedures are no longer immediately applicable.

The challenges of semi-parametric model estimation have led to the proposal of a variety of techniques, though some have gained much wider usage than others. This section aims to first provide a brief overview of a range of models developed for these problems, namely proportional and additive hazard models and the Buckley-James estimator. Other approaches to semi-parametric estimation for linear models have utilised generalised rank tests for estimators (Ritov, 1990; Tsiatis, 1990; Ying, 1993) and Inverse Probability of Censoring Weighted (IPCW) approaches (Koul, Susarla, & Van Ryzin, 1981), but are not described here for brevity's sake. The approaches described below have received the bulk of subsequent research attention and thus are given a more detailed description. After describing these models, extensions of the most prominent models are also outlined.

It should also be noted that techniques are categorised as semi-parametric for the purposes of this section if they assert a relationship between the dependent variable (hazard or event times) and the independent variables without making distributional assumptions. In some cases, this is distinct from how the authors proposing the techniques classify them. This is motivated by a lack of consistency in the definition of semi-parametric techniques and the need for a consistent framework for discussion in this review.

2.2.1.3.1 Cox Proportional Hazards Model

The most famous and widely used model in survival analysis, semi-parametric or otherwise (Liu, 2012), is the Cox regression model put forward by David Cox (1972). In this model, the hazard is the dependent variable considered, and is given as:

$$\lambda(t|x) = \lambda_0(t)\exp(\beta' x)$$

The hazard rate for some time t and covariate vector x is simply a product of the time-varying baseline hazard rate $\lambda_0(t)$ and the exponential of the linear predictor of covariates and coefficient vector. The exponential ensures that the hazard rate is constrained to be positive. This model is semi-parametric in that an assumption is made about how covariates affect the hazard rate, but no assumptions are made about the baseline hazard rate. As the baseline hazard rate is not assumed to follow a given distribution, the usual maximum likelihood procedure cannot be directly applied. To illustrate this, note how the likelihood would appear as the product of the probability density function for events observed to occur and the survival function for events known only to have occurred after some time point:

$$L(\beta) = \prod_{i=1}^{n} f(y_i|x_i;\beta)^{\delta_i} S(y_i|x_i;\beta)^{1-\delta_i}$$

As this likelihood contains the unspecified term it cannot be directly maximised to estimate the parameters of interest. Instead, a 'partial' likelihood is considered, based on the probability of patient i experiencing the event at time t_i conditional on an event occurring at that time. The patients at risk at time t_i are denoted as $\mathcal{R}(t_i)$, and so this conditional probability can be expressed as:

$$\frac{h(t_i|x_i)}{\sum_{j\in R(t_i)} h(t_i|x_j)} = \frac{\lambda_0(t_i)\exp(\beta' x_i)}{\sum_{j\in R(t_i)} \lambda_0(t_i)\exp(\beta' x_j)}$$

In this conditional probability, the baseline hazard rate at time t_i is a common factor in the numerator and denominator and can be removed from consideration.

$$\frac{\lambda_0(t_i)\exp(\beta' x_i)}{\sum_{j\in R(t_i)}\lambda_0(t_i)\exp(\beta' x_j)} = \frac{\exp(\beta' x_i)}{\sum_{j\in R(t_i)}\exp(\beta' x_j)}$$

Using this formulation, the partial likelihood is taken as the product over all event times (the term is set to one in the case of censored observations):

$$PL(\beta) = \prod_{i=1}^{n}\left(\frac{\exp(\beta' x_i)}{\sum_{j\in R(t_i)}\exp(\beta' x_j)}\right)^{\delta_i}$$

This product allows for maximisation with respect to the parameter vector β, avoids any specification of the baseline hazard $\lambda_0(t)$, and accounts for censored observations by considering them as part of the at-risk population in the denominator up until their time of censoring. Some estimation procedures for the baseline hazard have been put forward, such as the Breslow estimator suggested by Breslow in response to the original paper by Cox (1972). Nevertheless, interpretation of the estimated parameters is done on a relative basis. That is, the ratio of hazards for two patients should be a constant proportion over time as only the baseline hazard has a temporal component. Thus, this model is often referred to as Cox's proportional hazards model. As described here, Cox's model assumes there are no tied event times. Modifications have, however, been proposed by researchers such as Breslow (1974) and Efron (1977) for circumstances in which tied event times are present in the data.

2.2.1.3.2 Additive Hazards Model

Cox's proportional hazards model assumes a multiplicative relationship between the hazard rate and covariates. In some instances, however, an additive relationship may be more appropriate. Many authors have suggested additive hazard models, the most common of which is briefly described here.

Aalen proposed an additive model for the hazard function, allowing for time-varying covariates and effects (Aalen, 1989, 1993) without assuming a specific form. This model is expressed as:

$$\lambda\big(t\big|x(t)\big) = \beta(t)'v(t)$$

where $v(t) = \big(1, x(t)\big)$. Cumulative regression functions are easier to estimate than non-cumulative functions when no particular form is assumed, as is the case in Aalen's model. The j-th cumulative regression function at time t is given as:

$$\mathrm{B_j}(t) = \int_0^t \beta_j(s)ds$$

These cumulative regression functions are estimated using:

$$\mathrm{B}^*(t) = \sum_{t_k \leq t} W(t_k) I_k$$

where $W(t_k)$ is a generalised inverse of $v(t_k)$ and I_k is a vector with all elements equal to zero except for the patient with an event at time t_k. From the cumulative regression functions, survival functions can then be derived based on Nelson-Aalen's cumulative hazard or the product limit approach.

Beyond considering an additive rather than a multiplicative relationship, Aalen also asserted that this model addressed two limitations of the Cox model. First, this additive model is much better suited to describing time-varying covariate effects and, secondly, it does not rely on the potentially violated proportional hazards assumption. Diagnostic tests have also been developed to assess the appropriateness of the model, both by Aalen (1993) and other authors. For example, a goodness of fit test which can be adjusted for specific alternatives was proposed by Gandy and Jensen (2005).

Many other authors have also suggested estimators for additive hazard models, though with limited uptake given the prevalence of Cox's proportional hazards. Some have taken similarly unrestricted approaches to covariate effects as in Aalen's model, though others have proposed estimators under the case of constant covariate effects (for example, see Lin and Ying (1994)).

2.2.1.3.3 The Buckley-James Estimator

Another semi-parametric model was proposed by Buckley and James (1979), which has since been termed the "Buckley-James estimator". Rather than focusing on the hazard rate, this model considers the usual linear regression model for estimating event times, though again no explicit distributional assumptions are made beyond the additive relationship between event times and covariates:

$$t = \alpha + \beta' x + \epsilon$$

For censored observations, however, t cannot be directly used and must be modified. The authors propose that t is replaced with t^*, defined for the i-th observation as:

$$t_i^* = \delta_i t_i + (1 - \delta_i) E(t_i | t_i > c_i, x_i)$$

The expectation in this definition is a form of imputation for those observations which are only known to exceed their censoring time. The conditional expectation is based on the product limit estimator of the error CDF.

$$e_i(a, b) = y_i - a - b' x_i$$

$$\hat{F}_{a,b}(\varepsilon) = 1 - \prod_{i; e_{[i]} \leq \varepsilon} \left(\frac{n - i}{n - i + 1} \right)^{\delta_i}$$

Where a and b are estimates of the α and β parameters and $e_{[i]}$ denotes the i-th ordered error term. Using this CDF, the conditional expectation is given as:

$$E(t_i | t_i > c_i, x_i) = b' x_i + \sum_{k; \delta_k = 1, e_k > e_i} \frac{e_k(0, b) \Delta \hat{F}_{0,b}(e_k)}{1 - \hat{F}_{0,b}(c_i - b' x_i)}$$

where $\Delta \hat{\mathrm{F}}_{0,\mathrm{b}}(e_k)$ is the probability mass function for the error terms. The estimation procedure for the model described is an iterative one involving generating estimates a and b of the parameters α and β, recalculating the conditional expectation for censored observations, and then returning to parameter estimation. This process repeats until convergence (or in some cases until the estimates simply oscillate between two values, in which case it is advised to use the midpoint of the two values unless they are very different (Buckley & James, 1979)).

2.2.1.3.4 Extensions

After this brief overview of the types of semi-parametric models which have been proposed in the literature, their continued development and extensions are now discussed. These extensions are split across three focuses. One focus has been the functional form component of hazard rate models to allow for combinations of different forms. Another focus has been on the generalisation of variable selection and regularisation methods from uncensored data contexts to the Cox regression model. This is closely related to the third research focus, in which various models have been adapted to high-dimensional data contexts. This has been motivated by the advent of a big-data environment, particularly in genomic research.

(a) Hazard Models – Functional Form

In the earlier model descriptions, the relationship between the hazard rate and covariates was shown to be multiplicative in Cox's proportional hazards model and additive in Aalen's additive hazard model. The information provided by each model may be complementary, as the models provide insight into two different aspects of how covariates affect hazard. There is a need, however, to choose one as the final model, and this model choice should reflect what is believed to be the true nature of the relationship for the relevant problem. The relationship between covariates and hazard may be multiplicative or additive, and covariate effects within each may be fixed or time-varying.

Covariate Effects in the Additive Model

As mentioned above, additive hazard models can be estimated with covariate effects that are entirely unrestricted and time varying as in Aalen's model (Aalen, 1989) or with constant covariate effects as by Lin and Ying (1994). Combining these structures, McKeague and Sasieni (1994) put forward a version of Aalen's additive risk model in which some covariate effects are time-varying and unrestricted (as in Aalen's original formulation) but others are constant. The proposed model involves iterative estimation of non-constant and constant covariate effects and requires specification as to which covariates should be treated as constant or varying.

Covariate Effects in the Multiplicative Model

In the original formulation of the Cox model, all covariate effects are treated as constant. In some circumstances, however, varying effects may be present, and accordingly several authors have suggested modifications to the original Cox model to extend it to these situations. Examples of such modifications are discussed by Nan, Lin, Lisabeth, and Harlow (2005) and Tian, Zucker, and Wei (2005). Tian et al. (2005) used a kernel-weighted local partial likelihood to estimate time-varying covariate effects. Nan et al. (2005) allowed covariate effects to vary as a function of an index variable (age at a marker event) rather than time. Typically, time-varying covariate effects are approximated using cubic basis splines.

Combining Additive and Multiplicative Components

In cases where some covariates are believed to have an additive relationship with hazard while others have a multiplicative one, it may be desirable to blend the additive and multiplicative structures. Several researchers have endeavoured to develop such models. Lin and Ying (1995) suggested a model and estimation procedures in which the hazard is

given as an additive term with constant covariate effects plus the product of the baseline hazard with Cox's multiplicative term:

$$\lambda(t|x) = g\big(\alpha' V(t)\big) + \lambda_0(t) h\big(\beta' W(t)\big)$$

The vector of (potentially time-varying) covariates x_i is split into those in the additive term $V(t)$ and those in the multiplicative term $W(t)$. The link functions are assumed to be known and represented by g and h. An alternative model was also proposed by Scheike and Zhang (2002), in which hazard is given by the usual Cox regression equation but the baseline hazard is replaced by Aalen's additive hazards equation:

$$\lambda(t|x) = \big(\alpha'(t) V(t)\big) \exp\big(\beta' W(t)\big)$$

This model also differs from that of Lin and Ying (1995) in that time-varying effects are allowed for in the additive component as part of the baseline. Similarly, Martinussen and Scheike (2002) proposed an additive-multiplicative hazard model. In this model, the hazard consists of Aalen's additive term and the product of the baseline hazard with Cox's multiplicative term:

$$\lambda(t|x) = \alpha'(t) V(t) + \lambda_0(t) \exp\big(\beta' W(t)\big)$$

Again, this allows for more time-varying effects, and is different from that of Scheike and Zhang (2002) in that it takes the sum rather than the product of the additive and multiplicative components. The models proposed by both Scheike and Zhang (2002) and Martinussen and Scheike (2002) are estimated through the solving of the model's score equations.

Variable and Structure Selection

All the hazard models described above imply that two decisions need to be made in model selection processes. The first decision relates to which variables should be included in the model. The second decision relates to the model structure in terms of whether each covariate effect is modelled with a constant or non-constant effect. Methods considering both decisions or only the second are described here, while a more extensive overview of variable selection and regularisation methods is deferred to subsection (b) below.

For the Cox model, S. Zhang, Wang, and Lian (2014) proposed an approach for variable selection with non-constant covariate effects using the Smoothly Clipped Absolute Deviations (SCAD) penalty in conjunction with a B-spline basis expansion. They did not,

however, provide a method for determining the structure for covariate effects. Du, Ma, and Liang (2010) proposed an iterative approach for simultaneous variable and structure selection, in which non-constant and constant effects are estimated separately in a penalised framework with repeated substitution.

More commonly, variable and structure selection has been performed through decomposing every covariate effect into a non-constant and constant component, after which a doubly penalised likelihood is employed to simultaneously shrink both aspects. Such an approach was taken by J. Yan and Huang (2012), who used a B-spline basis expansion to approximate the non-constant effects and an adaptive LASSO penalty. A similar approach was also employed by Lian, Lai, and Liang (2013) with a SCAD penalty and by Honda and Yabe (2017) who used an orthonormal basis for covariate decomposition. If non-constant and constant effects are both non-zero, then the covariate is modelled with a non-constant effect. If only the constant effect is non-zero, then the covariate is modelled with a constant effect, and the covariate is not included if both effects are shrunk to zero. Unlike the iterative method proposed by Du et al. (2010), these methods consider both aspects of model selection simultaneously. An extension to moderately high-dimensional data was also considered in this context by Honda and Härdle (2014), who showed that their method using a group SCAD or LASSO is applicable to cases where the number of covariates is moderately large relative to sample size.

(b) Penalised Estimation Methods

This concludes the overview of research into different functional forms for hazard models as well as related model selection methods. The focus is now free to turn to penalised estimation methods that are employed to perform variable selection and regularisation.

Much of this research aims to extend failure time models (particularly the Cox model) to include popular methods used in uncensored data situations. A brief description of the common formulation of such estimation procedures is provided here to highlight how modifications to this general form are required for censored data. It is established how the estimation problems are usually formulated; the focus can then move to a discussion of the variety of methods put forward and the differences between them. For all the methods discussed in this subsection, it should be noted that it is common for covariates to first be rescaled to have common standard deviation. This makes the related coefficients unitless and thus directly comparable.

While variable selection and regularisation methods can take many forms, such as stepwise or Bayesian procedures (e.g. Faraggi & Simon, 1998), most research has focused on penalised estimation procedures. In such methods, variable selection and regularisation is achieved by adding a penalty term to the usual estimating equations which imposes a cost for coefficients relative to their magnitude. Parameters are most commonly estimated to minimise the residual sum of squares plus a penalty term, often referred to as penalised least squares.

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta' x_i)^2 + P_\lambda(\beta)$$

In this general equation, there is usual residual sum of squares coupled with a method-specific penalty term $P_\lambda(\beta)$, with the weight accorded to the penalty determined by the tuning parameter λ. These methods cannot be directly applied to censored data as the residual sum of squares cannot be computed. Alternative minimisation problems differ depending on the model of interest, as discussed in the following paragraphs.

For semi-parametric linear models, the formulation of the minimisation problem is very similar to that shown above, particularly for the Buckley-James model. Synthetic data are constructed by iteratively imputing the dependent variable for censored cases. Thus, the minimisation problem to be solved is different from the above only in that the residual sum of squares is calculated using synthetic, imputed data for censored observations. This type of approach was employed in a three-step algorithm put forward by Johnson (2009) to allow for application of popular penalties such as the LASSO and related extensions in parameter estimation. This three-step algorithm addressed numerical and theoretical property limitations of similar previous algorithms (for examples, see Johnson (2008) and Jin, Lin, and Ying (2006)) by initialising the usual Buckley-James estimator with estimates that are root-n consistent. While Jin et al. (2006) also used a root-n consistent initial estimator, they required stronger assumptions for theoretical properties than in the method described by Johnson (2009). The three steps of this algorithm can be succinctly summarised as first obtaining consistent initial parameter estimates, then using this to construct the synthetic response data, and finally applying the LASSO model to the observed and synthetic data.

For hazard models using a likelihood, specifically the Cox model, the likelihood function serves as an alternative to the least squares objective function. Accordingly, a penalised likelihood can be used. This is expressed in the general form as:

$$\arg\min_{\beta} Q(\beta) + P_\lambda(\beta)$$

where $Q(\beta)$ is the negative log-likelihood as a function of the parameters being estimated and $P_\lambda(\beta)$ has the same interpretation as previously. The logarithm of the likelihood is a monotonic transformation and is used to make computations more tractable. The negative of this term is taken to make this a minimisation problem. The penalised likelihood formulation makes it relatively straightforward to apply most penalised estimation procedures developed for fully observed data to the Cox model. As an aside, this formulation is also applicable to fully parametric models.

A large variety of penalties has been considered, though all penalties aim to regularise or shrink coefficients towards zero. This regularisation is intended to prevent or at least mitigate the effect of overfitting in model estimation and thus to improve the generalisability and predictive performance. The variety stems from the way different penalty functions influence the final coefficients estimated. Some penalties, such as the ridge penalty, will not cause coefficients to be shrunk to exactly zero. Others, such as the popular LASSO penalty, force some coefficients to be zero and thus also achieve variable selection in addition to shrinkage. The choice of penalty is driven by problem-specific considerations and objectives, with several notable penalties described below.

<u>The Ridge</u>

The ridge penalty (Hoerl & Kennard, 1970) is one of the most common regularisation methods used for penalised estimation. With ridge regularisation, the L_2-norm of coefficients on the p covariates are penalised:

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \beta_j^2$$

Because of the nature of this penalty, coefficients are shrunk but are not set to exactly zero, meaning that regularisation is performed without a variable selection component. This method also allows for the inclusion of a greater number of covariates than training samples in model estimation.

The LASSO

The Least Absolute Shrinking and Selection Operator, or LASSO (Tibshirani, 1996) is one of the most widely used penalties employed in statistical modelling and is a common alternative to the ridge penalty. The LASSO penalises the L_1-norm of coefficients on the p covariates:

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$$

As a result of the nature of the L_1-norm penalty and as implied by its name, estimation with the LASSO penalty causes some coefficients to be set to exactly zero, meaning that variable selection is performed as well. Its use was illustrated for the Cox model through a straightforward penalised likelihood procedure by Tibshirani (1997) and it has been extensively used since. Unlike ridge, it does not allow for the inclusion of more covariates than training samples.

Several modified versions of the LASSO have also been developed. The fused LASSO (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005) is a variant which penalises both the L_1-norm of coefficients as well as their successive differences:

$$P_\lambda(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$$

The additional term in this penalty encourages equivalence of successive coefficients. This form of penalty is applicable in the case where there is a meaningful order or grouping of predictors (Tibshirani et al., 2005), particularly for high-dimensional data. This feature of the penalty has led it to be applied in the case of gene expression data where individual genes may share a gene pathway as well as certain other problems such as those involving copy number data (Chaturvedi, de Menezes, & Goeman, 2014). An algorithm for implementing the fused LASSO with the Cox model was described by Chaturvedi et al. (2014) as part of a more general penalised likelihood implementation.

The adaptive LASSO is a variation which was developed to address the fact that the original LASSO model does not have the oracle property (Zou, 2006). A model is said to have the oracle property if it is consistent in variable selection as well as coefficient estimation. Rather than change the penalty structure or add an additional aspect like the

fused LASSO, the adaptive LASSO considers covariate-specific weights selected on an adaptive basis. Specifically, the weight for each covariate is set as the inverse of the unpenalised coefficient estimate, denoted w_j. The penalty function is thus given as:

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} w_j |\beta_j|$$

The adaptive aspect of this penalty is intuitively appealing, as covariates with a large effect are subject to a smaller penalty weighting than smaller, less influential covariates. It was extended to the Cox model by H. H. Zhang and Lu (2007) and remains a popular choice. Beyond the standard Cox model, it has also been employed for variable selection for the time-varying coefficient Cox model (J. Yan & Huang, 2012).

The Elastic Net

The ridge and LASSO penalties have also been considered as a composite penalty rather than being competing options. The Elastic Net (EN) penalty (Zou & Hastie, 2005) consists of a weighted average of the ridge and LASSO penalties, expressed as:

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \left[\alpha \beta_j^2 + (1-\alpha)|\beta_j|\right]$$

The implementation of the EN method for the Cox model (Simon, Friedman, Hastie, & Tibshirani, 2011) is important for the large set of high-dimensional problems such as genomics to which Cox and other survival models are applied. In these contexts, the variable selection feature of the LASSO is desirable to identify influential features but has the undesirable feature of arbitrarily selecting only one covariate from groups of highly correlated covariates which would be more faithfully represented as a collection of covariates. The ridge penalty better achieves this grouped representation, shrinking coefficients towards zero but not removing them entirely. While EN is only partially effective at addressing correlated attributes, particularly for high-dimensional data (Vinzamuri & Reddy, 2013), it represents a balance between variable selection and group representation from the two penalties.

The SCAD

A common alternative to the unmodified LASSO is the Smoothly Clipped Absolute Deviations (SCAD) penalty (Jianqing Fan, 1997; Jianqing Fan & Li, 2001). As with the

LASSO, the SCAD penalty performs variable selection, but it also has the oracle property. It is given as:

$$P_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| & if\ |\beta_j| \le \lambda, \\ \dfrac{2a\lambda|\beta_j| - \beta_j^2 - \lambda^2}{2(a-1)} & if\ \lambda < |\beta_j| \le a\lambda, \\ \dfrac{\lambda^2(a+1)}{2} & otherwise. \end{cases}$$

$$P_\lambda(\beta) = \sum_{j=1}^{p} P_\lambda(\beta_j)$$

This formulation allows for a reduction in penalisation of larger coefficient estimates, while other coefficients may be set to zero. The SCAD penalty has also been shown to retain its oracle property when adapted to the Cox model (Jianqing Fan & Li, 2002).

The OSCAR

Another composite penalty is the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR), which was proposed for contexts in which variable selection and variable group representation are of interest (Bondell & Reich, 2008). This penalty function combines both the L_1-norm and the pairwise L_∞-norm:

$$P_\lambda(\beta) = \lambda\left[\sum_{j=1}^{p}|\beta_j| + c\sum_{j<k}\max\left(|\beta_j|, |\beta_k|\right)\right]$$

As with the LASSO, OSCAR achieves variable selection by setting coefficients of unimportant predictors to zero, but also simultaneously encourages equality of coefficients for variables which are highly correlated. This latter feature of the function means that groups are identified and represented as such (with a single coefficient) without a need for specification of groups, making it distinct from related methods which have an explicit step for group identification (Bondell & Reich, 2008). It was adapted to Cox regression by Vinzamuri and Reddy (2013) and has found some use in high-dimensional data problems.

Penalty weight selection

The description of the above methods has focused on how variable inclusion is influenced by the form of penalty applied. A second element of penalised estimation procedures is the weighting, which controls how much the model is affected by the penalty. While not

as actively researched, several authors have provided suggestions for how these weightings should be selected. Verweij and Van Houwelingen (1994) described a general framework for implementing penalised likelihood estimation for Cox regression and they advocated the selection of the weight parameter by maximising the model's cross-validated predictive value. That is, the weight parameter which offers the best performance on the basis of an appropriate cross-validated metric should be employed, with the choice of metric being up to the user. Huang and Harrington (2002) suggested that a resampling approach be utilised instead, motivated by the instability of likelihood estimators when there are correlation structures in the data and the ratio of covariates to observations is high. They found that while both cross-validation and bootstrap procedures reduced the mean-squared error of the final model, the bias associated with bootstrapping was smaller (Huang & Harrington, 2002).

(c) High-Dimensional Data Adaptations

The third research focus relates to adapting survival methods to problems where data are high-dimensional, having more features than observations. Gene expression data represent a prominent example of very high-dimensional data, with thousands of features often associated with fewer than 100 observations. With improvements in data collection technology, high-dimensional data are becoming more common and represent unique problems for model estimation. Methods developed for such data aim to allow for model estimation through variable selection, dimension reduction, or a combination of the two.

Variable Selection

Several variable selection methods have been proposed for high-dimensional data, of which some represent standalone estimation procedures while others take place prior to model estimation. Among the latter type are screening procedures, which are typically used to complement subsequent estimation procedures. Screening procedures consider all features and quantify the strength of each feature individually with outcomes. A threshold value can then be used to select the subset of features having the strongest relationship with outcomes. Subsequent procedures, such as model estimation (Ma & Huang, 2007), dimension reduction or clustering (Bair & Tibshirani, 2004), consider only this reduced set of relevant features. For uncensored data this type of feature reduction can be performed using Pearson correlations, but for survival problems alternative measures must be employed such as univariate Cox scores (Bair & Tibshirani, 2004) or rank statistics (Song, Lu, Ma, & Jessie Jeng, 2014). The IPCW rank statistic proposed by Song

et al. (2014) can also be used and is applicable for a general class of survival models while also being robust to outliers in covariates through the nature of its construction.

Penalised estimation procedures are also employed for variable selection in high-dimensional problems, but they generally require modifications to the original algorithms. The ridge penalty can be used without modification, but as it does not perform variable selection it will potentially result in thousands of coefficients. The LASSO algorithm in its original formulation does not guarantee at least one solution because of the nature of the algorithm, but the later-developed Least Angle Regression (LARS) algorithm has the LASSO as a special case and is applicable to high-dimensional problems (Efron, Hastie, Johnstone, & Tibshirani, 2004). This has allowed for its application to problems in which the number of features is greater than the number of observations, though the final model estimated will have fewer non-zero coefficients than observations. This has been referred to as the LARS-LASSO and it has been used with both the Cox (Gui & Li, 2005) and additive hazard models (Honda & Yabe, 2017; Ma & Huang, 2007). Implementation of penalised estimation while allowing for inclusion of a greater number of features than observations has also been achieved by formulating the problem in terms of reproducing kernel Hilbert spaces (Li & Luan, 2003), functions ordinarily seen in the context of support vector machines to achieve their high-dimensional nature. They have been employed in the Cox model for general estimation (Li & Luan, 2003) and by Vinzamuri and Reddy (2013) to implement a kernel variation of the Elastic Net (KEN) penalty. The EN penalty for high-dimensional data has also been employed for the Buckley-James model (Sijian Wang, Nan, Zhu, & Beer, 2008).

An additional desirable feature of a penalised estimation procedure for many high-dimensional data problems is the consideration of feature groups. A common example of this is a collection of gene expressions (individual features) in a gene pathway (a grouping). One approach taken to dealing with the grouping of variables is to use a hierarchical penalty considering individual and groups of covariates (S. Wang, Nan, Zhou, & Zhu, 2009). For each variable j in group k, the variable effect is decomposed into group and variable-specific effects, $\beta_{jk} = \gamma_k \theta_{jk}$ with $\gamma_k \geq 0$. The penalty term is then expressed in terms of these effects:

$$P_\lambda(\beta) = \sum_{k=1}^{K} \gamma_k + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\theta_{kj}|$$

where there are K non-overlapping groups of variables and p_k variables in group k. The first term affects the entire group while the second considers individual variables. In addition to this formulation, an adaptive variant with covariate specific weights as well as the case of overlapping groups of covariates can also be used (S. Wang et al., 2009).

Other penalties such as the OSCAR or KEN have also been proposed as methods which can account for grouping in high-dimensional data, and unlike the hierarchical penalty they do not require prior knowledge of groupings (Vinzamuri & Reddy, 2013).

Lastly, while penalised estimation methods allow for variable selection in high-dimensional problems, they impose penalties on all covariates. In certain situations, it may be desirable to include select covariates in an unpenalised fashion to assess whether additional terms add new predictive power to the model. Motivated by such situations, particularly for considering whether microarray data add information to clinical, Binder and Schumacher (2008) proposed an offset-based boosting algorithm for the Cox model, referred to as "CoxBoost".

Dimension reduction

As most models cannot be estimated with a greater number of covariates than observations, dimension reduction methods have been employed to summarise many features into fewer aggregate features. Model estimation can then proceed in the usual manner. While interpretation of the new covariates is made more complicated because of the way they are constructed, these methods retain most information available from the original high-dimensional case. The two most prominent such methods are principal components analysis and partial least squares.

The principal components method, in essence, considers the design matrix x and constructs a set of vectors as linear combinations of the original features. Specifically, it constructs the eigenvectors of the variance-covariance matrix of the predictors. These new vectors are orthogonal to one another and explain the maximum possible variance from the original design matrix. These vectors are referred to as the principal components, which are then used as the covariates in model estimation. An example of a principal component analysis approach is provided by Bair and Tibshirani (2004), who proposed a method in which covariates are originally screened according to their univariate Cox scores before the principal components of the resulting subset of covariates are computed. These principal components are then used as predictors in model estimation. Such an

approach has also been used in conjunction with Aalen's additive risk model, which suffers from dimensionality limitations even in the case of only a moderately large numbers of features (Tan et al., 2006). The application of principal components after an initial screening step, in which features with univariate coefficients below some threshold are removed, has been termed "supervised" principal components in uncensored problems (Bair, Hastie, Paul, & Tibshirani, 2006) because only features related to outcomes are considered.

The method of partial least squares is closely related, and it similarly finds orthogonal components explaining the maximum variance, but unlike principal components the response vector Y is explicitly considered in their construction. In Li and Gui (2004), this method was extended to the Cox model through repeated fitting of the least square residuals. While previous authors had also combined partial least squares with the Cox models, direct applications were used whereas Li and Gui (2004) developed a general extension to the Cox model. In addition to the Cox, the Buckley-James algorithm has also been combined with partial least squares in an iterative implementation (Huang & Harrington, 2005).

2.2.1.4 Summary of Statistical Survival Literature

While statistical models were broadly categorised as being non-parametric, semi-parametric or parametric in this section, most survival analysis research has been conducted using semi-parametric models. Originally, research in this area focused on the development of models which did not rely on distributional assumptions while still allowing for the incorporation of covariates. As such models became relatively well established, the major research focuses have shifted towards extending those already developed. Extensions have come in the form of alternative specifications of the relationship between outcomes and explanatory variables, penalised estimation, and adaptations to high-dimensional data. These latter two areas in particular reflect the need to adapt the developed models for the types of data increasingly being encountered in modern problems.

Two factors have driven the emphasis accorded to semi-parametric methods for survival analysis. First, non-parametric methods in statistical frameworks are limited by their inability to adequately account for covariates. Secondly, parametric methods have not required substantial survival-specific research on account of the ease with which existing

models can be applied through likelihood estimation procedures. As such, for parametric models, developments would largely be consistent with developments in statistical modelling more generally. In terms of non-parametric developments, most of this research has occurred in machine learning rather than statistical contexts. Such research is the focus of the subsequent sections.

2.2.2 Trees and Ensembles

Decision trees (Breiman, Friedman, Stone, & Olshen, 1984) are a very commonly employed non-parametric technique for both classification and regression problems. They are based on repeated partitions of the data into increasingly homogeneous groups, most commonly using binary splitting rules. This technique provides non-parametric and highly interpretable structures, and automatically captures interactions present in the data without need for prior specification.

The algorithm can be described in the general sense simply. To start, consider all data to belong to a single group, the 'root' node of the tree. Next, consider all possible binary splits which can be made based on the covariates x, with the splits for numeric or ordinal variables being of the form $x_j < c$ versus $x_j \geq c$, where c is some threshold value and x_j is the j^{th} covariate. Nominal variables can also be used, though splitting rules based on thresholds are not applicable for this type of data. Observations are instead split according to a rule of the form $x_j \in \text{A}$ versus $x_j \notin A$ where A is a subset of the nominal variable's possible values. For both types of rules, each split will result in two 'child' nodes. An appropriate splitting statistic is used to evaluate possible splits and select the actual split used. For example, entropy is often used for classification tasks, and the split which reduces entropy the most is used for each node. The Gini coefficient is also commonly used instead of entropy. This splitting process is repeated, with each child node in turn being considered as a parent node, until some stopping criterion is met. Common stopping criteria include the number of observations in a node being less than some value or all observations having the same class. The terminal nodes (nodes with no children) are then taken as the final outputs of the tree. For classification tasks, the tree prediction is the most common class in the terminal node associated with a new data point. For regression tasks, the prediction is the mean of the terminal node associated with a new data point.

Traditional decision trees and splitting statistics suffer from the same limitations as other techniques for censored data in that they do not automatically consider partial information

available from censored observations. Modifications to the traditional methods for application to survival analysis has, however, been a very active field of research. Several sub-areas are identified in this section of the review, though some overlap does exist. The first sub-area considered relates to the initial development of 'survival trees', which were adapted to survival data by modifying the splitting functions used. The second discusses the use of censoring unbiased transformations in the survival tree construction process to retain direct relationships between partial and full data loss functions. The third considers ensemble techniques involving combining many trees. A brief overview of related research is also provided.

2.2.2.1 Splitting Statistics and Individual Survival Trees

The flexibility of decision trees has allowed for relatively straightforward application to survival data through modification of the splitting statistics. If a splitting method appropriately considers both uncensored and censored observations in parent and children nodes, then the construction of a decision tree can proceed in the usual manner, with terminal nodes summarised by non-parametric survival functions. Early research into the application of decision trees to survival data revolved primarily around the use of different splitting statistics, with many options now available. For example, Leblanc and Crowley (1992) proposed a splitting algorithm rooted in the proportional hazards model and its associated likelihood function. This algorithm chooses splits to maximise the one-step full likelihood of the proportional hazards model over the constructed tree, which explicitly accounts for the presence of censoring. The number of algorithms available has, however, raised the question of which should be used and in what circumstances, serving as the motivation for several comparative studies. Radespiel-Tröger, Rabenstein, Schneider, and Lausen (2003) considered seven splitting methods for building survival trees on a simulated data set and a clinical gallstone dataset with a relatively high level of censoring (35%) and found that a variant of the log rank statistic offered the lowest prediction error. In a later study, Radespiel-Troger, Gefeller, Rabenstein, and Hothorn (2006) used the same clinical data set to investigate whether there is a relationship between covariate split selection stability in the root node of the tree and the associated predictive error. Six splitting methods were compared, with the study concluding that there is such a relationship between covariate split selection stability and predictive error, finding once again that a log rank statistic offered the best performance in this dataset with high censoring. The generalisability of both studies is limited, however, by the small

number of scenarios considered. Both used only a single clinical dataset and while Radespiel-Tröger et al. (2003) supplemented this with simulated data, only one hazard structure was considered. More recently, Shimokawa, Kawasaki, and Miyaoka (2015) compared nine splitting methods on a much more comprehensive range of simulations. In addition to varying censoring prevalence and volume of data used in training, various hazard functions were also considered, finding that the decision as to which splitting method to use should be based on expectations of the shape of the hazard function being modelled and the prevalence of censoring. No single method was found to be optimal in all scenarios, and in the case of a 'bathtub' shaped hazard function it was recommended that tree structures not be used at all (Shimokawa et al., 2015).

2.2.2.2 Censoring Unbiased Transformations and Individual Survival Trees

A potential limitation of many splitting functions for survival trees is that they adjust for censored data with methods distinct from those which would be used if the full data were observed (Molinaro, Dudoit, & van Der Laan, 2004; Steingrimsson, Diao, & Strawderman, 2019). That is, the proposed risk measures for partial data are not directly related to the risk measures which would be employed for the full data scenario, resulting in sub-optimal estimates for the measures which would have been used otherwise (Molinaro et al., 2004). In response to this issue, several authors have proposed methods for adapting the loss function used in survival trees to retain a direct relationship to the appropriate full data loss function (quadratic loss in the case of many regression scenarios). The proposed methods can be described as "censoring unbiased transformations", or CUTs, in which the full data loss functions are generalised to the partial data scenario.

A function of the observed (partial) data is a CUT of a given risk measure if it shares a conditional expectation with that risk measure for all covariate combinations (Steingrimsson et al., 2019). More formally, if $L_F(t, x)$ is a loss function for the full data, then $L_P(y, \delta, x)$ is a CUT for L_F if $E(L_P|x) = E(L_F|x)$ for all possible x (as described by Steingrimsson et al. (2019), albeit with slightly different notation). The simplest such method used for survival trees is often referred to as Inverse Probability of Censoring Weighting (IPCW), which is a specific case of a CUT. This is essentially a weighted complete-cases approach to constructing a survival tree, in which censored observations are not explicitly included and the weighting of uncensored observations is inversely proportional to the probability of being censored after the observed event time. Under

quadratic loss and setting $G(t|x)$ to be the probability of no censoring by time t conditional on the covariates x, the new loss function is given by:

$$L_{IPCW} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{G(y_i|x_i)}(\hat{y}_i - y_i)^2$$

Demonstrating equivalent expectations between the IPCW loss function and the full data loss function in the presence of censoring is trivial, and in the case of no censoring the IPCW term is always equal. This method requires two assumptions: (1) event time t is conditionally independent of censoring time c given the covariates x and (2) $G(u|x) > 0$ for all $u > 0$. Molinaro et al. (2004) proposed such a method for constructing survival trees under general loss functions for univariate or multivariate outcome predictions, as well as for density estimation. They modelled the censoring survival function G using Kaplan-Meier estimates and compared their IPCW trees to the method proposed by Leblanc and Crowley (1992). Molinaro et al. (2004) found superior IPCW tree performance on simulated data (using a censoring proportion of 20%) with respect to quadratic and absolute loss functions. It should be noted that this approach also allows for informative censoring, in which the covariates may affect the likelihood of censoring, and thus it does not require that the censoring survival function be estimated using non-parametric methods.

IPCW trees were extended by Steingrimsson, Diao, Molinaro, and Strawderman (2016), who also aimed to address the gap between full data and partial data loss functions. Steingrimsson et al. (2016) proposed doubly robust survival trees which more efficiently use the partial information available and are, as the name indicates, more robust than the IPCW method. Under the IPCW method, the model is consistent provided the censoring mechanism is correctly specified. Under the doubly robust method, the estimator is consistent if either of the censoring or survival mechanisms are correctly specified. This is achieved by adding an augmentation term to the IPCW loss function which is an unbiased estimate of zero if the estimate of G is consistent. If only the estimate of S is consistent, then the augmentation term corrects for the bias introduced by mis-specifying G. Finally, if both estimates for population S and G are consistent, then the estimator is more efficient with respect to loss function and predictors than that of IPCW (Steingrimsson et al., 2016). A generalisation of this doubly robust survival tree was also developed more recently by Steingrimsson et al. (2019) who provide a more detailed

description of CUTs and propose Censoring Unbiased Regression Trees (CURTs). CURTs are trees constructed using a CUT of a full data loss function, having IPCW trees and doubly robust trees as two special cases.

2.2.2.3 Ensemble Methods

As is the case with decision trees more generally, survival trees tend to suffer from high variability with respect to the training data used in model construction (Hothorn, Lausen, Benner, & Radespiel-Tröger, 2004). Ensemble methods address this issue by aggregating the predictions of many individual trees. Hothorn et al. (2004) proposed a bootstrap aggregation ('bagging') algorithm for survival trees for any specified splitting rules. This algorithm generates many bootstrapped data sets, and for each a survival tree is constructed. For a new data point, predictions are generated by constructing an aggregated Kaplan-Meier survival function based on the data in the corresponding terminal nodes for each constructed survival tree. The authors distinguish this from the aggregation of point estimates, as the generation of the survival function is more informative (Hothorn et al., 2004). When compared with individual survival trees, the bagging approach was found to offer superior performance on several simulated scenarios (with censoring levels ranging from 0% to 50%) as well as on two clinical datasets.

Random forest approaches have also been considered in the literature, addressing the marginal correlation between constructed trees in bagging methods. This is done by inserting additional randomness into the tree construction process by considering a random subset of candidate covariates for splitting at each node in each tree. Ishwaran, Kogalur, Blackstone, and Lauer (2008) set out random survival forests and distinguished their approach from others by requiring that the splitting method employed, and by extension all other aspects of the tree, directly consider both event and censoring times. This is consistent with the original formulation of random forests, which requires that all aspects of random forest construction account for outcomes. In this case the outcomes of interest are both event and censoring times. While this does exclude several splitting methods, there remain a number which meet this criterion (such as the log rank splitting rule), with Ishwaran et al. (2008) demonstrating the application of random survival forests under four different methods. Application to eight distinct datasets demonstrated performance at least as good as an IPCW random forest model proposed by Hothorn, Bühlmann, Dudoit, Molinaro, and Van Der Laan (2006) and a Cox regression. A later study on the properties of this random survival forest established the consistency of this

method within the context of a finite feature space of discrete covariates (Ishwaran & Kogalur, 2010).

CUTs have also been employed in constructing related ensemble methods (Hothorn, Bühlmann, et al., 2006; Steingrimsson et al., 2019). Hothorn, Bühlmann, et al. (2006) described a random forest algorithm based on IPCW, which differs greatly from the later proposed random survival forest in that it does not consider censored observations in the component trees. Instead, the forest is constructed using multinomial bootstraps with IPCW. That is, every observation i is assigned a sampling probability equal to $\frac{\delta_i}{G(y_i|x_i)}$ where a Kaplan-Meier estimator is used for G. One consequence of this is that out-of-bag error is no longer a reliable measure of the robustness of the model if any observations have very high sampling probabilities, as these observations may appear in nearly all bootstrapped trees. Another random forest approach based on CUTs was proposed by Steingrimsson et al. (2019), using a form of response imputation as well as exchangeably weighted bootstrapping. The IPCW random forest is also a special case of the general Censoring Unbiased Regression Ensemble (CURE) framework proposed by Steingrimsson et al. (2019). The CURE framework can be used with any appropriate CUT, including IPCW and doubly robust methods.

Two other ensemble tree methods have been proposed in the literature, based on extremely randomised trees and Bayesian methods. R. Zhu and Kosorok (2012) put forward Recursively Imputed Survival Trees (RIST), which are based on the extremely randomised trees (ERTs) variation of the random forest. ERTs add further randomisation to the random forest by considering splits for each covariate chosen completely at random (R. Zhu & Kosorok, 2012). The RIST method uses ERTs in conjunction with imputation for censored event times in an iterative procedure. When compared to the random survival forest, IPCW random forest, and a Cox regression on five simulated and two clinical data sets, the RIST method had significantly better performance in almost all instances (R. Zhu & Kosorok, 2012). Exceptions to this were understandable, such as the Cox model outperforming when the simulated data satisfied the assumptions of this model. R. A. Sparapani, Logan, McCulloch, and Laud (2016) proposed Bayesian Additive Regression Trees (BART), which are based on adding many weak learner survival trees in a Bayesian framework with priors on many aspects of the model, including tree complexity, variables

used for splits, splitting rules, and terminal node summaries. The posterior distribution of the Bayesian method is solved through MCMC methods.

2.2.2.4 Related Research

Related research has also considered multivariate scenarios (Juanjuan Fan, Su, Levine, Nunn, & LeBlanc, 2006; Su & Fan, 2004), splitting methods based on hyperplanes (Kretowska, 2004), and use of decision trees for incorporating stepwise time-varying regression effects (Xu & Adak, 2002).

Multivariate survival analysis involves the usual scenario where the time to an event is of interest, but all observations fall into distinct groups, with intra-group correlation. An example of such a scenario would be tooth decay. The event of interest is time until a filling or similar measure is needed, but it would be expected that teeth for the same person are all correlated, resulting in a multivariate problem in the case that the teeth of multiple people are observed. Juanjuan Fan et al. (2006) considered such a problem and proposed the use of the robust log rank statistic for splitting, as this statistic accounts for intra-group correlation. It is also useful in the case that the groupings are 'nuisance' parameters, which are not of direct interest. This is contrasted with the approach of Su and Fan (2004), who employed frailty models and their associated likelihood functions. Under a frailty model, the normal Cox hazard function applies, but is multiplied by a group-specific frailty term typically assumed to follow a Gamma distribution. Formulating the problem in the context of an integrated log likelihood with Cox and Gamma distribution components allows for the use of a splitting rule which maximises this log likelihood. Application to simulated data appeared to favour this frailty model approach over a robust log rank statistic when the underlying data-generating mechanism is itself a frailty model (Su & Fan, 2004), but the robust log rank method may be more appropriate for retaining covariate-specific information (Juanjuan Fan et al., 2006).

2.2.2.5 Summary of Tree and Ensemble Literature

In summary, several themes of research have been identified in the field of survival analysis for tree and related ensemble techniques. In the 1980s and 1990s a great deal of research revolved around the use of different splitting statistics which extended decision trees to survival data. In more recent years, several authors have performed comparative studies to shed light on which of these splitting statistics performs best and in what scenarios. In the last two decades, however, research has moved towards the use of CUTs and ensemble methods, with some overlap between the two. Use of CUTs with survival

trees have aimed to allow the use of partial data loss functions which are directly related to full data loss functions. Ensemble techniques have been found to offer performance gains over individual trees in both regression and classification contexts elsewhere, and this has served to help motivate their adoption in survival analysis. A variety of distinct ensemble techniques has been proposed for survival data, including random survival forests, censoring unbiased regression ensembles, recursively imputed survival trees, and Bayesian additive regression trees. However, there does not seem to be a clear consensus in this literature as to which of the many techniques should be used in which situations. This lack of guidance may be exacerbated by the different principles underlying the different techniques.

2.2.3 Support Vector Machines

Support Vector Machines (SVMs) (Cortes & Vapnik, 1995) were originally developed for classification purposes and have also been successfully extended to regression problems as well. They are attractive as a method because they are rooted in statistical learning theory, can model both linear and non-linear relationships, and they also avoid the curse of dimensionality through kernel functions.

SVMs are most seen in classification tasks and aim to separate binary classes of examples with a linear classifier by maximising the margin between the two groups. Their widespread usage has been motivated by their use in conjunction with the 'kernel trick', which maps the original features to a higher dimensional space without explicit calculation of each individual feature in the new space. That is, a linear classifier can be defined in a high-dimensional feature space based on a transformation of the original input features, without requiring the new features to be explicitly calculated and defined. Support Vector Regression (SVR) is an extension of SVMs to regression problems, modifying the algorithm to ensure that all data points fall inside of the margin rather than outside of it as in classification. In regression, this margin is known as the error insensitive zone, as no error or loss is incurred provided the data points fall in this region.

The general framework of SVR is briefly outlined to contextualise their modification for survival data. SVR uses a linear combination of weights and covariates to perform regression. The covariates may be those of the original data vector x, but are more commonly based on transformations to higher dimensional spaces, $\varphi(x)$. The regression estimates for an SVR are given as:

$$\hat{y} = w'\varphi(x) + b$$

where b is a constant. The associated optimisation problem is

$$\min \ \frac{1}{2}w'w + C\sum_{i=1}^{n}(\varepsilon_i + \varepsilon_i^*)$$

$$\text{Subject to} \begin{cases} y_i - w'\varphi(x_i) - b \ \leq \ e + \varepsilon_i \\ w'\varphi(x_i) + b - y_i \leq e + \varepsilon_i^* \\ \varepsilon_{\mathrm{i}}, \varepsilon_i^* \ \geq 0 \end{cases}$$

where e is the error insensitive zone, C is the cost parameter applied to estimates for which $|\hat{y} - y_i| > e$, the ε are the penalised slack variables for over or underestimation and w is the vector of weights of the model. Slack variables are introduced to allow for solutions where some observations do not fall within the error insensitive zone, but incur a penalty as seen in the above optimisation problem. For reasons which will become evident later, special consideration should be given to the role of e in determining the error contribution of observations.

A second type of SVM considered in this section are ranking SVMs, which can be used for censored data encountered in survival problems. The estimates from a ranking SVM are given by the equation form as for SVRs, but the optimisation problem differs slightly:

$$\min \ \frac{1}{2}w'w + C\sum_{i,j;W_{ij}=1} W_{ij}\varepsilon_{ij}$$

$$W_{ij} = \mathbf{1}(y_i < y_j)$$

$$\text{Subject to} \begin{cases} w'\left(\varphi(x_j) - \varphi(x_i)\right) \leq 1 - \varepsilon_{ij} \\ \varepsilon_{ij} \geq 0 \end{cases}$$

A pair of points i, j are comparable in the above formulation if $y_i < y_j$, and errors are only incurred for incorrect ranking of comparable pairs. It is through modification of the W_{ij} indicator variable that SVMs under ranking constraints are extended to censored data, discussed further below.

Application of SVMs to survival analysis have been motivated by their excellent performance in classification and regression problems, as well as their applicability to high-dimensional data in which the number of features is large relative to the number of

observations. As with other techniques, however, the traditional SVM formulations do not account for the problem of censored data inherent to survival analysis and thus require modification. Several authors have proposed methods to utilise SVMs for survival analysis, which can generally be categorised according to whether they treat the survival problem as a regression or ranking task.

2.2.3.1 SVMs in Survival Analysis – Regression Constraints

Regression approaches are the most common formulation in which SVMs have been applied to survival analysis problems. Several such SVR formulations have been proposed.

An IPCW approach has been employed by several authors to apply variations of SVR with little modification to the original algorithm (Eleuteri & Taktak, 2012; Goldberg & Kosorok, 2017; Kim & Jeong, 2006; Shim & Hwang, 2009). Application of IPCW in this context is almost identical to the case of survival trees. Defining the censoring survival function as G, which may be non-parametric or may consider covariate effects, an application of IPCW to SVR results in the following optimisation problem:

$$\min \quad \frac{1}{2}w'w + C\sum_{i=1}^{n}\frac{\delta_i}{G(y_i|x_i)}(\varepsilon_i + \varepsilon_i^*)$$

$$\text{Subject to} \begin{cases} y_i - w'\varphi(x_i) - b \;\leq\; e + \varepsilon_i \\ w'\varphi(x_i) + b - y_i \leq e + \varepsilon_i^* \\ \varepsilon_i, \varepsilon_i^* \;\geq 0 \end{cases}$$

Note that no change was required other than the weighting of errors. While the IPCW approach requires estimation of the censoring survival function in addition to the SVR model, this is not seen as a prohibitive limitation, as the underlying censoring mechanisms are generally relatively simple (Goldberg & Kosorok, 2017). Estimation of the censoring survival function in SVR applications under IPCW is often done using the Kaplan-Meier estimator or a Cox proportional hazards model, and it is only in this censoring survival function that censored observations contribute to the loss function. The flexibility of the IPCW approach has been shown through its application to several variations of SVRs, including least squares, iteratively reweighted least squares, and quantile regression variations (Kim & Jeong, 2006; Shim & Hwang, 2009). Beyond these specific examples, the IPCW approach has also been discussed for SVMs in general under any appropriate loss function (e.g. classification, regression, median, etc.) (Goldberg & Kosorok, 2017).

A second approach to adapting SVR for censored data was presented by Shivaswamy, Chu, and Jansche (2007), termed "SVCR", with the definition of the error-insensitive zone being modified for censored data. The authors started from the scenario of interval censoring, in which the event is known to have occurred between two follow-up times. This is assumed to be a common problem in survival studies when events are reported with a delay. In this context, all censored observations are known to survive to some subject-specific minimum time, denoted as L, and have the event before a maximum time U, where $L < U$. This interval is set as the observation-specific error-insensitive zone in the SVR problem formulation, thus not penalizing predictions which fall into the relevant interval. This approach can then be generalized to all observations, regardless of whether censoring is present or what type of censoring is relevant. For uncensored observations, the error-insensitive zone is of width zero as the exact event time is known. For right censored observations, the minimum event time is the time at which the subject leaves the study, and the maximum event time is set to be infinite. Accordingly, no penalty is incurred for any prediction of event time after the occurrence of right-censoring. The case of left censoring follows the same logic. The associated optimisation problem is shown formulaically below.

$$\min \quad \frac{1}{2}w'w + C\sum_{i=1}^{n}(\varepsilon_i + \varepsilon_i^*)$$

$$\text{Subject to} \begin{cases} w'\varphi(x_i) + b - u_i \leq \varepsilon_i \\ l_i - w'\varphi(x_i) - b \ \leq \ \varepsilon_i^* \\ \varepsilon_{\mathrm{i}}, \varepsilon_i^* \ \geq 0 \end{cases}$$

Another approach focusing on customizing the error-insensitive zone was presented by Khan and Bayer-Zubek (2008). An asymmetric loss function is used instead of imposing no penalty on predictions falling anywhere in the region where the event is known to have occurred. This approach involves a greater number of hyperparameters but also affords greater control over the error contribution of censored observations. The error-insensitive zone parameters are modified to afford more leeway for predictions in the direction of censoring (i.e., higher upper limit in case of right censoring) without necessarily imposing no penalty. The regularisation parameters are similarly modified. Assuming some small degree of left censoring present in the reported events as a result of a delay in reporting, as well as the presence of right censoring, four combinations of the regularization and error bounds are defined. The four combinations for only two types of censoring are due

to the asymmetric nature of the loss functions being constructed. The optimisation problem is shown below.

$$\min \ \frac{1}{2}w'w + \sum_{i=1}^{n}(C_i\varepsilon_i + C_i^*\varepsilon_i^*)$$

$$C_i^* = (1-\delta_i)C_c^* + \delta_i C_n^* \text{ and } C_i = (1-\delta_i)C_c + \delta_i C_n$$

$$e_i^* = (1-\delta_i)e_c^* + \delta_i e_n^* \text{ and } e_i = (1-\delta_i)e_c + \delta_i e_n$$

$$\text{Subject to } \begin{cases} y_i - w'\varphi(x_i) - b \ \leq \ e_i + \varepsilon_i \\ w'\varphi(x_i) + b - y_i \leq e_i^* + \varepsilon_i^* \\ \varepsilon_i, \varepsilon_i^* \ \geq 0 \end{cases}$$

Recommendations are also made for the parameter relationships.

$$C_c^* < C_n < C_n^* = C_c \text{ and } e_c^* < e_n < e_n^* = e_c$$

For right censored data, the error insensitive zone has a larger upper bound and a smaller penalty is imposed on predictions beyond this upper bound. The reverse is true for the case of left censoring, with a more lenient lower bound on the error insensitive zone and regularisation parameter, though not to the same degree as in the case of right censoring. The smaller allowances for left censoring are motivated by the expectation that a reporting delay is less likely to cause a large difference between follow-up and event time compared to causes of right censoring.

2.2.3.2 SVMs in Survival Analysis – Ranking Constraints

Ranking approaches have also been considered for the application of SVMs to survival data. These approaches focus not on predicting the time of event occurrence, but instead on correctly predicting the relative order of the event time for each observation. This results in an index as an output, with higher values indicating longer times until event occurrence, but not the exact time. Model construction in the context of a ranking problem is equivalent to optimising the associated concordance index. This approach to the problem allows for relatively straightforward inclusion of censored data points by considering comparable pairs. In survival problems, any pair of data points i and j are considered comparable if event times are observed for both data points or if the event occurred for one of the data points before the other data point was subject to censoring. If two points are both censored, or the event occurred for one data point after the censoring of the other data point, then the pair is said to be incomparable. Model optimisation aims

to correctly rank all comparable points. This type of approach was used by Evers and Messow (2008), who noted that the ranking formulation aims to define a hyperplane separating observations that have and have not experienced the event at every event time. This hyperplane is translated rather than redefined at each event time, resulting in an implicit proportional hazards assumption.

The ranking formulation of Van Belle, Pelckmans, Suykens, and Van Huffel (2007) aimed to correctly rank each data point with respect to all other comparable points, and the associated formulation is shown below:

$$\min \ \frac{1}{2}w'w + C \sum_{i,j;W_{ij}=1} W_{ij}\varepsilon_{ij}$$

$$W_{ij} = \mathbf{1}(y_i < y_j \ \& \ \delta_i = 1)$$

$$\text{Subject to} \begin{cases} w'\left(\varphi(x_j) - \varphi(x_i)\right) \leq 1 - \varepsilon_{ij} \\ \varepsilon_{ij} \geq 0 \end{cases}$$

The redefining of W_{ij} results in a natural adaptation of the ranking SVM to survival data. Errors are incurred in this model only when two comparable points are incorrectly ranked. The intercept term in this formulation is omitted, as it does not alter the relative ranking of observations. Solving the resulting minimization problem is, however, highly computationally intensive given the number of comparisons made for every point. A modification was proposed by Van Belle, Pelckmans, Suykens, and Van Huffel (2008), in which each point is compared with only its K nearest observations with smaller event times. The modified model was shown to train approximately ten times faster, though point estimate performance was slightly lower compared to the original model (Van Belle et al., 2008).

2.2.3.3 SVMs in Survival Analysis – Regression and Ranking Constraints

In addition to applying SVMs to survival data under regression or ranking constraints, several authors have considered the case of SVMs in which both constraints are used (Van Belle, Pelckmans, Suykens, & Van Huffel, 2010; Van Belle, Pelckmans, Van Huffel, & Suykens, 2011). Combining the two approaches is relatively simple, with two error terms being used in the optimisation problem (regression and ranking errors) rather than one. Using five clinical and 3 high-dimensional data sets with a wide range of censoring levels, Van Belle et al. (2011) compared SVMs under regression constraints, ranking constraints,

and a combination of the two. Models using ranking constraints either alone or in addition to regression constraints did not perform as well as models using only regression constraints. This was despite a theoretical preference for models using ranking constraints, which can be tied to existing statistical models.

2.2.3.4 Related Research

Related research in this area has considered advanced SVM methods such as Learning Using Privileged Information (LUPI) and learning using uncertain labels (Shiao & Cherkassky, 2013). These methods were applied by treating the survival data in a classification context, however.

Several authors have also utilised the kernel trick prominent in SVMs in other models. Li and Luan (2003) used inner product kernels in the context of a Cox regression model, which allows for estimation of coefficients in scenarios where the number of features p is greater than the number of observations n. This approach does not reduce the number of features meaning that the solution is dense in both the observation and feature space. Evers and Messow (2008) proposed a potential solution to this issue through Import Vector Machines, a method analogous to forward stepwise techniques but over the observation space rather than feature space.

2.2.3.5 Summary of SVM Literature

Research extending SVMs to survival data problems have largely done so through adjustments to the optimisation problem solved in estimating ranking or regression models. They have not, however, modified the form of the estimates provided by the final model. Ranking approaches were applied through the definition of comparable pairs, with predictions serving as a risk index. Regression approaches were applied using IPCW or modifying the error-insensitive zone, with predictions representing estimates of event timing. No research was identified under either approach or the combination of ranking and regression approaches that provided survival curves with probabilistic estimates of event occurrence across time.

2.2.4 Artificial Neural Networks

One of the most popular modern machine learning techniques is the artificial neural network (ANN), which is loosely based on the biological nervous system and neurons. ANN models have been successfully applied in many contexts, and there is a substantial body of literature suggesting that ANNs can match or exceed the performance of many

commonly used statistical methods, particularly on complex problems. A particular advantage they have over commonly employed models is that they are able to automatically capture interactions and non-linear relationships.

Neural networks take a sequence of input features in the form of an input layer, which are then passed to a sequence of neurons in the first hidden layer. Each neuron in this first hidden layer takes a linear combination of weights and the input features and then applies a transformation known as the activation function to the result. The neurons in the first hidden layers then serve as the collection of inputs to the second hidden layer where this process is repeated. This process can be repeated for several hidden layers before an output layer is eventually reached. An appropriate cost function is then used to assess agreement of network outputs with the true outcomes. Weights in the network are iteratively updated to optimise performance on the specified cost function, a process referred to as backpropagation. This type of model has been shown to be effective for problems in which the relationship between inputs is highly complex and non-linear. They also offer a great degree of flexibility in their architecture to accommodate different objectives. To achieve regression or binary classification, the output layer may contain a single output neuron with the model's prediction, with model training involving iterative weight adjustments to minimise some appropriate loss function. For other problems such as multinomial classification, more than a single number output is required and so the output layer may consist of multiple neurons. For a more comprehensive overview of ANNs outside of survival analysis problems, readers are directed to Negnevitsky (2005).

While regression and classification ANNs cannot be directly applied to survival problems because of the partial data available as a result of censoring, they are very flexible in terms of construction. This has resulted in a range of different implementations for survival data. Laurentiis and Ravdin (1994) considered ANNs to fall into one of three categories: time-coded, single time point, and multiple time point ANNs. The same categorisation has been used in subsequent research. Additionally, hybrid ANNs integrating with established statistical techniques have also been proposed and have gained relative popularity in recent years.

The following review focuses on the differences in ideas underlying different approaches to applying ANNs to survival data and accordingly de-emphasises network design (e.g., number of neurons, training algorithm, etc.).

2.2.4.1 Time-Coded ANNs

In time-coded models, as originally put forward by Ravdin and Clark (1992), predicted outcomes are available at multiple defined times. This is achieved through considering discrete time intervals covering the entire period of the study, with time being an additional input to a network with a single output neuron. Time is included as a discrete input variable rather than a continuous one, representing the start of each interval. The outcome of interest is the event status (coded as 1 or 0) at the end of the interval, and the original data vectors are replicated for every interval. That is, event status is 0 before the event occurred, and is 1 for all other intervals. Censored observations are only included up until their time of censoring, however. As an illustrative example, consider intervals $(1, 2, 3)$ with end points of (2.01,4.01,6.01). The example data of Table 2 is converted to the format of Table 3 by duplicating feature vectors (covariates) for each observation for all intervals or until censoring occurs. The target variable in the transformed data indicates whether the event occurred in the current interval or any previous intervals.

Table 2. Example Survival Data

Observation ID	Covariates				Time	Event
1	…	…	…	…	6	1
2	…	….	…	…	3	1
3	…	…	…	…	2	0

Table 3. Example Survival Data - Time-Coded Format V1

Observation ID	Covariates				Interval	Target
1	…	…	…	…	1	0
1	…	….	…	…	2	0
1	…	…	…	…	3	1
2	…	…	…	…	1	0
2	…	….	…	…	2	1
2	…	….	…	…	3	1
3	…	…	…	…	1	0

One drawback with this approach is that it leads to a bias in later intervals towards observations in which events were observed. This is because observations where the event

is observed are persistent (duplicated for all subsequent intervals) in contrast to observations which are censored, which are only included until the censoring time. Later intervals will thus have an over-representation of observations with the event observed. Ravdin and Clark (1992) used a time-coded ANN to model the survival of breast cancer patients and addressed the bias towards events in later intervals by randomly deleting event vectors until the prevalence of event and non-event vectors in each interval were balanced. Another method of addressing the bias is simply to predict conditional hazard rather than survival probability, allowing for patient vectors to be replicated until the event-time only, after which they are no longer included in modelling (Franco, Jerez, & Alba, 2005). Such a model, with weight decay regularization, was compared with Cox regression by Franco et al. (2005) and was shown to have statistically significantly improved performance on a clinical dataset. Using the same raw data and intervals as above, the transformed data when predicting conditional hazard is shown for illustrative purposes in Table 4. Note that the second observation is only duplicated for the first two intervals despite the event being observed, in contrast to Table 3.

Table 4. Example Survival Data - Time-Coded Format V2

Observation ID	Covariates				Interval	Target
1	…	…	…	…	1	0
1	…	….	…	…	2	0
1	…	…	…	…	3	1
2	…	…	…	…	1	0
2	…	….	…	…	2	1
3	…	…	…	…	1	0

Another variation on the time-coded ANN was considered by Eleuteri, Tagliaferri, Milano, De Placido, and De Laurentiis (2003), who incorporated Bayesian elements and placed constraints on the weights associated with the time input. By imposing restrictions on the sign of these weights to ensure they are positive, it is ensured that the resulting survival function is monotonic with respect to time.

While time-coded models consider time in discrete intervals rather than as a continuous variable, an approximation of survival curves can be achieved by combining the survival probabilities or conditional hazards output by the model at the endpoint of each interval.

While this is a more appropriate treatment of time than not including it in any form, there may still be some bias in the model resulting from the treatment of time as a discrete variable. If censored observations are considered to survive to the end of the interval in which they are censored, then the earlier they are censored in the interval the greater the bias introduced. Likewise, if they are not considered in the interval in which they are censored, the later they are censored in the interval the greater the bias introduced. This bias can be mitigated, though not eliminated, by using reasonably narrow time intervals.

2.2.4.2 Single Time Point ANNs

As with logistic regression and other binary classification approaches, single time point ANNs predict event status by a fixed follow-up time. Time is not an input to the model and is not accounted for in outputs. Examples include the models used by De Laurentiis, De Placido, Bianco, Clark, and Ravdin (1999) and Jerez-Aragonés, Gómez-Ruiz, Ramos-Jiménez, Muñoz-Pérez, and Alba-Conejo (2003), who both considered prognosis of breast cancer patients. Censored data is addressed as in typical classification models, either through removal (which introduces bias) or imputation via another technique (De Laurentiis et al., 1999). If the imputation technique is not entirely suitable, such as not being complex enough, then the performance of the final network will also suffer. Jerez-Aragonés et al. (2003) provided an example of excluding censored observations in constructed models, but at least partially included censored data by considering distinct single time point models for different time intervals. Unlike time-coded models, however, multiple distinct models were used without time as an explicit input. The limited ability of single time point ANNs to incorporate the information present in censored observations, coupled with considering only one time point, is partially overcome by multiple time point ANNs.

2.2.4.3 Multiple Time Point ANNs

Multiple time point ANNs also do not allow for time as an input variable, but instead make predictions for more than one time point using a single model. These models aim to predict a vector of outcomes for a single input instance, with the outcome vector representing survival status at several different time points. For example, a multiple time point model may predict event probability at the end of one, two, three, or four years for a given set of inputs. The probability of survival or death at multiple time points again allows for an approximation of a survival curve, albeit a crude approximation unless many time points are used. A variety of treatments is possible for censored observations past

their time of censoring in multiple time point ANNs. Values can be imputed with estimated survival status (Lapuerta, Azen, & Labree, 1995) or hazard rate (Baesens, Gestel, Stepanova, Poel, & Vanthienen, 2005). Alternatively, the loss function can be defined such that output neurons dealing with outcomes post-censoring do not contribute and thus do not influence the model training process (Brown, Branford, & Moran, 1997). This avoids the need for imputing with reasonable values, while still retaining the partial data for earlier time points.

The multiple time point approach to adapting ANNs for survival analysis has been widely adopted. They have been applied for breast cancer prognosis (Chi, Street, & Wolberg, 2007; Ripley, Harris, & Tarassenko, 1998; Street, 1998), coronary artery disease (Lapuerta et al., 1995), risk of loan default (Baesens et al., 2005), AIDS (Ohno-Machado, 1997), and metastatic cancer (Gensheimer & Narasimhan, 2019). While the flexibility of multiple time point ANNs have led to differences between the implementations of different authors in terms of activation functions, treatment of censored observations, network depth, or regularisation, the core idea has remained constant. They are also conceptually similar to time-coded ANNs, with time discretised into intervals and predictions generated for each of these intervals. Like time-coded ANNs, the discretisation of time introduces bias into the model depending on the width of the time intervals considered.

2.2.4.4 Hybrid ANNs

Another approach to adapting ANNs to survival analysis has been through integrating them with established semi-parametric and parametric statistical techniques. At their core, this has been done through replacing the linear predictor of statistical techniques with the output of an ANN.

2.2.4.4.1 ANNs with Semi-Parametric Survival Models

The first of these models was originally put forward by Faraggi and Simon (1995). In terms of network design, a simple three-layer feedforward network was used with an input layer for covariates, a hidden layer with a logistic activation function, and a single output neuron in the output layer. This model was not used to directly predict survival status, however, and instead was used to replace the linear combination of covariates and coefficients in the Cox proportional hazards model. To illustrate this, the partial likelihood of the Cox proportional hazards model is restated using previously defined notation:

$$PL(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} \right)^{\delta_i}$$

In this formulation, the model predictor is the linear combination of the coefficient vector β and the covariate vector x_i for the i-th observation. Next, let the output of a single neural network, $g(x_i, \theta)$, with H hidden neurons, no bias neurons, and a logistic activation function be expressed as

$$g(x_i, \theta) = \sum_{h=1}^{H} \frac{\alpha_h}{1 + \exp(-w_h' x_i)}$$

where α_h represents the weight applied to the output of the hidden neuron h, and w_h is the vector of weights applied to inputs to the hidden neuron h. The θ in $g(x_i, \theta)$ denotes the vectors of weights to be estimated for the hidden and output layers. These two model equations are linked in the approach proposed by Faraggi and Simon (1995), by replacing the linear predictor in the partial likelihood of the Cox model with the non-linear network function $g(x_i, \theta)$ to give:

$$PL(\theta) = \prod_{i=1}^{n} \left(\frac{\exp\big(g(x_i, \theta)\big)}{\sum_{j \in R(t_i)} \exp\Big(g\big(x_j, \theta\big)\Big)} \right)^{\delta_i}$$

$$= \prod_{i=1}^{n} \left(\frac{exp\left(\sum_{h=1}^{H} \frac{\alpha_h}{1 + \exp(-w_h' x_i)}\right)}{\sum_{j \in R(t_i)} exp\left(\sum_{h=1}^{H} \frac{\alpha_h}{1 + \exp(-w_h' x_j)}\right)} \right)^{\delta_i}$$

The maximum likelihood estimates can then be solved through iterative methods, with the popular Newton-Raphson method being used by Faraggi and Simon (1995).

This approach to implementing a neural network for survival analysis is useful because of its links to existing statistical techniques, both Cox regression and others. The linear or non-linear predictor functions of these other models need simply be replaced with the network predictor function. This approach also leverages the ability of ANNs to identify highly non-linear relationships without requiring prior specification.

The approach has been widely used and built upon by later research. Faraggi, Simon, Yaskil, and Kramar (1997) extended this approach to include Bayesian regularisation, aiming to reduce the issue of overfitting common to neural network models and, to a

lesser degree, maximum likelihood estimation procedures. Mariani et al. (1997) compared the proposed technique to a linear Cox regression on breast cancer patients, finding similar performance between the two models. More recently, Katzman et al. (2018) and Ching, Zhu, and Garmire (2018) employed the technique proposed by Faraggi and Simon (1995), but using more modern network designs. Katzman et al. (2018) used deep learning, a self-normalising activation function, and optimised a range of hyperparameters used in the network training process. The resulting model was compared to a traditional Cox proportional hazard model and a random survival forest on a diverse range of datasets, with variation in patient type and regions. The new model, referred to as "DeepSurv", outperformed the Cox proportional hazard model and was generally at least as good as the random survival forest. Ching et al. (2018) applied the technique to high-dimensional genomics data, once again using a more sophisticated model construction process, though in this case the regularisation aspects were the key feature. In this study, the proposed "Cox-nnet" model was at least as good as any other considered model, which included a variety of Cox proportional hazard models and a random survival forest. These results were achieved through consideration of 10 different cancer datasets.

2.2.4.4.2 ANNs with Parametric Survival Models

The second type of hybrid ANN is a special case of a time-coded ANN. This hybrid ANN was originally put forward by Biganzoli, Boracchi, Mariani, and Marubini (1998), and takes a similar approach to that of Faraggi and Simon (1995). Instead of integrating ANNs with Cox regression they are integrated with parametric generalised linear models. While the method is general in nature and can be applied in conjunction with various piecewise parametric models (Biganzoli, Boracchi, & Marubini, 2002), the partial logistic regression model for grouped time data is most common. To motivate the partial logistic ANN, or PLANN, consider the situation of τ discrete intervals of time with discrete conditional hazard rates $h_l(x_i)$ for each observation i and interval l. If we also consider δ_{il} to be the event indicator for observation i in interval l, the total likelihood is expressed as

$$L = \prod_{l=1}^{\tau} \prod_{i \in R_l} h_l(x_i)^{\delta_{il}} (1 - h_l(x_i))^{1-\delta_{il}}$$

where R_l is the set of observations at risk in interval l. From this, the proportional odds model for grouped survival times is given by the following:

$$\frac{h_l(x_i)}{1-h_l(x_i)} = \frac{h_l(0)}{1-h_l(0)} \exp\left(\beta' x_i\right)$$

This can be re-expressed as

$$h_l(x_i) = \frac{\exp(\theta_l + \beta' x_i)}{1+\exp(\theta_l + \beta' x_i)}, \theta_l = \log\left(\frac{h_l(0)}{1-h_l(0)}\right)$$

which is the partial logistic regression model for grouped time data, predicting interval-specific hazard rates based on a predictor which is a linear combination of covariates and coefficients. Biganzoli et al. (1998) noted that this model structure can be replicated by a specially designed ANN using a logistic activation function with no hidden neurons. Minimising the commonly used cross-entropy cost function is then equivalent to minimising the negative log likelihood of the partial logistic regression. This link underlies the subsequently proposed PLANN, which replicates the partial logistic regression model with an ANN and then adds a hidden layer. By adding a hidden layer, the linear predictor is replaced with a non-linear one, giving the following estimator of interval-specific hazard:

$$h_l(x_i) = \frac{exp\left(\alpha_0 + \sum_{h=1}^{H} \alpha_h \frac{\exp\left(w_0 + w_h' x_i^l\right)}{1+\exp\left(w_0 + w_h' x_i^l\right)}\right)}{1 + exp\left(\alpha_0 + \sum_{h=1}^{H} \alpha_h \frac{\exp\left(w_0 + w_h' x_i^l\right)}{1+\exp\left(w_0 + w_h' x_i^l\right)}\right)}$$

Bias neurons in the hidden layer and the output layer are included in this formulation and are denoted as w_0 and α_0 respectively. x_i^l represents the usual covariate vector for the i-th observations as well as the time interval, making this a special case of a time-coded ANN. This is the final formulation of the PLANN model, which is trained using cross-entropy error. A description of this approach for other GLMs is described by Biganzoli et al. (2002). The PLANN model has since been used for cancer prognosis (Biganzoli, Boracchi, Coradini, Grazia Daidone, & Marubini, 2003; Taktak et al., 2009) and in conjunction with rule extraction techniques (Lisboa et al., 2008). It has also been extended to the case of competing risks (PLANN-CR) (Biganzoli, Boracchi, Ambrogi, & Marubini, 2006; Boracchi, Biganzoli, & Marubini, 2001) and to include a Bayesian regularisation method known as Automatic Relevance Determination (PLANN-ARD) (Lisboa, Wong,

Harris, & Swindell, 2003). These developments have not been exclusive of one another, with Lisboa et al. (2009) combining PLANN-CR and PLANN-ARD to develop PLANN-CR-ARD.

2.2.4.5 Summary of ANN Literature

Driven by the flexibility of ANNs and their encouraging performance in other settings, a range of extensions has been proposed for survival data. These extensions have included time-coded and multiple time-point ANNs, which divide time into discrete intervals, as well as hybrid approaches integrating ANNs with statistical techniques, such as Cox or logistic regression. Of these, time-coded and multiple time point ANNs incur bias through the discretisation of time, but this is mitigated through considering narrower intervals. There is also overlap between hybrid and time-coded approaches, with the relatively popular partial logistic ANN technique having a direct link to logistic regression but also considering time in discrete intervals. There has been evidence that these ANN extensions to survival data can improve on the most common statistical survival technique of Cox regression in some settings.

2.2.5 Overall Summary of Survival Techniques

The challenge posed by censored survival data has given rise to multiple streams of research across both statistical and machine learning fields, resulting in a variety of distinct techniques. Semi-parametric methods have been afforded the most research in statistical contexts, where the challenge is to avoid making distributional assumptions beyond functional forms while still capturing covariate effects. For decision trees and related ensembles, the focus has been on selecting splitting rules which account for both censored and uncensored observations. New approaches to splitting rules have either been with rules specific to censored data contexts or with transformations which maintain links to the measures that would have been used if data were not censored. Different ensembles have used distinct variations of survival trees, but the primary modifications made to censored data have been in terms of these underlying splitting rules. Support vector machines have been adapted to censored data either by taking a ranking approach over comparable observation pairs or by taking a regression approach with creative definition of error-insensitive zones and penalties for censored observations. Unlike trees and statistical methods, SVMs do not provide survival or hazard functions for the considered data, but instead directly estimate event times (in the case of regression) or provide a risk index (in the case of ranking). Finally, the flexibility of ANN construction in terms of

data structure and model architecture has led to a variety of different extensions for survival data. Time-coded ANNs take time intervals as inputs and output either cumulative event or conditional hazard probabilities for those intervals. Multiple time point models use multiple output nodes to similarly predict survival or hazard across discrete time intervals. Hybrid ANNs provide a link back to statistical models, semi-parametric or otherwise, by replacing the linear predictor of statistical techniques with an ANN model.

Cox's proportional hazards model remains the most used benchmark against which novel survival techniques are compared. Such comparisons have shown that machine learning techniques are useful for many problems. However, different categories of machine learning survival models have not been extensively compared to one another. In the case of neural networks, even the different approaches to survival data within this category have rarely been compared. This is in contrast on the one hand to SVMs, for which there is some evidence that the regression rather than ranking approach is more successful (Van Belle et al., 2011), and decision trees, on the other, where intra-category comparisons are more common (Steingrimsson et al., 2019; R. Zhu & Kosorok, 2012).

3 Research Questions

For both research questions described below, the study cohort considered will be hospital admissions of adult patients presenting to the Emergency Department. This setting is characterised by a need for dynamic decision-making as well as diversity in the reasons for presentation. Such characteristics make tools improving risk assessment and management valuable.

3.1 Research Question 1

In hospital readmission research, the development and proposal of new models continue. This has been driven by heterogeneous study contexts, poor performance, and increasing policy focus on this area. These drivers have also led to the increased consideration of machine learning techniques as an avenue for potentially improving model performance. This research trend has, however, been limited to classification approaches. Studies that have taken survival approaches have not seen the same diversity of techniques employed, despite the motivation for consideration of machine learning techniques under classification approaches also being applicable to survival approaches. That is, the ability of machine learning techniques to account for complex and non-linear relationships between covariates and outcomes is potentially valuable under both survival and classification approaches. While many studies have taken classification approaches and compared machine learning and statistical techniques, only two studies were identified that considered machine learning survival techniques (Hao et al., 2015; Padhukasahasram et al., 2015) and comparative performance was not a focus in either. The first research question is thus motivated by the fact that only one type of machine learning survival technique was identified in the literature. There is a lack of studies investigating the potential for machine learning techniques to improve on statistical survival techniques for readmission prediction.

RQ1: Can machine learning survival techniques improve upon statistical survival techniques when predicting 30-day hospital readmissions?

To answer this question, a range of models will be considered for predicting 30-day readmissions. The benchmark survival modelling technique of Cox regression will be considered along with survival techniques from major machine learning categories. The literature review of Chapter 2 motivates the modelling techniques considered, outlined in

more detail in Section 5.1. These include survival trees under different splitting rules, doubly robust CURTs, doubly robust CUREs, random survival forests, recursively imputed survival trees, Bayesian additive regression trees, and three variations of neural networks for survival data. Of these machine learning techniques, prior research has considered only random survival forests, making the range of techniques considered a key contribution of this study. Lastly, the benchmark classification technique of logistic regression will also be considered.

Currently, evaluation of readmissions models is typically based on performance at a single time point. Accordingly, all models will be evaluated based on 30-day readmission prediction. While survival models provide probabilities associated with more than one time point, readmission models have not been used in the more flexible contexts allowed by survival approaches. The gap addressed by this research question is the lack of consideration of these techniques rather than how they are applied and assessed. In evaluating 30-day readmission prediction, the potential value of the newly considered techniques is established under current research methods.

3.2 Research Question 2

The simplification of what is a time-to-event problem to a binary regression problem has been identified as a limitation in previous research with a corresponding information loss (Futoma et al., 2015). Two limitations of classification models under a binary problem formulation are relevant in comparison with survival models. First, when using a fixed time point measure as in 30-day readmission prediction, it is implicitly assumed that all patients face an elevated risk of readmission related to their discharge for the same length of time. If this assumption is violated, then the predicted risk of patient readmission related to index admission will be under- or over-estimated. The second limitation is that the prediction from a classification model is usable only at the time of discharge. Though the potential for more flexible measures from survival models has been noted previously (Yu et al., 2015) no research was identified that assessed such alternatives. The only survival-specific application suggested in the identified literature was a daily ranking of 30-day readmission risk for discharged patients (Hao et al., 2015), but models were assessed in this study only based on the standard 30-day readmission risk as at the time of discharge. The only survival-specific measure of performance identified in the

reviewed literature was the c-index, which is analogous to AUC for binary classification models (Padhukasahasram et al., 2015). This was not, however, accompanied by consideration of any survival-specific model applications. The absence of any research that considered both model applications and assessment reflecting survival approaches provides the motivation for the second research question:

RQ2: How well can various survival modelling techniques capture aspects of hospital readmission risk over time relevant to managerial decision-making?

To answer this research question, the same survival techniques will be considered as in the first research question. The additional elements of answering this question are in identifying the aspects relevant to managerial decision-making and associated performance measures. This is done through proposing several novel applications and discussing the relevant aspects of model performance within those applications. Discussed in greater depth in Section 5.2, the proposed applications are the following:

- **Dynamic Risk Ranking (DRR)** – This application considers the probability of readmission for all at-risk patients (such as those discharged recently) for a specified time-period and ranks them. It is a dynamic ranking in that it can be updated daily and is not limited in application to patients only at the time of discharge.
- **Elevated Risk Period (ERP)** – This application assesses the length of time before a patient's risk of readmission reaches some acceptable level. This provides insight into the length of time that a patient faces elevated risk related to their discharge and thus how long they are of interest to the institution for post-discharge management decisions. This application also allows for customisation in the definition of an acceptable risk level, driven by the specific context and goals of the institution, though this is expected to be done on a relative basis.
- **Elevated Risk Period Probability (ERPP)** – As implied by the name, this proposed application is closely linked to ERP. The ERPP represents the probability of a patient's being readmitted within a period after discharge, as is typically done using classification approaches. Unlike previous approaches, however, the period considered can vary between patients based on their determined ERP. This addresses one of the core limitations of fixed-time period

approaches to predicting readmissions in that ERPP does not assume homogeneous evolution of risk over time.

- **Expected Readmissions (ER)** – This fourth application aims to assist with the forecasting of readmission numbers to facilitate efficient allocation of resources to service demand. Using survival curve outputs from a suitable model, the probability of each at-risk patient being readmitted within a period can be calculated conditionally on their being readmission-free between their discharge and the beginning of the period of interest. An expected number of readmissions in the period can then be found through the summation of these patient-specific probabilities.

These proposed applications enable managerial decision-making to address needs from both institutional and patient perspectives. The first application enables prioritising of patients to maximise the value of interventions or follow-up care. The second measure defines patient-specific periods over which risk of admission is pronounced, and the third measure computes the risk within this period. Finally, the fourth measure enables forecasting of future readmissions to better inform resource and staff planning decisions. Of these, only the first has been previously suggested by Hao et al. (2015) but using only a 30-day period and without related performance assessment. Based on these applications, relevant aspects of model performance can be identified to then inform selection of appropriate performance measures for model evaluation.

This research question aims to address the limited consideration of model applications specific to survival approaches and associated measures for model assessment in prior readmission research.

3.3 Contributions

Several contributions are made through addressing the two research questions. Given the inherently practical nature of readmission modelling, contributions are to both practice and the research literature.

For both research questions, a range of machine learning survival models not previously considered in the readmission literature are identified and empirically assessed. Considering the first research question, this empirical assessment is based on current application of models predicting readmission risk and is relative to the current survival

benchmark, Cox regression, and the current classification benchmark, logistic regression. This is expected to inform future model usage though highlighting available techniques as well as providing empirical evidence as to how currently used techniques compare with alternatives. Additionally, the comparison of survival models with logistic regression provides an indication of the trade-off involved in having predictions across time but potentially poorer single time point prediction.

Considering the second research question, the empirical comparison encompasses several additional contributions, as the basis for model assessment is less established. First, several novel applications of survival models in readmission modelling are proposed to support managerial decision-making. These represent at least part of the unexplored potential value of survival models in readmission modelling. Such applications are expected to be explored further in future research and should be of practical value to hospitals in better managing discharged patients. Secondly, the relevant aspects of performance for survival models used in such applications are identified as well as measures capturing these aspects of performance. Only one survival-specific measure of performance was noted in the existing literature and it was not linked to an associated practical application of a survival model. This contribution addresses the research gap and provides guidance to practitioners for evaluating survival models for practical applications. Lastly, leveraging the identification of important elements of model performance and appropriate measures, the range of machine learning survival models and the statistical survival benchmark are empirically assessed based on the identified measures.

For both research questions, model assessment and comparisons are expected to inform consideration of modelling techniques in future research and practice, albeit in different applications. Additionally, both research questions provide a comparison of machine learning and statistical models. Due to a lack of comparability between studies, it has not yet been definitively established whether machine learning techniques can be expected to consistently improve on statistical techniques and, if so, in what circumstances. The model comparisons of both research questions thus add to the available evidence around whether and when machine learning techniques can improve on statistical techniques. This contribution is strengthened by use of a wide range of techniques and separate consideration of two hospitals.

Two secondary contributions, though not the primary focus of this work, will also be made through addressing both research questions. First, study heterogeneity in the hospital readmission literature has made generalisation of results difficult. The emphasis on the USA context has exacerbated this problem for generalising findings across regions with distinct healthcare characteristics, such as Australia. Answering both research questions will add to the existing literature on Australian readmission models, which is particularly important given recent policy proposing pricing models to incentivise institutions to reduce readmissions (Independent Hospital Pricing Authority, 2021a). Secondly, in the context of the survival modelling literature, this project will address on an empirical basis the current lack of research extensively comparing different categories of machine learning survival models. These models have typically been motivated by the ability of the relevant machine learning techniques to capture complex and non-linear relationships (Biganzoli et al., 2002; Ishwaran et al., 2008; Laurentiis & Ravdin, 1994; Shimokawa et al., 2015), but have primarily been compared to Cox regression or to a limited set of alternatives. As these models share a common motivation for their development, their performance relative to one another is of interest for informing model-choice decisions in other contexts. Using two hospitals, this project provides two such scenarios in which the relative performance of these models along several metrics will be evaluated.

3.4 Research Principles

Finally, the underlying principles guiding decisions made in the process of addressing these research questions should be specified. These relate to practical application and comparability and will be referred to in subsequent sections where relevant decisions are made.

Practical Application

For both research questions considered, the goals of this project are to assess the viability of models for use in practice. Thus, the methods employed in this project aim to reflect the practical application of developing, selecting, and applying readmission models.

Comparability

For both research questions considered, the performance is assessed of a range of models in predicting readmissions. As the goal is to compare models of different types, the methods should ensure that when models are compared, the differences can be attributed to differences in models rather than differences in model selection or evaluation. To promote comparability of research more generally, methods should ensure that other authors could replicate this work if given the same data. This precludes the use of subjective adjustments or other decisions in data processing and model selection.

4 Data

The data used in the project are outlined throughout this chapter. It is first described in terms of pertinent aspects such as its source, outcome definition, and period over which it was collected. Secondly, the processing applied to the data to prepare it for modelling is detailed. Finally, descriptive statistics are presented and differences between admission characteristics for the two hospitals considered are highlighted.

4.1 Data Description

In this project, hospital discharges are the observations of interest. A discharge relates to the event where a patient who has been admitted to the hospital is released. This project specifically considers the hospital discharges of adult patients who were admitted to hospital after presenting to the Emergency Department for either Gold Coast University Hospital or Robina Hospital. Both hospitals service the Gold Coast region. After exclusions (described in Section 4.2), the data provided consists of information available at the time of discharge for 70,635 observations in the period ranging from April 30th, 2016 to April 30th, 2018. The following subheadings describe the pertinent aspects of the data provided.

Data Source:

Data are provided by Gold Coast University Hospital (GCUH) and Robina Hospital (RH) and used with permission from Healthcare Logic Pty Ltd (HCL). These two hospitals have entered into agreements with HCL to make their Power Performance Manager (PPM) costing data available for the project. Both are major public teaching hospitals servicing the Gold Coast area. The patient populations serviced by each are quite different, however, on account of differences in specialisations and capacity. This is illustrated through descriptive statistics produced for discharged patients for each hospital in Section 4.3. Because of the differences in patient populations, readmissions were modelled for each hospital separately, consistent with the way institution-specific models would be constructed in practice and discussed further in Section 4.2. After exclusions, also described in Section 4.2, there were 70,635 discharges available for modelling, of which 46,659 corresponded to GCUH and 23,976 corresponded to RH.

Outcome Definition:

For this study, discharges for patients admitted to either GCUH or RH are considered. Readmissions were deemed to occur if:

- readmission type was Acute; and
- readmission status was Emergency.

Readmission type was determined to be acute if the initial stage of the patient's journey required acute care rather than Rehabilitation. Readmission status was determined as Emergency if the readmission as an inpatient occurred via the Emergency Department. These conditions aim to ensure that readmissions are only considered where they are unplanned and urgent. The goal of predicting readmission risk in this work is to avoid or measure these types of readmissions, rather than those which are routine, planned or otherwise low severity. For research question 1, the outcome of interest was whether such a readmission occurred within 30 days. This represents a binary variable. For research question 2, the outcome of interest is a continuous variable indicating time until readmission or censoring and a binary variable indicating whether readmission or censoring.

Data Period:

Discharges from both hospitals from April 30th, 2016 to April 30th, 2018 are included in the set of observation available for modelling. Several factors contribute to the data period considered. First, to construct covariates relating to a patient's past involvement with health services such as number of previous admissions, one year is used as a lookback period. Thus, while only the period 2016-2018 is usable in modelling, data from 2015 are also used implicitly. Secondly, a trade-off is made between the volume of data used in model training and its relevance for future patients. The wider the time frame considered, the greater the heterogeneity in medical practice and systems which are expected to develop and improve over time. This is undesirable as it inhibits the degree to which a model will accurately reflect readmission dynamics for the future patients to whom it will be applied. Thus, it was felt that the period used balanced the need for a sufficiently large dataset to train the considered models while remaining recent enough to be representative.

Data Fields:

Of the fields available, 19 were used in modelling (see Table 6 below). The consideration of these fields was motivated by the intended use of models considered in this research and to remain faithful to the principle of practical application. Several aspects of these decisions bear discussion.

First, the type of information captured by the fields are highlighted. The fields used in modelling primarily relate to prior use of health services, as well as sociodemographic factors including sex, age, and the region of the patient's home address. Regarding the fields capturing prior use of health services, it should be noted that this includes usage beyond the emergency department, such as number of inpatient admissions through any avenue in the last year. Length of stay for the index admission is also used, serving as a proxy for illness severity. These fields, provided and derived, were included because they were readily available at the time of patient discharge as well as representing potentially important predictors of readmission risk.

Second, the quality of the data corresponding to these fields is relevant. After the exclusions described in Section 4.2, both those performed by HCL prior to data provision and those performed as part of this work, no fields had any missing data. The construction of fields through the use of hospital costing data also ensures that the quality of data used in this study reflects the quality of data available to hospitals for developing predictive models. This equivalence is particularly important given this work's emphasis on informing the practical application of such models within hospital settings.

Third, several types of information have been considered in prior research that are not captured by the fields outlined in Table 6. The fields capture information about prior utilisation of health care services, demographics, and limited information about illness severity. No fields relate to medical comorbidity, mental health comorbidity, and overall health and function. Given that fields capturing such information would be expected to be influential, as has been found in prior research, models constructed in this work may incur some bias from their exclusion. Despite this, such fields are not included as they do not reflect the actual data available (at least in the data source used in this work) when the model would be applied in practice. Prominent examples of fields with predictive power unavailable at the time of discharge include the International Classification of Disease (ICD) codes and Diagnosis Related Groups (DRG) codes. Existing readmission modelling research has found ICD and DRG codes to contain valuable information for readmission prediction (Futoma et al., 2015). Unfortunately, these fields are often not

recorded until four to six weeks after discharge in the data sources used in this work and so are not appropriate for a model intended for application at discharge or shortly after discharge.

Lastly, while feature engineering involving other fields available at discharge could potentially add predictive information, the innovations of this work's research questions relate to the techniques used and the practical application of resulting models. The potential value of more extensive feature engineering is left as an avenue for future research.

Data Splitting:

For the purposes of separating model training and final evaluation, the data were split into a training and testing set containing 70% and 30% of the data, respectively. In keeping with the principle of model evaluation reflecting practical application, as stated in Section 3.4, a longitudinal split was used. In practice, models will be trained on past patients before application to new patients, with the longitudinal aspect being made more important by the relevance of temporal trends relating to medical practice and systems. This split and the relevant date ranges are shown in Table 5, with overlapping dates reflecting splits part-way through the day.

Table 5. Train and Test Data – Size and Dates

Hospital	Split	Quantity	Start Date	End Date
GCUH	Train	32,661	2016-04-30	2017-09-30
	Test	13,998	2017-09-30	2018-04-30
RH	Train	16,783	2016-04-30	2017-09-29
	Test	7,193	2017-09-29	2018-04-30

Table 6. Features Used in Modelling

Covariate Name	Covariate Description
AdmitWardCode1 (*Derived Field*)	An aggregated version of the AdmitWardCode field. This derived field is described in Section 4.2. AdmitWardCode: WardCode patient is admitted to.
Age	Age of a patient calculated from DOB to Discharge Date
ED_NumPresPrevYear	Number of ED presentations that occurred during the year prior to the current row's admission date
ED_NumPresSincePrevAdm	Number of ED presentations that occurred since the patient's previous inpatient admission via ED
ED_NumPresSincePrevAdmALL	Number of ED presentations that occurred since the patient's previous inpatient admission (via Outpatients, ED etc.)
GenderCode	Gender of a patient (M or F)
iGC (*Derived Field*)	A grouped version of the Postcode field specifying the region of the Gold Coast the patient's home address is in. This derived field is described in Section 4.2. Postcode: Postal Code of a patient's home address
Inpat_NumAdmPrevYearALL	The number of all inpatient admissions (via Outpatients, ED etc.) that occurred during the year prior to the current row's admission date
Inpat_PrevAdmLOSPrevYear	Length of stay of previous inpatient admission via ED in days
Inpat_PrevAdmLOSPrevYearALL	Length of stay of previous inpatient admission (via Outpatients, ED etc.) in days
Inpat_TimeSincePrevAdmALL	Days since the previous inpatient admission (via Outpatients, ED etc.) that occurred during the year prior to the current row's admission date
Inpat_TotalAdmInICU	Number of Inpatient Admissions that the patient had in the ICU within the previous year
Inpat_TotalAdmInICUALL	Number of Inpatient Admissions (via Outpatients, ED etc.) when the patient was in ICU within the previous year from the current row's admission date
Inpat_TotalTimeAdmPrevYear	Cumulative length of stay in days as an inpatient admission via ED during the year prior to the current row's admission date
Inpat_TotalTimeAdmPrevYearALL	Cumulative length of stay in days as an inpatient admission (via Outpatients, ED etc.)

Covariate Name	Covariate Description
	within hospital during the year prior to the current row's admission date
LOSCalc (*Derived Field*)	Difference in days between the date elements of AdmitDateTime and DischargeDateTime fields. AdmitDateTime: Timestamp of inpatient admission DischargeDate: Timestamp of inpatient discharge
Outp_NumApptPrevYear	Number of outpatient appointments that occurred during the year prior to the current row's admission date
Outp_NumApptSincePrevAdm	Number of outpatient appointments that occurred since the patient's previous inpatient admission via ED
Outp_NumApptSincePrevAdmALL	Number of outpatient appointments that occurred since the patient's previous inpatient admission (via Outpatients, ED etc.)

4.2 Data Processing

Before being provided for the project, data were extracted from databases of GCUH and RH by HCL as part of other business purposes. The data were provided in a form already conducive to modelling, with little additional processing required other than for certain model-specific requirements. The data-processing steps that were required before any type of model training was performed are described here, relating to exclusion criteria (carried out by HCL and the researcher) and feature manipulation for high-dimensional categorical variables.

Data Exclusions

The data originally extracted from the hospital databases included all patient discharges over the period from April 30th, 2016 to April 30th, 2018.

Before data were provided for this research project, several exclusions were made by HCL. Discharges were excluded if they related to admissions where:

- The index discharge destination was an episode of care change or hospital transfer
- The index discharge was for Chemotherapy or Dialysis DRGs
- The index discharge was against medical advice

- The admissions were classified as being either a “Routine readmission not requiring referral” or “Outpatient Appointment” and consisted of a day procedure (less than 1-day length of stay) discharged from a day unit
- The index discharges were from the ED Short Stay Unit (ED.SSU) or Clinical Decision Unit (CDU) and coded as an inpatient
- The Admission Source was an ‘Episode change’
- The Admission Source was ‘Non-admitted patient referred from another hospital”

For this project, a model predicting readmission risk is desired for patients admitted directly to the hospital for reasons which are not routine or otherwise simple and are associated with discharges to the community in line with medical advice. The purpose of these exclusions was to ensure that each discharge was associated with this type of admission. While the reasoning for most exclusions is self-evident, some bear additional explanation. First, admissions where discharge was for Chemotherapy or Dialysis DRGs were excluded because these patients tend to be frequently admitted to hospital for routine procedures, but in some cases may be classified as being in the Emergency Department for administrative reasons. Thus, they would be coded as a readmission but would actually be a routine visit. Secondly, discharges were excluded if the index discharge was from ED.SSU or CDU and coded as an inpatient. This was because patients moved to either of these locations in the ED are technically classified as inpatients in hospital systems but were not actually admitted to an inpatient ward outside of the ED.

The resulting dataset provided by HCL consisted of 130,743 discharges. Three exclusion criteria were then applied after its provision. Discharges were excluded if they related to admissions where:

- The same admission was already present in the data (130,716 discharges)
- The admission source was not “Accident and Emergency” (77,316 discharges)
- The age of the patient at admission was under 18 years (70,635 discharges)

The first exclusion criterion removed duplicated observations, of which there were few. The consideration of only admissions with admission source “Accident and Emergency” was motivated by a need to balance homogeneity of patients and size of the applicable population of constructed models. For homogeneity of patients, better model performance can be expected when the cohort of admissions do not contain several distinct populations. For example, readmission dynamics would be expected to be very different

for newborns as compared to patients admitted to the emergency department. Including multiple populations in the same data for modelling purposes would adversely affect performance or would require much more complexity in modelling to achieve similar performance. It also would not be an accurate reflection of how these models would be constructed and applied in practice. The consideration of all admissions with admission sources of "Accident and Emergency" is arguably still a broad definition of cohort, as admissions are not further broken down by condition. This is done to ensure that the constructed models are applicable and thus add value for a wider range of patients as well as to ensure sufficient data for robust construction of the models considered in this project. Certain machine learning techniques, with ANNs being notable examples, require large amounts of data to reliably capture complex readmission dynamics. Restriction of admissions to a single important admission source strikes a balance between objectives relating to the homogeneity of the modelled population, quantity of data for model estimation, and model applicability.

The exclusion of admissions relating to patients under the age of 18 was similarly motivated by a desire for a more homogeneous population. Patients under the age of 18 may face different readmission dynamics for a range of factors, such as reduced agency compared to adults. In practice and in the literature, paediatric patients are often modelled as a distinct group (Jovanovic et al., 2016). This exclusion also removed a relatively small proportion of patients after previous exclusions, with 70,635 admissions in the final dataset.

Discharges by Hospital

After exclusions, the 70,635 discharges were further split according to whether each discharge related to either GCUH or RH. The rationale for splitting the data according to institution is briefly outlined here and relates to the principle of ensuring model development reflects practical applications set out in Section 3.4.

If the goal of model construction was to arrive at a final model which could be applied to a range of institutions without modification, then the model should be constructed using discharges from multiple institutions. Such models are relevant when the intended application is to capture risk associated with patients for the level of care provided by the 'average' institution, such as for readmission rate measurement or similar policy measures. A model constructed using many institutions will reflect the average

characteristics of the data, including underlying ones such as processes and staff capability. Thus, deviations from expected outcomes or rates are of interest as they can be attributed to better or worse than average performance. In this work, however, the goal of the research questions is to provide a comparison and evaluate statistical and machine learning survival models for use on an institution-specific basis.

The suggested value of survival approaches in this work relates to the additional information provided that allows for model applications assisting managerial decision-making. Model applications leveraging the richer information of survival models are reliant on a model providing a good fit for the relevant institution. A model constructed on data from multiple institutions captures average performance, which contrasts with the institution-specific fit needed for the applications considered in this work. Unless the two institutions were very similar and there were a clear need for more data in model training, the final model would be expected to perform worse in the institution of interest if trained on data from another institution. This was highlighted by Yu et al. (2015), who found that one-size-fits-all models performed worse than institution-specific models, though in their study different features were used in the scenarios because of differences in data collection between institutions.

After splitting the data based according to the relevant hospital for each discharge, there were 46,659 discharges from GCUH and 23,976 discharges from RH.

Feature Processing

In addition to exclusions, adjustments were also made to high-dimensional categorical features characterised by many rare values. These adjustments aimed to reduce their dimensionality while retaining important information. Adjustments were made for the Postcode and AdmitWardCode fields and are briefly described here.

After all exclusions, the Postcode field contained 1,024 unique values, with the 20 most frequent values making up almost 90% (88.94%) of all observations. The number of possible values and scarcity of observations for most values was considered excessive for the information content of this field. The field was reduced to three possible values in a new feature, iGC, representing the region of the Gold Coast the discharged patient was from. The three possible values for this new feature, the postcodes they include, and their frequencies are shown in Table 7 and Table 8 for GCUH and RH respectively.

Table 7. iGC Feature Definition (GCUH)

iGC Field Values	Included Postcodes	Frequency (Training)	Relative Frequency (Training)
InnerGC	4214-4220, 4226-4230	20592	63.05%
OuterGC	4208-4210, 4212, 4221, 4223-4225, 4270-4272, 4275	7766	23.78%
Other	All others	4303	13.17%

Table 8. iGC Feature Definition (RH)

iGC Field Values	Included Postcodes	Frequency (Training)	Relative Frequency (Training)
InnerGC	4214-4220, 4226-4230	12483	74.38%
OuterGC	4208-4210, 4212, 4221, 4223-4225, 4270-4272, 4275	2973	17.71%
Other	All others	1327	7.91%

The AdmitWardCode field details the ward code the patient was admitted to. This field contained 70 unique values across both hospitals, with the seven most frequent values making up almost 90% (88.28%) of all observations. The possible values differ between the two hospitals and thus the recoding for this field was done for each hospital separately as well. For each hospital, the relative frequency of values was generated using the training data. All codes with a relative frequency of at least 5% were kept without change. Those codes with a relative frequency of less than 5% were grouped into an aggregate "Other" category. This was done to reduce the dimensionality of this field while retaining common values. For GCUH, the AdmitWardCode field was reduced from 40 possible values to six. For RH, the AdmitWardCode field was reduced from 22 possible values to five. Table 9 and Table 10 present the values retained in the field for each hospital, the descriptions provided by HCL, and the frequency of each possible value in the recoded field AdmitWardCode1.

Table 9. AdmitWardCode1 Feature Definition (GCUH)

AdmitWardCode	Description	Count (Training)	Relative Frequency (Training)
GEAC	Emergency Acute Block D Lower Ground	8931	27.35%
GED	Emergency Department (DLG) Block D Level Lower Ground	8172	25.02%
GMA	Medical Assessment Unit (BLGS) Block B Level Lower Ground Sth	6071	18.59%
GCDU	Clinical Decision Unit (Dlg) Block D Level Lower Ground	4802	14.70%
GESS	Ed Short Stay (BLGN) Block D Level Lower Ground	2733	8.37%
Other	-	1952	5.98%

Table 10. AdmitWardCode1 Feature Definition (RH)

AdmitWardCode	Description	Count (Training)	Relative Frequency (Training)
RMAU	Med Assess Unit (Robina CG)	10241	61.02%
RCDU	ROB ED Clinical Decision Unit (A Block Ground Floor)	2799	16.68%
RESSU	Ed Short Stay Unit (Robina G)	1614	9.62%
RED	Emergency Department (Robina G)	1588	9.46%
Other	-	541	3.22%

The processing of these two categorical variables was the only feature processing which was carried out and applicable for all models. While additional feature processing was later applied for specific models, those model-specific steps are described where relevant in Chapter 6.

4.3 Descriptive Statistics

Lastly, descriptive statistics are shown for both hospitals using the full data.

Table 11. Descriptive Statistics (Full Data)

	GCUH	RH
Data		
Total admissions	46,659	23,976
Readmissions in 30 days	14.41%	15.65%
Censored Observations	61.62%	58.02%
Selected Features used in Modelling		
Age: Mean (SD)	59.16 (20.50)	66.48 (19.94)
Female (%)	48.13%	52.25%
Region		
Inner Gold Coast	62.94%	74.51%
Outer Gold Coast	24.14%	17.86%
Other	12.92%	7.63%
LOS: Mean	4.53	3.98
Inpatient Admissions in Previous Year: Mean (Median)	1.30 (0)	1.41 (0)
Outpatient Appointments in Previous Year: Mean (Median)	5.42 (1)	4.41 (0)
ED Presentations in Previous Year: Mean (Median)	1.94 (1)	2.16 (1)

The above statistics provide further support for the decision to consider the two hospitals separately for modelling, with notable differences between the two hospitals in several respects. RH is characterised by patients who are older, are admitted for shorter times, have less frequent inpatient and outpatient admissions, and are more often from the inner Gold Coast region. It should also be noted that repeat admissions are included in Table 5 and in subsequent modelling. These are included to ensure that the data considered reflects that available to hospitals, where repeat readmissions are relevant, consistent with the focus on institution-specific models.

Regarding censored observations for the survival formulation in the second research question, these include any admission without a subsequent readmission by the end of the data period (2018-04-30). No restriction on the time between admission and subsequent readmission was imposed. Instead, and as will be described further in Chapter 5, it is of interest whether survival models can demonstrate elevated risk of readmission related to the index admission as well as the return to baseline risk.

5 Methodology

The structure of this chapter is driven by the key decisions. The first decision to be made is which models identified in the literature are included in the set of models considered for each research question. The decision process for model inclusion or exclusion is described in Section 5.1. Greater detail about each included model, its implementation and relevant decisions are then set out in Chapter 6 to retain the focus in this chapter on the core methodological components rather than on the peculiarities of each model. Having defined the set of models considered for each research question, the second major decision relates to model assessment, both for model selection and final model evaluation. The relevant aspects of model performance for the goals of each research question and the corresponding choice of metrics used in model selection and evaluation are discussed in Section 5.2.

5.1 Modelling Techniques Considered

The machine learning survival techniques considered in this project for each research question are a subset of those discussed throughout the literature review. They were selected through their link to the goals of the two research questions. Both research questions, ignoring the consideration of time, relate to estimating the probability of readmission for discharged patients. Accordingly, techniques in which model outputs are not probabilistic were not considered. For the first research question relating to 30-day readmission prediction, a model with non-probabilistic output is usable only as a risk prioritisation tool. For the second research question, models without probabilistic outputs are not usable even as a risk prioritisation tool unless it is assumed that risk rankings are time-invariant. As a result, no SVM techniques were considered as they provided only estimates of event times or time-invariant risk scores. After excluding SVM techniques, the distinct and relevant machine learning survival techniques identified in the literature review are:

- Survival Trees
- Censoring Unbiased Regression Trees
- Random Survival Forests
- Censoring Unbiased Regression Ensembles
- Recursively Imputed Survival Trees

- Bayesian Additive Regression Trees
- Time-coded ANNs
- Multiple time point ANNs
- Hybrid Cox-ANNs

It should be noted that the Censoring Unbiased Regression Trees and Ensembles do not offer probabilistic outputs as part of their original algorithms, but they are included because this can be remedied with minor adjustments. No clear adjustment was identified to produce probabilistic outputs from SVM models. The above machine learning techniques provide probabilistic estimates as functions of time and are thus used for both research questions.

In addition to the machine learning techniques, two statistical techniques are also considered. The first considered is Cox regression. This is the most common survival technique, both in hospital readmission modelling and other fields. It represents the relevant statistical survival benchmark for both research questions. The second statistical technique is logistic regression, considered here because the first research question relates to the value of survival models in predicting 30-day readmissions. Logistic regression serves as an additional point of comparison for the survival approaches as it represents the most frequently employed technique for predicting 30-day readmission. It also provides additional insight into the performance of survival models predicting risk across time, relative to the most common alternative model predicting risk only at a single time. This is valuable for better understanding the trade-off faced in gaining the additional risk-over-time information from survival models that may come at the cost of fixed-point prediction. Logistic regression is not considered for the second research question as it is unable to produce the survival functions which are the focus.

More thorough descriptions of each modelling technique for the project, as well as how they were implemented and their relevant hyperparameters, are deferred to Chapter 6.

5.2 Model Assessment

Having specified the set of models considered under each research question, the focus turns to how model performance is assessed for model selection and final model evaluation. In detailing model assessment, two decisions are made. The first relates to the

use of data in estimating out-of-sample performance while the second relates to the actual measures of performance. While the choice of performance measures is necessarily determined by which of the two research questions is being answered, the method for generating out-of-sample performance is held constant for both questions. Consequently, the method for generating out-of-sample performance estimates is discussed and justified first, followed by separate discussions of the performance measures used for each research question.

5.2.1 Out-of-Sample Performance

To avoid optimistic estimates of model performance resulting from overfitting, the data used for model training should not be used for performance measurement. A variety of methods for estimating out-of-sample performance is available to address the issue of overfitting. For this work, a combination of cross-validation and a longitudinal train and test split are used, with the latter being described first.

Once final models for each technique are selected, the longitudinal training and testing data split is used for final model evaluation. As mentioned in Section 4.1, 70% of the discharges for each hospital were assigned to the training set with the remaining 30% assigned to the testing set. The exact numbers and date ranges are reported in Section 4.1, Table 5. This split aimed to balance the need for sufficient training data to construct data-intensive models, as well as the need for sufficient testing data to provide reliable estimates of performance. The use of a longitudinal split of the data into training and testing sets reflects the fact that practical application of these models in a hospital would be carried out through training on historical patients for use on future ones. A known feature of readmission prediction models in healthcare is the tendency for their performance to deteriorate over time through changes in the medical environment, including technological and procedural factors, as well as changes in the broader environment. Broader changes include sociological and economic, and those in disease prevalence. A longitudinal split better captures the temporal dynamics relevant to the healthcare system for final model evaluation than a random split.

While a longitudinally defined test dataset is most appropriate for evaluating the final models produced in this project, it is not appropriate for selecting the final model of each type. To mitigate the issue of overfitting, the final testing set should be used only once for evaluation of the final models. As many of the included models set out in Section 5.1 involve numerous hyperparameters, five-fold cross-validation of the training data is used

to assess candidate models and determine the final model of each type. While a tertiary split may better capture the longitudinal element to the data, it would reduce the number of discharges available for modelling and be less robust to overfitting, compared to cross-validation for the number of models considered.

Five folds were chosen over the common alternative of 10 folds because of the computational burden involved in constructing many of the models considered. Using 10 folds would double the number of models constructed and evaluated for the same set of hyperparameters considered. The slight benefit in accuracy from 10 folds would thus come at a large cost. It should be noted that five-fold cross-validation is used for assessing all models in which hyperparameters are varied to ensure differences in final model performance are attributable to the models themselves rather than differences in model selection. This includes models such as the random survival forest which would ordinarily be assessed using out-of-bag data. This is consistent with the principles set out in Section 3.4.

In summary, five-fold cross-validation is used for model selection for all models in which hyperparameters are varied. Having selected the final model of each type, these models are constructed on the entirety of the train data before being evaluated using the test data. The following subsections set out the performance measures used in conjunction with these procedures.

5.2.2 Performance Measures – RQ1

Can machine learning survival models improve upon statistical survival models when predicting 30-day hospital readmissions?

In general, the intended application of a model should be reflected in the performance measures used to assess it. For the first research question, a range of new models is applied for the purpose of predicting readmission probabilities at a fixed time point. Such models are typically used in practice by individual institutions to identify patients at the greatest risk of readmission. The associated performance measures for such models in the readmission modelling literature were described in Section 2.1.2.3 and are briefly outlined again here. These performance measures were evaluated based on predicted probability of readmission in 30 days for all models. Predictions from survival models at other time points were not assessed.

The primary metric for model performance for the research question is discrimination as measured by the area under the ROC curve (Hanley & McNeil, 1982), often referred to as AUC. Model discrimination relates to the ability of a model to differentiate between instances where readmission did or did not occur. Good discrimination is required of any model intended to identify those patients at the greatest risk of readmission for targeted interventions.

The second metric for model performance is calibration, measured by the Hosmer-Lemeshow statistic (Hosmer Jr, Lemeshow, & Sturdivant, 2013). The calibration of a model estimating the probability of the event occurring by a time point relates to the agreement between predicted and observed risk. Beyond the prioritisation of patients with a discriminative model, a well calibrated model also allows for more nuanced application and practical evaluation. As examples, calibration is a prerequisite for tasks involving measuring the effectiveness of interventions, directing resources towards patients based on risk thresholds, and cost-benefit analyses. This is also consistent with common practice in the literature.

Finally, accuracy, sensitivity and specificity are used as high-level descriptors of model performance.

5.2.3 Performance Measures – RQ2

How well can various survival modelling techniques capture aspects of hospital readmission risk across time relevant to managerial decision-making?

This research question relates to the potential value of survival models in novel applications for hospital readmissions. Previous research has focused either on investigating risk factors or on producing models predicting risk of readmission by a fixed time point. The survival models considered here are predictive but not limited to a single time point, instead estimating risk over time. In answering the research question, the potential applications of a suitable survival model to assist managerial decision-making must first be considered to determine the desirable model characteristics. Four such applications are proposed and described here to motivate the selection of performance metrics corresponding to the determined model characteristics.

The first proposed application, Dynamic Risk Ranking (DRR), serves a similar purpose to the models employed in previous research and in the first research question. A DRR produced by a survival model should rank patients by risk of readmission to enable better

targeting of interventions and prioritisation of follow-up care. This risk ranking is distinct from those produced by classification models because the ranking is dynamic. That is, patient risk can be calculated after readmission. Classification models are limited to ranking patients at the time they were discharged or relying on assumptions on the evolution of patient-specific risk. The key characteristic of a survival model applied in this fashion would be its discriminative ability, as in the first research question, because this relates to the ability of the model to differentiate between patients who were and were not readmitted. Secondary to discrimination, more nuanced evaluation of the effectiveness of decisions made based on a DRR application requires calibration. While these aspects of model performance are also relevant to classification models, the measures of discrimination and calibration for survival models must reflect estimates of risk over time rather than at a single time point.

The second and third proposed applications are closely related. The second proposed application is the identification of how long patients face elevated risk of readmission post-discharge, termed the Elevated Risk Period (ERP). The third proposed application is the prediction of readmission probability in the ERP, termed the Elevated Risk Period Probability (ERPP). Again, the value of these applications is best described through comparison to classification model applications in the reviewed literature. Previous readmission models have focused on predicting readmission within a fixed period, often 30 days. This carries the implicit assumption that all patients experience the same ERP and ERPPs are estimated based on that simplifying assumption. Survival models, through the provision of estimated survival and hazard functions, enable determination of patient-specific ERPs rather than assuming the evolution of readmission risk is homogeneous for all discharged patients. These applications thus provide a less standardised implementation of current tools which are better able to account for patient differences. It should also be noted that the determination of ERP itself is likely contingent on the needs of the user. It may be determined by focusing on the time where risk is substantially elevated, the time until it begins to approach a baseline, or the time until the baseline is reached. It may be used in determining whether to provide follow-up care, for how long, and for communicating and managing patient expectations. Several ERPs relating to different goals for a single patient are illustrated in Figure 2. Regardless of the exact method for determining ERP and thus ERPP, this application requires a model in which the predicted risk of readmission over time reflects the actual risk. Thus, a necessary

model characteristic in the context of these proposed applications is the calibration of estimated risk over time. Secondary to calibration, discrimination is also desirable to better reflect patient-specific features. As a motivating example, a Kaplan-Meier function would be expected to be characterised by good calibration but poor discrimination as it would provide the same predictions for all patients.

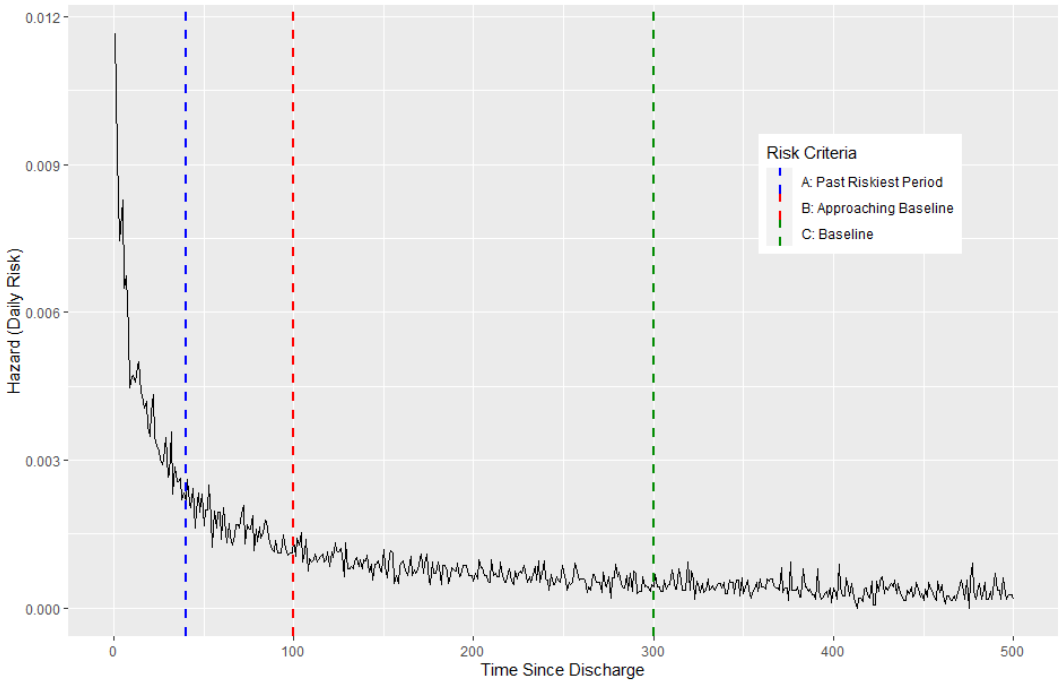

Figure 2. Differences in Elevated Risk Period (ERP)

The fourth application proposed for survival models relates to demand forecasting and is the calculation of Expected Readmissions (ER). Given survival curves from a well fitted model, it is straightforward to calculate an expected number of readmissions in a period conditional on patients being readmission-free up to the start of the period. Reliable forecasting of readmissions is useful from planning and resource allocation perspectives to gauge expected demand (number of readmissions) in a period. Given the probabilities involved in the calculation of expected readmission numbers, calibration is again a necessary model quality for the model to be applied in this way. The precision of the estimates for patients currently at risk is then dependent on the ability of the model to discriminate between patients in the predicted survival functions, again making discrimination a secondary but still important model characteristic.

While the relative value of discrimination and calibration will depend in practice upon the specific application and context, these are two aspects of model performance

determined to be most relevant for practical applications supporting managerial decision-making. Having motivated the aspects of performance considered, the selection of metrics capturing calibration and discrimination suitably accounting for risk predictions over time is now described.

5.2.3.1 Time-Dependent Concordant Index

The most popular measure of model discrimination is AUC, which is applicable for binary classification problems. The formulation of AUC as a concordance index is briefly outlined to motivate measures of discrimination applicable to survival data.

For a given model, the AUC can be interpreted as the probability that a higher risk or probability is assigned to a randomly selected observation where the event occurred than a randomly selected observation where the event did not occur. This can be computed as the number of concordant pairs divided by the number of comparable pairs in the data. For the binary context, let the outcome be denoted as y^b taking values of 1 or 0 corresponding to whether the event did or did not occur. A pair of observations i, j is considered comparable if $y_i^b = 1$ and $y_j^b = 0$. Letting model risk predictions for the i-th observation be denoted $z(x_i)$, this same pair of observations can be considered concordant if $z(x_i) > z(x_j) \mid y_i^b = 1 \,\&\, y_j^b = 0$. Outlining these terms more formally, we can define the comparability and concordance status of all pairs as the following:

$$comp_{ij} = I(y_i^b = 1 \,\&\, y_j^b = 0) \tag{1}$$

$$conc_{ij} = I\left(z(x_i) > z(x_j)\right) \times comp_{ij} \tag{2}$$

The concordance index or area under the ROC curve can then be computed as:

$$AUC = \frac{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} conc_{ij}}{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} comp_{ij}} \tag{3}$$

With this serving as the basis, the idea of a concordance index has been extended to survival data. The most popular such measure for survival data is Harrell's C-index (Harrell et al., 1996), which was originally developed in the context of a Cox regression

model. In this survival-specific formulation, the ideas of concordant and comparable pairs remain but they are redefined to consider time in addition to event occurrence. Pairs of observations are now considered comparable where one observation is known to have remained event-free for longer than the other. More formally, let t_i, c_i, and $y_i = \min(t_i, c_i)$ denote event time, censoring time, and observed time respectively for the i-th observation. Further, let $\delta_i = I(t_i \leq c_i)$ denote the event indicator taking a value of 1 if the event is observed and 0 if the observation was censored. Finally, making the reasonable assumption for many scenarios that for any given observation $t_i \neq c_i$, the comparability of two observations i, j is defined as the following:

$$comp_{ij}^{Harrell} = I(y_i < y_j \,\&\, \delta_i = 1) + I(y_i = y_j \,\&\, \delta_i = 1 \,\&\, \delta_j = 0) \qquad (4)$$

Using this definition, the concordance of pairs can now be determined by establishing the agreement between the model-predicted risk ranking of the pair and the observed ranking. If using a Cox regression model as in Harrell et al. (1996), then concordance can be defined as before in (3), with $z(x)$ now referring to the linear predictor of the Cox model. It should be stressed that, because of the proportionality assumption of the Cox model, the assigned risk ranking of observation pairs is time invariant. Restating the definition of concordance in a form more closely linked to survival models in general and reinforcing the time-invariant nature of the Cox model, we have the following:

$$conc_{ij}^{Harrell} = I\left(S(t|x_i) < S(t|x_j)\right) \times comp_{ij}^{Harrell} \qquad (5)$$

Note that the use of the survival probabilities reverses the inequality and that all choices of $t > 0$ are equivalent for the Cox model. Having respecified the definitions of comparable and concordant pairs for survival data under the Cox model, the concordance index $\mathcal{C}$ can be specified as previously:

$$C^{Harrell} = \frac{\sum_{i=1}^{n} \sum_{j=1;j\neq i}^{n} conc_{ij}^{Harrell}}{\sum_{i=1}^{n} \sum_{j=1;j\neq i}^{n} comp_{ij}^{Harrell}} \quad (6)$$

This metric has been widely employed to assess survival models. It has been stressed that it is appropriate where the ranking of observations is time-invariant, but it has also been applied for models where this time-invariant property is not guaranteed. This has been done through considering survival probabilities at a fixed time point for ranking or by considering cumulative hazard across the entire time horizon (Ishwaran & Kogalur, 2021; Padhukasahasram et al., 2015). It is argued, however, that an appropriate measure of discrimination for a model with time-varying rankings should account for this temporal aspect.

One variation of the concordance index for survival that does account for time-varying rankings of comparable pairs was proposed by Antolini, Boracchi, and Biganzoli (2005). This variation has been employed in various studies utilising survival models for health analytics in which risk rankings are not constant across time (Farrell, Mitnitski, Rockwood, & Rutenberg, 2020; Long & Mills, 2018; F. Yan, Lin, Li, & Huang, 2018). In this variation of the concordance index, time-dependent concordance is introduced as a condition and defined as "the predicted survival probability, at the time where the subject i developed the event, [being] greater for the subject j who actually is still free from the event" (Antolini et al., 2005). In this formulation, the determination of whether a pair of comparable observations i, j is concordant is based on the ranking at the relevant event-time. More formally, and omitting the superscript from $comp_{ij}^{Harrell}$ to reflect that the definition of comparability status in (4) is relevant to survival data in general, the concordance status of two observations is expressed as the following:

$$conc_{ij}^{td} = I\left(S(y_i|x_i) < S\left(y_i\middle|x_j\right)\right) \times comp_{ij} \quad (7)$$

This definition does not assume that the risk ranking of a pair of observations is constant across time. Instead, concordance is determined by comparing predicted survival probabilities at the time when one of the observations experienced the event. This neatly

allows for time-varying rankings, which may characterise non-proportional models. Using this new definition, the time-dependent concordance index is calculated as below:

$$C^{td} = \frac{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} conc_{ij}^{td}}{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} comp_{ij}} \quad (8)$$

This is the concordance index used to assess model discrimination for this research question. It was implemented as part of this project using the above expressions in R.

5.2.3.2 D-Calibration

Turning to the second aspect of model performance, calibration for survival models in this context relates to the agreement of predicted and actual risk across time. This contrasts with what is sometimes termed "1-Calibration" (Andres et al., 2018; Haider, Hoehn, Davis, & Greiner, 2020) which focuses on the agreement between predicted and observed risk only at a single time point. While 1-Calibration measures such as the Hosmer-Lemeshow test are well established and appropriate for predicting n-day readmissions (where n is constant), they lack a direct extension to risk estimates beyond a single time point. Simply applying the test at multiple time points would be inappropriate because the tests would not be independent. Motivated by the prognostic value of individualised survival curves and the need for tests of their calibration, a measure of the distribution termed "D-Calibration" has been outlined (Andres et al., 2018; Haider et al., 2020). The prognostic value of individualised and well-calibrated survival curves stems from settings where decisions are made for patients independent of the risk profiles of other patients, where the goal is to measure the effectiveness of actions taken to reduce avoidable readmissions, and for patient communication.

Deferring consideration of censoring, the core idea of the D-calibration measure is that the model producing individual survival functions acts as a mapping of observed event times on the interval $[0, \infty)$ to survival probabilities on the interval $[0,1]$. It is then expected that the proportion of probabilities in a subset $[a, b]$ of the interval $[0,1]$ will be equal to the width of the interval. To motivate this, for a well calibrated model it can be reasonably expected that 60% of the observed event times were associated with survival probabilities below 60%. Similarly, it would be expected that 20% of the observed event times were associated with survival probabilities below 20%. Combining these, the

expected proportion of observed event times mapped to probabilities in the interval $[0.2,0.6]$ would be expected to be 40%. Borrowing notation from Haider et al. (2020), this can be expressed more formally as:

$$\frac{|D_{\Theta}([a,b])|}{|D|} = b - a \tag{9}$$

Where D is the total data, $|\,.\,|$ is the size of a data set, and $D_{\Theta}([a,b])$ is the subset of the data for which the model Θ maps observed event times onto survival probabilities in the interval $[a,b]$. From the principle that a well-calibrated model should produce a mapping of observed event times onto survival probabilities that are uniformly distributed, a straightforward χ^2 test can be used to test the null hypothesis of D-calibration[2]. This is done by specifying k groups of equal widths over the interval $[0,1]$ and considering the number of observations mapped to that interval compared with the expected number.

The extension of the above method to censored observations can now be described. The adjustment for censored observations is based on allowing partial contribution of censored observations to each of the k groups according to the conditional probability of group membership under the null hypothesis. As a similar example to that used in Haider et al. (2020), consider 10 groups of width 10% and a censored observation where the survival probability associated with the censoring time is 65%. The survival probability associated with the true event time is thus known to be between 0% and 65%, and if it is uniformly distributed (as under the null hypothesis), the conditional probability of its being in any of the 10 groups can be calculated. More formally,

$$\Pr(S_{\Theta}(t_i|x_i) \in [a,b) | S_{\Theta}(t_i|x_i) < 0.65) = \frac{\Pr\left(S_{\Theta}(t_i|x_i) \in [a,b)\ \&\ S_{\Theta}(t_i|x_i) < 0.65\right)}{\Pr(S_{\Theta}(t_i|x_i) < 0.65)}$$

[2] An implicit but reasonable assumption of this method is that the true stochastic process underlying the event times is not characterised by notable probability masses, such as in the case of a Heaviside step function.

This is straightforward to compute. Using the interval $[0.6,0.7)$ for $[a, b)$ and omitting the dependence on x_i, the following solution is obtained:

$$\begin{aligned}
&\Pr(S_\Theta(t_i) \in [0.6,0.7) | S_\Theta(T_i) < 0.65) \\
&= \frac{\Pr(S_\Theta(t_i) \in [0.6,0.7) \ \& \ S_\Theta(t_i) < 0.65)}{\Pr(S_\Theta(t_i) < 0.65)} \\
&= \frac{\Pr\left(S_\Theta(t_i) \in [0.6,0.65)\right.}{\Pr(S_\Theta(t_i) < 0.65)} \\
&= \frac{0.05}{0.35}
\end{aligned}$$

For other intervals, it can be shown that the contribution of this censored observation is either $0.1/0.35$ or 0, depending on whether the starting point of the interval is less than or greater than the censoring time, respectively. Rather than replicate the final algorithm summarising the overall method exactly, the reader is directed to the appendix of the work of Haider et al. (2020).

This serves as the measure of calibration used in this project for the second research question and was implemented using R code from https://github.com/haiderstats/ISDEvaluation.

For completeness, it should be noted that one limitation from the incorporation of censored observations under the null hypothesis is that it results in a smoothing effect. As the null hypothesis of a uniform distribution is used, censored observations are evenly spread across those intervals that the true survival probability may have been associated with. Consequently, a dataset with a high degree of early censoring will result in more even interval sizes than might otherwise be the case, increasing the p-value associated with the test. Figure 3 shows the distribution of follow up times associated with censored patients for the entire dataset considered in this work, showing only a slightly greater frequency of censoring at earlier times.

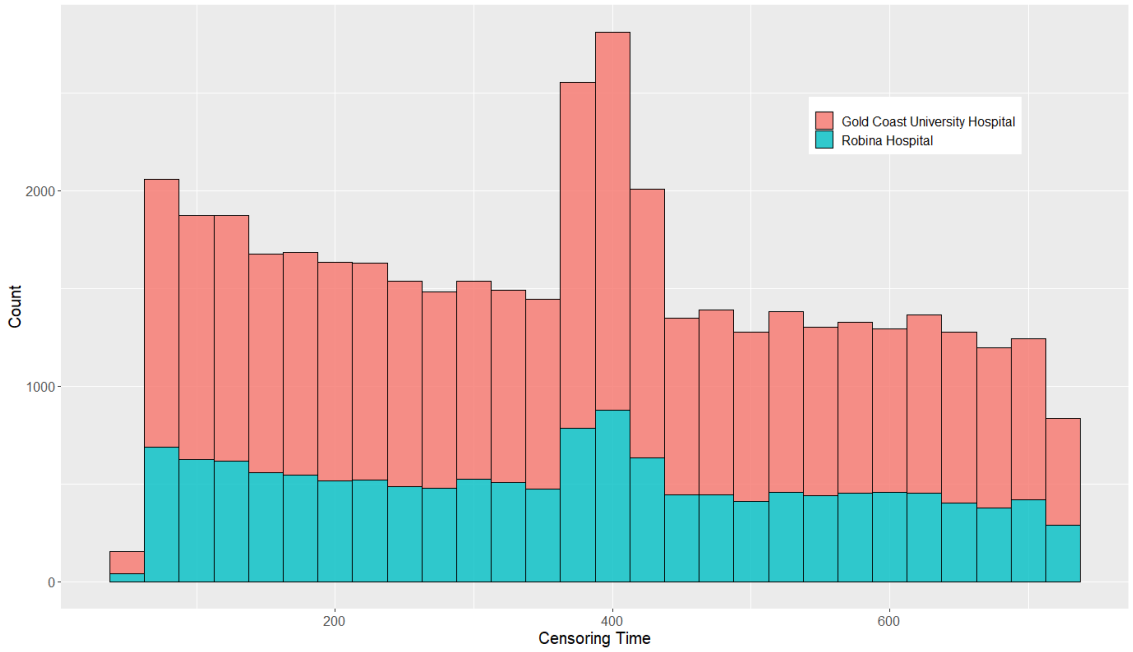

Figure 3: Censoring Distribution by Hospital

5.2.3.3 Integrated Brier Score

The above has outlined appropriate measures of discrimination and calibration identified in the literature for use in model evaluation under the second research question. Both appropriately incorporate the temporal component of risk predictions under survival models. Neither measure, however, is appropriate for model selection in isolation. Ignoring the fact that p-values are not intended for ranking, particularly given the omnibus nature of the underlying χ^2 test, using the D-Calibration measure would select a model with good calibration but ignore the need for discrimination. For example, a Kaplan-Meier function would be expected to be characterised by a very large p-value but is of limited use given its lack of individualised predictions. Similarly, the time-varying concordance index would result in a high-discrimination model but it ignores the need for calibration. C-statistics (which include the time-varying concordance index) are also characterised by low sensitivity for comparing many predictive models (Uno, Cai, Pencina, D'Agostino, & Wei, 2011). Rather than use either measure individually or combine them in an ad-hoc manner, the Integrated Brier Score is employed for model selection. Use of this measure is consistent with common practice in survival model assessment. This metric also has the attractive feature of considering both discrimination

and calibration, as it can be shown to be decomposed into a sum of these two components[3] provided that a distinct number of model predictions exist (Steyerberg et al., 2010). This measure is described here.

The Brier score represents the squared error of probabilistic predictions for binary outcomes and was originally proposed as a measure of accuracy in weather prediction (Graf, Schmoor, Sauerbrei, & Schumacher, 1999). For binary classification problems with outcome y^b equal to 1 or 0 and probabilistic model predictions $P_\Theta(y_i^b = 1|x_i)$, it can be expressed as the following:

$$BS = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^b - P_\Theta(y_i^b = 1|x_i)\right)^2 \tag{10}$$

Using this expression, and temporarily assuming K unique predictions given as $P_{\Theta,\mathrm{k}}$ for $k = 1, \dots, K$, the decomposition into calibration and discrimination components can be shown. Let η_k denote the number of observations with prediction p_k and let λ_k denote the proportion of these η_k observations where the event occurred. With the assumption of discrete and distinct predictions, (10) can be expressed as:

$$BS = C + D = \frac{1}{n}\sum_{k=1}^{K}\eta_k\left(\lambda_k - P_{\Theta,k}\right)^2 + \eta_k\lambda_k(1-\lambda_k) \tag{11}$$

With the decomposition of the Brier Score into calibration (C) and discrimination (D) demonstrated, the focus returns to how (10) can be applied for survival problems. Assuming no censoring, or at least none prior to time t, the extension is relatively straightforward through replacing the actual outcome and model prediction terms with the relevant survival analysis quantities. Returning to the treatment of y as the minimum

[3] As the Brier score can be shown to be a linear sum of the calibration and discrimination, the relative weighting of these two components could also be adjusted according to context- or application-specific needs.

of censoring and event times and assuming no censoring at this point, the Brier Score for survival data is given by the following:

$$BS_{surv}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(I(y_i > t) - S_{\Theta}(t|x_i)\right)^2 \quad (12)$$

Extending this further to the case where an overall score is desired rather than at a single time point t, the Integrated Brier Score integrates over all times with respect to a weighting function $W(t)$.

$$IBS_{surv} = \int_0^{\max(t)} \frac{1}{n}\sum_{i=1}^{n}\left(I(y_i > t) - S_{\Theta}(t|x_i)\right)^2 dW(t) \quad (13)$$

Finally, censored data were accounted for by Graf et al. (1999) using an IPCW approach. Relaxing the assumption of no censoring, the IPCW approach to computing the Brier Score produces the following equation:

$$BS_{cens}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{I(y_i \leq t) \times \delta_i}{G(y_i)}\left(0 - S_{\Theta}(t|x_i)\right)^2 + \frac{I(y_i > t)}{G(t)}\left(1 - S_{\Theta}(t|x_i)\right)^2 \quad (14)$$

where $G(t)$ is Kaplan-Meier estimate of the censoring survival function. This expression gives a weight of zero to those observations censored before t, but inflates the weighting of other observations commensurate with the probability of having been censored. Those observations which were censored before t are implicitly included in the expression by inclusion in the estimation of G. The consistency of the estimator with this weighting scheme was shown by Gerds and Schumacher (2006).

Following the same idea as previously, for a reasonable weighting function $W(t)$ the IBS with modification for censoring can be expressed as the following:

$$IBS_{cens} = \int_0^{\max(t)} \frac{1}{n} \sum_{i=1}^{n} \frac{I(y_i \leq t) \times \delta_i}{G(y_i)} \left(0 - S_\Theta(t|x_i)\right)^2 + \frac{I(y_i > t)}{G(t)} \left(1 - S_\Theta(t|x_i)\right)^2 dW(t) \quad (15)$$

This expression defines the version of the IBS used in this project, with $W(t) = \frac{1}{\max(t)}$, implemented using R code from https://github.com/haiderstats/ISDEvaluation.

5.2.3.4 Summary of RQ2 Measures

In summary, several measures are used in answering the second research question. The IBS measure, containing both discrimination and calibration components, is used for model selection. For the evaluation of the final models of each type, the IBS is used in addition to the time-dependent concordance index and tests of D-Calibration. The wider range of metrics employed for final evaluation reflects the smaller universe of models under consideration and provides a richer description of their performance.

6 Modelling Techniques and Implementations

In the previous chapter, the set of modelling techniques considered for each research question and technique-agnostic tools for model selection (namely the use of cross-validation and choice of metrics) were described. The purpose of this chapter is to provide more detailed descriptions of modelling techniques used in the project, including how they were implemented, decisions in model construction, and the selection process used to determine the final model. The value of reproducible research in the area was used as a guiding principle in all decisions involved in the selection of final models. Reproducibility promotes comparability with future research, which is of particular importance given the lack of comparability within the area of hospital readmissions and was highlighted as a principle for this work in Section 3.4. Linked to this, search grids are defined where hyperparameters are relevant for determining the final model.

6.1 Logistic Regression and Cox's Proportional Hazards

The model fitting processes for logistic regression and Cox regression are discussed together because of the large overlap in data adjustments and the model selection process, linked to the statistical nature of these two techniques. As both logistic and Cox regression are well established, the techniques themselves are not described here. Readers are instead directed to the excellent textbooks of Hosmer, Lemeshow, and Sturdivant (2013) and Kleinbaum and Klein (2012) for logistic and Cox regression respectively

6.1.1 Data Adjustments

Given that statistical models make assumptions about the nature of the underlying data, adjustments to the training data are a natural part of a reasonably sophisticated model implementation. This is relevant to this work as many of the candidate features exhibit high positive skew with large outliers. The presence of extreme outliers or skewness leading to sparse regions in the predictors is problematic as a result of the large effect that extreme values can have on coefficient estimates. As a simple example, a feature may have an even distribution of values between 0 and 100 and a well-defined coefficient estimate for this range. An outlier value of 600 or 1000 may have a disproportionately large positive or negative effect on the final coefficient value. As the extreme observation can make a large contribution to the deviance of the model, the model is incentivised to better fit that specific point at the expense of fit in the region where most of the data fall.

As would be the case in practice, the data are adjusted prior to model fitting to mitigate these issues. This was only done for statistical models, as machine learning models are purported to be more flexible and better able to handle such data characteristics without requiring extensive pre-processing. To ensure reproducibility, adjustments were made according to two rules:

Rule 1 – Let $o_{1,j}, o_{2,j}, \dots, o_{U,j}$ be the U unique and ordered values of the j-th covariate. If the relative frequency of $o_{1,j}$ is greater than 85%, then an upper bound for this covariate is set at the value $o_{2,j}$. More formally, if the relative frequency of $o_{1,j}$ is greater than 85%, the variable is transformed with the equation $x_{i,j}^{*} = \mathbf{1}(x_{i,j} = o_{1,j}) \times o_{1,j} + \mathbf{1}(x_{i,j} > o_{1,j}) \times o_{2,j}$ where $\mathbf{1}$ is the indicator function taking a value of 1 if the conditions is satisfied and 0 otherwise.

Rule 2 – Where the relative frequency of values in a right-looking window of width 5 is less than $1/U$, impose an upper bound at the first value in the window. More formally and indexing the unique and ordered values by u, if the combined relative frequency of $o_{u,j}, \dots, o_{u+4,j}$ is less than $1/U$ and u is the minimum value for which this condition is true, the variable is transformed with the equation $x_{i,j}^{*} = \mathbf{1}(x_{i,j} \leq o_{u,j}) \times x_{i,j} + \mathbf{1}(x_{i,j} > o_{u,j}) \times o_{u,j}$.

Adjustments were determined through application of these rules to the training data but applied to both the train and test data for each hospital. For example, these rules dictated that an upper bound of 95 be applied for the age variable when considering the training data of GCUH. This upper bound was applied to the age variable of both the training and testing data for GCUH. The modifications to the variables for both GCUH and RH are shown below in Table 12.

Table 12. Statistical Model Data Transformations

Feature	Upper Bounds - GCUH	Upper Bounds - RH
Age	105 → 95	107 → 97
ED_NumPresPrevYear	74 → 11	76 → 11
ED_NumPresSincePrevAdm	38 → 1	23 → 1
ED_NumPresSincePrevAdmALL	38 → 1	23 → 1
Inpat_NumAdmPrevYearALL	34 → 8	34 → 7
Inpat_PrevAdmLOSPrevYear	195 → 14	150 → 12
Inpat_PrevAdmLOSPrevYearALL	195 → 16	154 → 14
Inpat_TimeSincePrevAdmALL	365 → 162	365 → 203
Inpat_TotalAdmInICU	6 → 1	7 → 1
Inpat_TotalAdmInICUALL	6 → 1	7 → 1
Inpat_TotalTimeAdmPrevYear	273 → 31	152 → 30
Inpat_TotalTimeAdmPrevYearALL	297 → 44	270 → 37
LOSCalc	303 → 22	489 → 20
Outp_NumApptPrevYear	140 → 30	185 → 27
Outp_NumApptSincePrevAdm	105 → 12	103 → 8
Outp_NumApptSincePrevAdmALL	114 → 12	86 → 9

In adjusting the data to avoid undue influence of extreme outliers or skewness, a decision was made between truncation and case deletion. Case deletion was not used for two reasons. First, removing instances with uncharacteristic values would mitigate the statistical difficulties but would reduce the reliability of the models when applied to any future admissions with values considered uncharacteristic. As this work aims to reflect practical considerations for the development and use of models predicting readmission, this is undesirable. Second, case deletion reduces the sample size available with which the model can learn relationships between fields and outcomes. When applying the truncation rules, 25,200 instances (77.16%) of the GCUH training data and 12,741 instances (75.92%) of the RH training data were subject to modification of at least one feature value. While rules defined specifically for case deletion could reduce these numbers, they still serve to illustrate that a large portion of data would be lost with a case deletion approach.

6.1.2 Decisions in Model Construction

To construct a realistic benchmark logistic regression or Cox regression model predicting readmission risk, a thorough modelling process was followed. The process of arriving at a final model encompasses two major aspects:

- Variable Selection
- Parameter Estimation

Variable selection is the process of determining the independent variables related to outcomes, being 30-day readmission for logistic regression and readmission status in conjunction with follow-up time for Cox regression. Further, decisions must be made as to what interactions between these independent variables are relevant and whether non-linear relationships between the dependent variable and independent variables are present. Including unnecessary variables in any form will result in an overparameterised model which may generalise poorly to the test data. Excluding variables, interactions, or non-linear terms that are relevant will also adversely affect the model, limiting its ability to capture real relationships and introducing bias.

Parameter estimation is stated here as a separate aspect but it can overlap with variable selection. This is the process of estimating the relationships between the outcome and independent variables. It can be modified through regularisation techniques. These techniques can help mitigate the issue of correlation structures between the independent variables and, depending on the technique, can also perform variable selection.

In this project, variable selection was the primary focus. Stepwise variable selection procedures allow for the selection of a subset of relevant covariates through considering many combinations. Using these procedures in conjunction with regularisation as well is inappropriate, however. Regularisation would be used if there was an a priori expectation around the features, interactions, and non-linear terms that were related to readmission outcomes. This is because using regularisation procedures for both shrinkage and variable selection would necessitate definition of an initial candidate set of covariates consisting of main effects, interactions, and non-linear terms. To avoid biasing the model through an incorrect specification which may fail to include important terms, however, the author opted for a more thorough stepwise variable selection procedure considering a wider range of main effects, interactions, and non-linear terms. The term selection process described below, using stepwise procedures, allows for consideration of each of these components in three stages. Stepwise procedures also partially address correlation between variables through the exclusion of covariates sharing information content with covariates already included.

For this variable selection process, three types of terms were considered that cover a wide range of possible effects. These are main effects, interactions, and polynomial terms. Despite the focus on considering a broad range of terms, it is not realistic to define a covariate set involving all possible main effects, interactions, and polynomial terms at once. Instead, a greedy-style approach to determining the terms to include in a final model is proposed and used, in which stepwise procedures are repeatedly employed by breaking the variable selection problem into three stages.

Stage 1: The 19 candidate covariates (three factor variables and 16 numeric) will be considered as main effects. A hybrid forward and backward stepwise procedure beginning from a full model will be applied to find the set of covariates maximising an appropriate information criterion. The set of covariates included after this stepwise procedure will then be used in Stage 2.

Stage 2: The main effects included at the end of Stage 1 will then form the basis for potential interaction terms in Stage 2. Using the reduced covariate set from Stage 1, all possible 2-way interactions will be considered. A hybrid forward and backward stepwise procedure will again be applied to find the set of main effects and interactions maximising an appropriate information criterion.

Stage 3: In Stage 3, the covariates associated with main effects still included after Stage 2 will be considered with non-linear effects. For each numeric variable with more than 10 unique values and retained at the end of Stage 2, a squared and cubic term are defined and added to the set of candidate terms. Once the non-linear terms have been defined, a hybrid forward and backward stepwise procedure is again applied. This will determine the final terms to include in the model, with the upper limit being all the main effects and interactions from Stage 2 as well as all non-linear terms defined in this stage.

This 3-stage strategy does not consider all possible terms which could be included in the model because of the size of such a covariate set, but instead focuses on identifying the main effects, interactions, and then non-linear terms sequentially. While it is possible this may miss important interaction or non-linear terms involving covariates which were excluded in Stage 1 or Stage 2, it is unlikely that these more complex terms would be important while the main effects were not. Hence, the method is a greedy approach to defining the final set of covariates for the model as an exhaustive approach is not feasible and would increase variance.

As part of the described procedure, the entire training data set for each hospital will be used in the stepwise procedures. Stepwise procedures aim to maximise information criteria which are intended to represent estimates of out-of-sample performance. In view of this, using cross-validation procedures as well is unnecessary. This makes the model development procedure distinct from that used for most of the machine learning techniques, but it reflects the lack of hyperparameters relevant to logistic regression and Cox regression beyond the information criterion being maximised in the stepwise procedures.

For logistic regression, the procedure described above was conducted once using the Akaike Information Criterion (AIC) for all three stages and once under the Bayesian Information Criterion (BIC) for all three stages. AIC represents the less restrictive criterion while BIC encourages a more parsimonious model. Convergence issues were encountered in the fitting algorithm when using AIC for all three stages for the Cox regression model. Accordingly, and consistent with the greedy nature of the described process, AIC was used only in Stage 1 and BIC was used for Stage 2 and 3. This was done using the survival (Therneau, 2020) and stats (R Core Team, 2020) R packages.

6.2 Survival Trees

Decision trees are a commonly used technique in classification and regression contexts. Decision trees repeatedly partition data into increasingly homogeneous subgroups according to sequences of binary rules. The result is a highly interpretable and flexible non-parametric model. They have been adjusted for survival contexts through straightforward application of splitting functions accounting for censored data (unlike squared error and entropy) and appropriate terminal node summaries such as Kaplan-Meier or Nelson-Aalen survival functions. For this project, they were implemented using two R packages, randomForestSRC (Ishwaran & Kogalur, 2021) and rpart (Therneau & Atkinson, 2019).

6.2.1 Decisions in Survival Tree Construction

Compared with logistic and Cox regression, survival trees are less sensitive to the distribution of independent variables. Further, variable selection is performed automatically as only variables improving the model at each split are used. If a variable does not hold any predictive power, it will not be used for splitting within the model. The

entire candidate variable set can thus be used to construct the tree and the resulting model can be inspected to identify which variables were found to be important. Neither variable selection nor data distribution is considered in constructing survival trees. The two aspects of model construction considered as hyperparameters in this project and described below are:

- Splitting Function
- Tree Size

Splitting Function

Selection of a splitting function appropriate for survival data has allowed for straightforward modification of decision trees for survival data. The goal of the splitting function is to measure how well a given split separates the data into homogeneous groups, with many such splitting functions having been suggested. For the purposes of this project, two popular splitting functions are considered. These are the log-rank statistic and the one-step full likelihood approach of (Leblanc & Crowley, 1992).

The log-rank statistic is a two-sample statistic used to compare the survival distributions of two samples or populations. It is implemented as part of randomForestSRC, an R package for building random forests for survival, regression, and classification. While a tree-specific implementation was not found in which the log-rank statistic was used and could be reproduced by the author, a tree can be constructed in the forest framework of randomForestSRC through appropriate parameter specification. This is done by specifying that bootstrapping not be used, only a single tree be constructed, and all covariates be considered at each split point. Other packages using log-rank splitting statistics for survival trees or forests included the party package (Hothorn, Hornik, & Zeileis, 2006) and ranger package (Wright & Ziegler, 2017). The split points and statistics of these packages could not be replicated by manual calculation and so were not used.

The one-step likelihood approach evaluates candidate splits according to the likelihood or, equivalently, deviance associated with a one-step full likelihood (Leblanc & Crowley, 1992). More specifically, the proportional hazard model is given by the following:

$$\lambda(t|x) = \lambda_0(t)s(x) \tag{16}$$

Where $s(x)$ is typically a log-linear function of covariates and parameters. For the one-step full likelihood approach, $s(x)$ is replaced with a factor for terminal node membership and the cumulative baseline hazard is approximated by the Nelson estimate. This substitution means the full likelihood function can be evaluated (rather than the often-seen partial likelihood) and could be maximised through an iterative procedure. In the one-step likelihood splitting function, each potential split at a given point in the tree is evaluated by considering the full likelihood after the first step of the iterative maximisation procedure. The split associated with the highest one-step likelihood is then selected. This splitting function was implemented as part of the popular rpart package through an equivalence to Poisson methods after rescaling the data to have exponential baseline hazard.

Tree Size

The second major consideration in constructing a survival tree is the tree size. In general, the tree can learn increasingly complex relationships as the number of splits increases. On training data, this means performance continues to increase as new splits are added to the tree and terminal nodes have fewer observations in them. This increase in performance corresponds to increased ability to capture real relationships, which is expected to improve performance on new data, as well as to provide increased ability to capture natural variation in the training data, which causes overfitting. The goal in tree construction is to find a size large enough to provide a good fit to the training data but not so large that it overfits and thus does not generalise well to new data.

The randomForestSRC and rpart implementations of survival trees differ in the control afforded in the tree construction process, particularly with respect to tree size. Consequently, tree size was varied according to the relevant hyperparameter in each implementation. Rather than compare dissimilar survival trees from these two implementations to choose a final survival tree, the two implementations are treated as distinct models for this project and a final model from each considered. This is because while the models under both splitting functions will be functions of tree size, this aspect of the trees will be varied in different manners and so are not directly comparable.

For the log-rank splitting function in the randomForestSRC implementation, tree size is controlled through stopping conditions based on minimum node size and node depth.

From the documentation for the randomForestSRC package (Ishwaran & Kogalur, 2021), three conditions must be met for a node to be split:

1. *The current node depth must be less than the maximum node depth allowed*
2. *The current node size must be at least 2 times the node size specified*
3. *The current node must be impure*

The third condition simply means that if a node is perfectly homogeneous, no further splits are made. The first two conditions relate to parameters which can be specified when building the tree. Node depth is measured by the number of splits between the root and terminal nodes. The minimum node size parameter controls the minimum number of observations that should be in a terminal node.

Given the size of the datasets involved in this project and the expectation that the subgroups for discharged patients created in the recursive partitioning procedure may realistically be of different sizes, node depth was used to control tree size. This parameter was varied between two and 20 in steps of one to generate 19 trees of various sizes under the log-rank statistic. Each distinct tree was grown and evaluated five times as part of the five-fold cross-validation procedure.

For the one-step likelihood splitting approach in the rpart implementation, tree size is controlled through cost complexity pruning. In essence, this involves building a large tree before pruning it by removing weak subtrees. This approach is more robust than using early stopping conditions because of the potential for valuable splits to follow weak splits. More formally, consider a large tree $\mathcal{T}$ where the number of terminal nodes is given by $|\mathcal{T}|$ and the error of the tree is given by $\mathcal{R}(\mathcal{T})$. Next, the cost-complexity function is defined as $\mathcal{R}_\alpha(\mathcal{T}) = \mathcal{R}(\mathcal{T}) + \alpha|\mathcal{T}|$, with α being a regularisation parameter. For a given value of α, the goal is to find the subtree $\mathcal{T}_\alpha$ of the original large tree $\mathcal{T}$ that minimises $\mathcal{R}_\alpha(\mathcal{T})$. This is achieved by repeatedly removing the least valuable internal nodes in the tree. The sequence of subtrees obtained through this repeated pruning of the weakest internal nodes can be shown to contain the unique $\mathcal{T}_\alpha$ minimising $\mathcal{R}_\alpha(\mathcal{T})$. This makes the sole parameter of interest for controlling tree size the cost-complexity parameter α. As the value is increased, smaller trees are selected.

For generating candidate trees in this project under the one-step likelihood splitting function, the cost-complexity parameter is varied between 0.0001 and 0.01. Rather than use a sequence of values for the cost-complexity parameter with even differences, two

sequences are used. The first sequence consists of values from 0.0001 to 0.001 in steps of 0.00005 and the second sequence consists of values from 0.002 to 0.001 in steps of 0.01, resulting in 28 unique values for the complexity parameter. The purpose of using two sequences is to avoid the under-representation of lower magnitude values which would be caused by considering a single step size. The greater emphasis on the range between 0.0001 and 0.001 was driven by preliminary results which indicated trees with complexity parameters in this range exhibited superior performance.

The set of hyperparameter values considered for the two survival tree implementations in this project are shown in Table 13.

Table 13. Search Grid Hyperparameters (Survival Tree)

Model Type	Parameters Varied	Parameter Values Considered
Survival Tree – One Step Likelihood	Cost-complexity parameter	0.00010, 0.00015, 0.00020, …, 0.0090
		0.00100, 0.00200, 0.00300, …, 0.01000
Survival Tree – Log Rank Statistic	Node depth	2, 3, 4, …, 20

6.3 Censoring Unbiased Regression Trees (CURT)

Motivated by the fact that survival trees use splitting rules distinct from those which would be used if the full data were observed, an alternative approach is to use Censoring Unbiased Transformations (CUTs) of the loss function. A CUT of a given measure (e.g. squared error) can be defined as any function of the observed data (including censored observations) if it shares a conditional expectation with the measure for all covariate combinations (Steingrimsson et al., 2019). Using a similar example to Steingrimsson et al. (2019), Buckley-James regression (Buckley & James, 1979) aims to estimate the continuous dependent variable t. In the presence of censoring, however, t is replaced with t^*, which is defined as:

$$t_i^* = \delta_i t_i + (1 - \delta_i) E(t_i | t_i > c_i, x_i) \quad (17)$$

It can then be shown that t_i^*, conditional on the observed data, shares an expectation with t_i for all i.

Censoring Unbiased Regression Trees, referred to hereafter as CURTs, are an alternative extension of decision trees to survival data. Whereas the survival trees described in Section 6.2 use a splitting function that is specific to survival data, CURTs employ CUTs of the loss function which would be used if the full data were observed. The basic tree algorithm applies when constructing a CURT, with the key difference being that the splitting function is a CUT of the full-data loss function rather than a function specific to survival data. As with survival trees, the terminal nodes of a CURT can be summarised using non-parametric estimators of the survival or hazard function.

Several CUTs have been used in the literature. In the following paragraphs, a brief overview of the relevant CUTs for this project is given, but greater detail is available in the cited studies. A relatively simple example is Inverse Probability of Censoring Weighting, or IPCW. Using the notation of this work, Steingrimsson et al. (2016) expressed the IPCW CUT for a given loss function for a decision tree as:

$$L_{IPCW}(O|G) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} I(x_i \in \mathrm{N_k})\frac{\delta_i L(z_i, \beta_k)}{G(y_i|x_i)} \tag{18}$$

where the IPCW loss is a function of the observed data O conditional on the estimator for the censoring survival function G. K represents the number of terminal nodes, while $I(W_i \in N_k)$ is an indicator function returning 1 if x_i is associated with the k-th terminal node. δ_i is the event indicator of the i-th observation, y_i is the minimum of the censoring and event time and $z_i = h(y_i)$, where h is a strictly increasing function (e.g. a logarithm or the identity function). Finally, $L(z_i, \beta_k)$ is the original loss function and β_k is the prediction associated with the k-th terminal node. The equivalence of the expectations for (18) and the full data loss function is straightforward to demonstrate. This specific CUT has been used in tree construction previously in the literature (Molinaro et al., 2004), and makes two assumptions: 1) the event time is conditionally independent of the censoring time given the covariates and 2) $G(\mu|x) > 0$ for all $\mu > 0$.

For this project, the doubly robust survival trees proposed by Steingrimsson et al. (2016) are used, which employ a doubly robust CUT that more efficiently uses the partial information available in censored observations through an augmentation term. This augmentation term depends on both the censoring and conditional survival functions. The key property of this doubly robust CUT is that it is a consistent estimator of the full data loss function if at least one of the survival function and censoring function are correctly specified. Steingrimsson et al. (2016) present the doubly robust CUT of the loss function as the following:

$$L_{DR}(O|G,\hat{Q}) = L_{IPCW}(O|G) + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} I(x_i \in N_K)\left(\frac{(1-\delta_i)\hat{Q}_k(y_i,x_i)}{G(y_i|x_i)} - \int_0^{\tilde{T}_i}\frac{\hat{Q}_k(u,x_i)}{G(u|x_i)}d\hat{\Lambda}_G(u|x_i)\right) \quad (19)$$

The details of this loss function and its derivation are available in Steingrimsson et al. (2016). The pertinent aspect of the function for this work is the addition of the augmentation term to the IPCW loss function. This term depends on the estimator of the censoring survival function G and an estimator of the conditional survival function $\hat{S}$, which can be seen in the definition of $\hat{Q}(u,w)$.

$$\hat{Q}_k(u,x) = -\frac{\int_u^{\infty} L(h(r),\beta_k)d\bar{S}(r|x)}{\bar{S}(u|x)} \quad (20)$$

For implementing this model, R code was provided by the authors of the original doubly robust survival tree article (Steingrimsson et al., 2016). This code used the rpart package to construct trees after applying a response imputation procedure to the data which allowed for application of more traditional tree implementations while retaining an equivalence to the CURT procedure described. As the original code was intended for performing regression with the final CURT, the code was modified in this work to generate survival curves rather than regression estimates. This was done through modifying the code to first predict terminal node membership for all data used in model

construction. A Kaplan-Meier estimator was then constructed for each terminal node using all data falling into that terminal node. Finally, new observations were assigned the relevant Kaplan-Meier function by first predicting terminal node membership and then outputting the associated function. This is the same process used in survival trees to create and then assign terminal node summaries.

6.3.1 Decisions in CURT Construction

For the purposes of this project, the process of constructing a CURT is set out in two major steps:

1. Choose the CUT of the loss function
2. Construct the decision tree using the transformed loss function

Each of these steps involves decisions for model construction. For the first step, given that the doubly robust CUT of (19) is chosen, this CUT requires specification of the censoring and conditional survival functions. For the second step, the size of the tree constructed must be chosen. In making both decisions in this project, faithfulness to the original implementation was considered.

Choosing the Censoring and Conditional Survival Functions

As outlined already, the doubly robust CUT requires an estimator of the censoring survival function G and the survival function S. Consistent with the provided code and original paper of Steingrimsson et al. (2016), the censoring survival function was estimated using the Kaplan-Meier estimator, making the assumption that censoring is independent of the covariates. This assumption is justified by the censoring mechanism present in the data used. Data were collected from the 30th of April, 2016 to the 30th of April, 2018. Outcomes are censored only if readmission has not occurred by the 30th of April, 2018. This mechanism does not depend on the covariates beyond their influence on readmission time.

Of greater interest for modelling is the survival function. Given the objective is to model readmission risk as a function of time and covariates, the Kaplan-Meier estimator was not used. Instead, the model used to estimate the conditional survival function was treated as a hyperparameter in the CURT model, with the candidate models being those considered in Steingrimsson et al. (2016) and allowed for in the provided code. These consisted of a decision tree using the rpart package as described in Section 6.2, a random survival forest

using the randomForestSRC package as described in Section 6.4, and two parametric AFT models.

Choosing the depth of the CURT

The rationale for determining tree size is the same as for survival trees more generally. For the implementation of this CURT, the rpart package is used and tree depth is determined by the cost-complexity parameter. Two approaches for selecting the best value of this cost-complexity parameter were considered.

The first approach is that used by Steingrimsson et al. (2016). Under this approach, for a given conditional survival model, the complexity parameter for the CURT is chosen using a simulation approach with 10-fold cross-validation based on a quadratic loss function. For each possible depth of the tree, the 10-fold cross-validated error for that cost-complexity parameter value across simulations is averaged. The cost-complexity parameter value associated with the lowest error across all simulations is used for constructing the final model.

While this simulation approach is the method employed by the authors proposing the doubly robust CURT, it performs model selection on a different basis than other models in this project. Thus, while the internal simulation method is used to remain faithful to the original implementation, a second set of models are constructed in which the cost-complexity parameter value is varied and treated as a prespecified hyperparameter. That is, for each candidate cost-complexity parameter value (and conditional survival model), a full model will be constructed and evaluated using 5-fold cross-validation. This allows for the choice of final cost-complexity parameter value to be made with reference to the performance measure specific to the research question rather than with reference to squared error. Smaller values of the complexity parameter were considered than in the case of survival trees based on preliminary results indicating that smaller values resulted in better performance.

The tables below show the hyperparameters varied when generating candidate models for model selection. Table 14 reflects the hyperparameters when tree depth is determined via simulations internally (denoted CURT V1), while Table 15 reflects the hyperparameters when tree depth is explicitly considered in candidate model generation (denoted CURT V2).

Table 14. Search Grid Hyperparameters (CURT V1)

Model Type	Parameters Varied	Parameter Values Considered
CURT	Model for conditional survival function	Survival Tree – Log Rank Statistic
		Random Survival Forest
		Log-normal AFT
		Log-logistic AFT

Table 15. Search Grid Hyperparameters (CURT V2)

Model Type	Parameters Varied	Parameter Values Considered
CURT	Model for conditional survival function	Survival Tree – Log Rank Statistic
		Random Survival Forest
		Log-logistic AFT
		Log-normal AFT
	Cost-complexity parameter	0.000010, 0.000015, 0.000020, …, 0.000095
		0.000100, 0.000150, 0.000200, …, 0.000950
		0.00100, 0.00200, 0.00300, …, 0.01000

6.4 Random Survival Forests (RSF)

Random forests are an ensemble extension of decision trees. Random forests aim to improve upon decision trees through combining many decision trees constructed on resampled data. For a forest of m trees, the algorithm can be described at a high level as:

1. Sample n observations with replacement from the training data, where n is equal to the number of observations in the training data.
2. Construct a deep decision tree on the sampled observations. At each split in the decision tree, consider a random subset of size $mtry$ from the p available covariates. Only splits using these $mtry$ covariates are considered.
3. Repeat steps 1 and 2 until m trees have been constructed on m bootstrapped datasets.

Once the forest has been constructed, predictions for new data can be generated by passing each new data point through all m trees and averaging the predictions of each tree. The underlying principle of this ensemble technique relates to the high variance but

low bias of decision trees and the reduction in variance achieved by averaging many predictions. By averaging many decision trees, the result is a low variance and low bias ensemble model that may perform better than any single tree, albeit with a loss of interpretability. The use of a random subset of covariates at each split in each tree adds randomness to the construction process, reducing the issue of correlation between trees and consequently reducing the variability of the model.

Random survival forests, as described by Ishwaran et al. (2008), accommodate survival data by using survival trees in the ensemble. Beyond the use of survival trees with suitable splitting functions rather than regression or classification trees, the construction of the ensemble proceeds in the same manner as usual. Ishwaran et al. (2008) distinguished their random survival forest from previous work applying random forests to survival data through considering all observations for sampling and tree construction. In particular, this is distinct from the work of Hothorn, Bühlmann, et al. (2006), who used a weighting approach that meant censored observations were not included in the construction of any trees.

When generating predictions from the random survival forest, the cumulative hazard function of the ensemble is used. The Nelson-Aalen estimator for the cumulative hazard function (CHF) is given by:

$$\widehat{\Lambda}(t) = \sum_{j;t_{(j)} \leq t} \frac{d_{(j)}}{r_{(j)}} \tag{21}$$

where $t_{(j)}$ represents the j-th ordered event time, $d_{(j)}$ is the number of events at time $t_{(j)}$ and $r_{(j)}$ is the number of observations at risk at time $t_{(j)}$. The cumulative hazard function is estimated in this way separately for each terminal node in each tree in the ensemble. That is, for a given tree and given terminal node, only the data used in constructing the tree (the bootstrapped dataset) that also fell within that terminal node are used in constructing the CHF. To produce a prediction for a new data point, it is first passed through all m trees and thus associated with m terminal nodes, one for each unique tree.

The CHF of the m terminal nodes is then averaged to produce an ensemble CHF which applies to the new observation.

More formally, let $\widehat{\Lambda}_k(t|x_i)$ represent the conditional CHF for survival tree k for observation i with covariate vector x_i. Then, the ensemble cumulative hazard is given by:

$$\widehat{\Lambda}_{ensemble}(t|x_i) = \frac{1}{m}\sum_{k=1}^{m}\widehat{\Lambda}_k(t|x_i) \tag{22}$$

From the ensemble CHF, it is simple to calculate survival as:

$$\hat{S}(t|x_i) = \exp\left(-\widehat{\Lambda}_{ensemble}(t|x_i)\right) \tag{23}$$

When implementing RSFs in this project, the randomForestSRC package was used. This was written by Ishwaran and Kogalur (2021), who are both authors of the article which first set out random survival forests as considered here. Additionally, while random forest performance is usually validated using out-of-bag estimates in which data not sampled in the bootstrapping is treated as a validation set, performance of the RSF models in this work are assessed using five-fold cross-validation. This is done to promote comparability between models as per the guiding principle set out in Section 3.4 by ensuring performance is measured in the same manner for all modelling techniques.

6.4.1 Decisions in RSF Construction

While a range of parameters can be adjusted in the randomForestSRC implementation, including the sampling scheme for constructing trees and whether random splitting is used, the primary parameters of interest were the following:

- The number of survival trees in the ensemble
- The number of covariates considered at each split
- The number of observations in terminal nodes

All trees were constructed using the log-rank splitting function.

Number of Trees

The number of trees in a random forest is treated as a hyperparameter in this project, though there are mixed views as to whether it should be varied or simply set to a sufficiently large value, potentially driven by what is computationally feasible (Probst & Boulesteix, 2017). The latter view is motivated by the fact that each new tree fitted on the dataset is independent of previous trees and so risk of overfitting is not increased, in contrast with methods like boosting where tree fitting is sequential. The choice to vary forest size in this work was driven first by a need to establish whether the ensemble was sufficiently large to extract the full benefit of averaging many trees. That is, when considering several forests of similar magnitude, performance should stabilise at higher values as the upper bound of variance reduced is reached. The second factor in the decision is that despite a larger forest being better in theory, there are cases in practice where performance increases before decreasing as a function of forest size (Probst & Boulesteix, 2017). Lastly, while it can be argued that larger forests are always better, varying the parameter should not impede the selection of a final model. Even in the case that larger forests truly are always better, this would simply mean the largest model is selected. Ignoring the parameter in the case that it should be varied would be expected to result in a worse final model selected.

For model selection, forest sizes of 500, 750, and 1000 were considered.

Covariates Considered at each Split

At each split in each tree constructed as part of the overall forest, candidate splits are considered only for a random subset of all available covariates. This is driven by the inability of trees to consider future potential splits when determining the best split at a specific node and the inability to go backwards after the split is made to determine whether an alternative split would have improved the final model. Adding a stochastic element to the selection of splitting rules increases the variety of trees produced. The consequent reduced correlation between trees also lowers the variability of the final ensemble further. While a common default for this parameter is to set it equal to the square root of the number of covariates, the optimal value depends on the problem (Hastie, Tibshirani, & Friedman, 2009) and so it is treated as a hyperparameter in this project. For model selection, the number of covariates considered for splitting at each point ranged from 1 to 8. This includes the common default, which would be $\sqrt{19} \approx 4$.

Terminal Node Size

Random forests are typically made up of many high-complexity trees. As mentioned in Section 6.2, the minimum node size parameter controls the minimum number of observations that should be in a terminal node. Given the tendency for random forest to perform well when using deeper and thus more complex trees, a terminal node size of three was trialled in addition to the default value of 15. This means that for a node to be split, its current size must be at least twice the value specified. The value of three was used to assess the effect of further increasing tree depth and thus complexity from the default of 15.

The parameters varied and the considered values for the RSF model are shown in Table 16.

Table 16. Search Grid Hyperparameters (Random Survival Forest)

Model Type	Parameters Varied	Parameter Values Considered
Random Survival Forest	Number of trees	500
		750
		1000
	Covariates considered at each split	1, 2, 3, …, 8
	Terminal node size	3
		15

6.5 Censoring Unbiased Regression Ensembles (CURE)

Censoring Unbiased Regression Ensembles, referred to hereafter as CUREs, are the ensemble extension of CURTs set out by Steingrimsson et al. (2019) and share the same underlying principles in their use of CUTs. They are analogous to other tree ensemble methods, such as random forests or bagging. Drawing on the algorithm described in Section 6.4, the CURE algorithm differs primarily in that CURTs are constructed in step 2 rather than more traditional decision trees. For clarity, the algorithm for CURE is provided as detailed on page 372 of Steingrimsson et al. (2019), with modification to reflect censoring:

1. *Generate M independent sets of exchangeable bootstrap weights $w_1, \ldots, w_n$.*

2. *For each set of bootstrap weights, build a fully grown* CURT with a CUT of the full data loss function *where, at each stage of splitting,* $mtry$ *covariates are randomly selected from the* p *available covariates for candidate splits.*
3. *For each tree in the forest, calculate an estimator at each terminal node and average over the results obtained for the* M *sets of bootstrap weights to get the final ensemble predictor.*

The link between this algorithm and random forests is clear where a non-parametric bootstrap is used in combination with $mtry < p$, as is the case in this work.

The above algorithm describes the general class of models considered as CUREs, with specific implementation requiring choices as to bootstrapping, value of $mtry$, and the CUT of the loss function. In this project, non-parametric bootstrapping will be used, $mtry < p$ will be considered with the specific value considered as a hyperparameter, and the doubly robust CUT of equation (19) is used for constructing CURTs. The use of non-parametric bootstrapping corresponds to the default choice for random forest methods. The doubly robust CUT is selected for two reasons. First, it has more attractive properties than the most obvious alternative in that it is expected to be more efficient compared to IPCW and only one of the censoring and event survival functions needs to be correctly specified for the estimator to be a CUT (Steingrimsson et al., 2019). Secondly and more practically, constructing the CURTs in the ensemble with the additional random element involving consideration of a random subset of the p covariates make implementation using existing R software (namely the rpart package) more difficult. Related to this implementation difficulty, an imputation procedure was developed by Steingrimsson et al. (2019) such that existing forest implementations can be applied to the imputed data to achieve an equivalent result in the case of a doubly robust CUT. This equivalence does not hold for the IPCW CUT of the loss function. For details regarding the response imputation procedure, see Steingrimsson et al. (2019).

As with the CURT model, R code was provided by the authors of the original article outlining the CURE model. This code used the randomForest package (Liaw & Wiener, 2002) to construct the ensemble on the data after imputation of the response was performed. As with the CURT model, modifications were made to the provided code to allow for extraction of survival curves rather than estimates of event times. To do this, trees in the ensemble were constructed individually. For each tree, survival curves were

computed for validation or test observations based on Kaplan-Meier curves fitted to training data observations falling into each terminal node. These survival curves were averaged across all trees in the ensemble to produce the final survival curves.

6.5.1 Decisions in CURE Construction

In constructing a CURE model, given that a doubly robust CUT of squared error is used, four decisions are of interest. These are:

1. The censoring and conditional survival functions
2. The number of CURTs in the ensemble
3. The number of covariates considered at each split.
4. The number of observations in terminal nodes

The first decision is tied to the use of the doubly robust CUT as an objective function in each CURT model. The explanation of this decision is identical to that set out for the CURT model and so is not repeated here. The censoring survival function will be estimated using the Kaplan-Meier estimator, while the model used to estimate the conditional survival function will be treated as a hyperparameter in the CURT model. The candidate models consist of a random survival forest, which was implemented in the provided code, and a survival tree, which was implemented by the author. These models for the conditional survival function used the randomForestSRC and rpart packages as described in Section 6.4 and Section 6.2 respectively.

The second, third and fourth decisions are those associated with random forests more generally. As these were described in Section 6.4.1 and to avoid repetition, only the ranges of hyperparameter values considered are presented here. CURE is a more novel technique with less research available to inform the choice of reasonable hyperparameter values. The greater uncertainty motivates a greater number of trialled values. A wider range of values for the number of trees is thus considered compared to the random survival forest. Similarly, a slightly wider range of terminal node sizes was considered.

The hyperparameters varied and the considered values for the CURE model are shown in Table 17.

Table 17. Search Grid Hyperparameters (CURE)

Model Type	Parameters Varied	Parameter Values Considered
CURE	Model for conditional survival function	Survival Tree – Log Rank Statistic
		Random Survival Forest
	Number of trees	100
		250
		500
		750
		1000
	Covariates considered at each split	1, 2, 3, …, 8
	Terminal node size	3
		10
		20

6.6 Recursively Imputed Survival Trees (RIST)

Put forward by R. Zhu and Kosorok (2012), Recursively Imputed Survival Trees (RIST) are another tree-based ensemble approach to modelling survival data. They are similar to random survival forests in structure in that they generate M datasets by bootstrapping the original data and constructing a survival tree on each generated dataset. The two main differences between RIST models and RSF models are:

1) RIST models increase randomisation beyond that of an RSF by constructing M extremely randomised trees (ERTs) (Geurts, Ernst, & Wehenkel, 2006) on the entire training data. Where survival trees in an RSF consider a random subset of the p covariates at each split, ERTs take this further and consider random split points for the considered covariates.
2) RIST models aim to extract additional information from censored observations through an iterative process of model construction and imputation.

Building on these points of difference, the RIST model building procedure can be broken into two broad steps. The first step consists of constructing the ensemble of ERTs for survival data (in the same way as standard survival trees). The second step imputes the censored observations conditional on their censoring time and survival function output from the model in the first step. That is, for a given censored observation, an event time is imputed between the censoring time and the maximum follow-up time using random

sampling from the conditional survival distribution output from the model for that observation. This conditional imputation is repeated to generate M imputed datasets to be used in the next application of step 1. Both steps are repeated a pre-specified number of times, with step 1 using the original data for the first iteration and the results of the previous iteration after this point.

Letting τ represent the maximum follow-up time, the exact RIST model algorithm as set out in Table 2.1, page 16 of R. Zhu (2013) is as follows:

1. ***Survival tree model fitting****: Generate M extremely randomised survival trees for the raw training data set under the following settings:*
    a. *For each split, K candidate covariates are randomly selected from p covariates, along with random split points. The best split, which provides the most distinct daughter nodes, is chosen.*
    b. *Any terminal node should have no less than $n_{min} > 0$ observed events.*
2. ***Conditional survival distribution****: A conditional survival distribution is calculated for each censored observation.*
3. ***One-step imputation for censored observations****: All censored data in the raw training dataset will be replaced (with correctly estimated probability) by one of two types of observations: either an observed failure event with $Y < \tau$ or a censored observation with $Y = \tau$*
4. ***Refit imputed dataset and further calculation****: M independent imputed datasets are generated according to 3, and one survival tree is fitted for each of them using 1.a and 1.b*
5. ***Final prediction****: Steps 2-4 are recursively repeated a specified number of times before final predictions are calculated.*

This model has been reported to exhibit better performance than the Cox model, a random forest based on an IPCW CUT, and the RSF model described in Section 6.4 on a range of simulated and clinical datasets (R. Zhu & Kosorok, 2012).

For implementing this model, R code was obtained from the website of the primary author of the article proposing the model at https://sites.google.com/site/teazrq/software. This implementation allowed only for numeric covariates and thus all data were converted to numeric before application of the model.

6.6.1 Decisions in RIST Construction

For implementing the RIST model, four hyperparameters were varied:

- The number of trees
- The number of covariates considered at each split
- The minimum number of observed failures in each terminal node
- The number of imputation cycles

Number of Trees

The number of trees in the ensemble has been discussed previously as a hyperparameter for random survival forests. For RIST, this hyperparameter takes smaller values for RIST because of the additional variability in ERTs. In the original paper, only 50 trees were considered, though this is varied up and down as part of this work.

Covariates Considered at each Split

The number of covariates considered at each split for each tree has also been discussed in the context of random survival forests, though for RIST the hyperparameter affects the construction of ERTs rather than survival trees. This hyperparameter has not been varied in other work by the authors who originally proposed the method (R. Zhu, 2013; R. Zhu & Kosorok, 2012), but is varied here to remain consistent with decisions made for the random survival forest models. This promotes comparability of the results. The RIST model does, however, have long running times[4] and an additional hyperparameter, and so a coarser range of values is considered for this hyperparameter, specifically values of three, five and seven. The central value of five was determined as the square root of the number of covariates (19) rounded up.

Terminal Node Size

Related to the number of trees in the ensemble, the minimum number of observed failures in each terminal node was also treated as a hyperparameter. It has increased relevance for RIST models compared to the equivalent hyperparameter in RSFs as a result of the smaller ensemble size. By using fewer trees, each individual tree and their construction

[4] For example, setting the number of trees to 50, number covariates considered at each split to 5, terminal node size to 40, and imputation cycles to 2, the RIST model took 2.93 hours to train when using 80% of the training data (4/5 cross-validation folds). This was recorded when using a machine with an Intel® Core™ i9-9900K CPU @3.6GHz process. The coarser hyperparameter settings described in this section still require the construction of 1350 such RIST models.

parameters have greater influence. This hyperparameter was fixed at six in the original paper which considered datasets with fewer than 1000 observations, but larger values were considered in this project given the much larger datasets involved.

Imputation Cycles

Finally, the number of imputation cycles refers to the number of times the model fitting and imputation elements of the algorithm are repeated. R. Zhu and Kosorok (2012) reported that most of the value of the procedure is found in the first few iterations, with potentially incremental improvements afterwards. An explicit recommendation for not using more than five iterations was also provided.

The set of hyperparameter values considered for RIST in this project are shown in Table 18.

Table 18. Search Grid Hyperparameters (RIST)

Model Type	Parameters Varied	Parameter Values Considered
RIST	Number of trees	30
		40
		50
		60
		70
	Covariates considered at each split	3
		5
		7
	Terminal node size	10
		20
		30
		40
		50
		100
	Imputation cycles	1
		2
		3

6.7 Bayesian Additive Regression Trees (BART)

Bayesian Additive Regression Trees, or BART, is a Bayesian model proposed by Chipman, George, and McCulloch (2010) that is based on an ensemble of trees. This

model was originally proposed in the context of regression and classification but has also been extended to survival data (Bonato et al., 2011; R. A. Sparapani et al., 2016). Bonato et al. (2011) extended the BART model to survival data under parametric or semi-parametric assumptions. R. A. Sparapani et al. (2016) relaxed the need for these assumptions in their proposed model which centres on treating the survival problem in a discrete-time setting.

The core BART model can be presented as a summation of m trees, each of which has two components: the tree structure A and terminal node values N. For an outcome y, we can express the general BART model as the following:

$$y = f(x) + \epsilon \tag{24}$$

$$y = \sum_{j=1}^{m} g(x; A_j, N_j) + \epsilon \tag{25}$$

To make this a Bayesian expression, priors are assumed for the tree structure and terminal node values. The prior for the tree structure has elements for whether a given node is internal or terminal, the probability of each covariate being used in a split, and the split for a given covariate. The latter two elements are both assumed to be uniform, while the probability of a node being internal is given by the expression $\alpha(1 + depth)^{-\gamma}$, where $depth$ is the depth of the node, $0 < \alpha < 1$, and $\gamma \geq 0$. For further details regarding the method, including the N prior, see Chipman et al. (2010).

To adapt the original BART model to survival data, R. A. Sparapani et al. (2016) took a discrete time approach in which subject outcomes are considered at all unique follow up times until the subject-specific event or censoring time. Using an example from R. A. Sparapani et al. (2016), let Table 19 represent a survival data set of only three observations:

Table 19. Example Survival Data (modified from R. A. Sparapani et al. (2016))

Subject ID	Covariates				Follow up time	Event
1	…	…	…	…	2.5	1
2	…	….	…	…	1.5	1
3	…	…	…	…	3.0	0

From this table, we have unique follow up times (1.5,2.5,3.0). The dataset is then transformed by repeating each observation for all unique times before the actual follow-up time with indicator equal to 0. This then produces a dataset in which time becomes an input to a model and the outcome is binary. The transformed data set can be seen in Table 20 below.

Table 20. Discrete Time Transformed Example Survival Data (modified from R. A. Sparapani et al. (2016))

Subject ID	Covariates				Time	Event
1	…	…	…	…	1.5	0
1	…	….	…	…	2.5	1
2	…	…	…	…	1.5	1
3	…	…	…	…	1.5	0
3	…	….	…	…	2.5	0
3	…	…	…	…	3.0	0

As with the original BART method, further technical details of the BART model for survival data are not stated here to avoid a lengthy and unoriginal description. The interested reader should instead see R. A. Sparapani et al. (2016). For this work, the BART model for survival data was implemented through the associated BART R package (R. Sparapani, Spanbauer, & McCulloch, 2021).

6.7.1 Decisions in Survival BART Construction

The R implementation of BART for survival data allows for varying a range of hyperparameters in model construction. Previous research has, however, found that the BART model exhibits excellent performance on many problems using the default hyperparameter without need for extensive tuning. Accordingly, the default hyperparameter settings are used in this project with one exception (described below). This is also linked to more practical considerations, as the model is extremely

computationally intensive in terms of memory and time for datasets as large as those considered here (see Table 23). These considerations are also relevant for practical implementation of this model in a hospital setting, where computational resource constraints are present. The key hyperparameters affecting model construction in this work are the following.

- The number of trees in the ensemble
- The number of draws from the posterior distribution returned
- The number of periods to treat as the 'burn in' sample
- The degree of thinning to use when returning draws from the posterior

These parameters were considered when fitting the BART model as they directly affect the running time and size of the resulting model. Other than the number of draws from the posterior distribution returned, all hyperparameters were set to the recommended defaults for both hospitals. The ensemble model consisted of 50 trees. For the number of periods treated as burn-in, this parameter was set to 250, meaning the first 250 MCMC iterations were discarded. This parameter affects only the computational time involved in constructing the model. For the degree of thinning, the $keepevery$ parameter was set to 10, meaning only every 10^{th} draw from the posterior was returned. This reduces correlation between draws as well as reducing the size of the final model object.

The number of draws from the posterior distribution defaulted to 1000 in the used R implementation. This, in conjunction with the other default parameters, was not possible with the available computational resources. This parameter was thus reduced for both datasets until model construction became feasible. For RH, being the smaller dataset, this parameter was set to 500. For GCUH, being the larger dataset, this parameter was set to 200.

For clarity, the hyperparameter values used for this model for Robina Hospital and GCUH are shown in Table 21 and Table 22. Further, the size of the resulting R objects and the time taken on a virtual machine with two Intel® Xeon® CPU E5-4640 @ 2.40GHz processors are also reported for each hospital in Table 23.

Table 21. Parameters used in the BART Model (GCUH)

Model Type	Parameter	Parameter Values Considered
BART (GCUH)	Number of trees	50
	Draws from the posterior	200
	Burn-in sample	250
	Thinning	10

Table 22. Parameters used in the BART Model (RH)

Model Type	Parameter	Parameter Values Considered
BART (RH)	Number of trees	50
	Draws from the posterior	500
	Burn-in sample	250
	Thinning	10

Table 23. BART Model Object Sizes and Run Times

Hospital	Model Construction		Train Predictions		Test Predictions	
	Time	Size	Time	Size	Time	Size
GCUH	39.43 hours	41 Gb	3.66 hours	110.3 Gb	1.56 hours	47.3 Gb
RH	34.36 hours	33.8 Gb	6.89 hours	138.8 Gb	2.94 hours	59.5 Gb

6.8 Multiple Time Point ANNs

One approach to adapting neural networks to the censored data characterising survival analysis can be broadly referred to as multiple time point (MTP) ANNs. These approaches capture the temporal element of predictions using multiple output nodes predicting survival status or hazard and by discretising time into intervals. A variety of treatments for censored observations has been considered in these approaches, including imputing censored observations outcomes in later intervals. Alternatively, the objective function can be defined such that output neurons dealing with outcomes post-censoring do not contribute and thus do not influence the model training process (as in the definition of the objective function in the following paragraphs). Beyond the variation in what predictions

represent and how censored observations are accommodated, the usual network decisions remain relevant in terms of architecture, regularisation, batch size and epochs for training.

For this project, a recent implementation of an MTP ANN by Gensheimer and Narasimhan (2019) will be used, referred to hereafter as NNET Survival. Code was obtained from the corresponding GitHub repository https://github.com/MGensheimer/nnet-survival. This implementation provides a starting point for model construction in terms of the objective function used for network training and aspect of survival predicted. The underlying objective function dictating the interpretation of predictions is now detailed, after which more traditional ANN design decisions are described. The primary result is that the objective function of NNET Survival is equivalent to a statistical model's negative log likelihood. In this objective function, observations are considered only up to the time of censoring or event occurrence and accordingly model predictions can be interpreted as conditional hazards for each interval.

6.8.1 NNET Survival

To describe the approach of the NNET Survival model, the likelihood function of Gensheimer and Narasimhan (2019) is shown. Motivating the likelihood function, consider τ discrete intervals of time with discrete conditional hazard rates $h_l(x_i)$ for each observation i and interval l. The probability of a person's remaining event-free up to the end of interval k can be expressed as:

$$S_k(x_i) = \prod_{l=1}^{k} \big(1 - h_l(x_i)\big) \tag{26}$$

Using this probability statement, the likelihood contribution of a single observation i where the event was observed in interval k is:

$$L_{event,i} = h_k(x_i) \prod_{l=1}^{k-1} \big(1 - h_l(x_i)\big) \tag{27}$$

While the contribution of uncensored observations to the likelihood is straightforward, the appropriate contribution of censored observations depends on when censoring occurred within an interval. Simply assuming all censored observations were event-free until the end of the censoring interval would bias the final model, as would ignoring this interval entirely (Brown et al., 1997; Gensheimer & Narasimhan, 2019). To avoid this, for a given observation censored in the second half of interval $k-1$ or first half of interval k, the likelihood is the probability of surviving the first k intervals. This assumes that observations censored in the latter half of an interval survived the remainder of the interval but it does not assume this for observations censored in the first half of the interval. To formalise this for observations more generally, let $t_1^\tau, t_2^\tau, \dots, t_\tau^\tau$ represent the upper limits of the τ intervals and $t_0^\tau = 0$. Further, quantities are defined indicating whether an observation contributes to the likelihood for a given interval and indicating whether the event occurred in a given interval:

$$surv_s(l,i) = \begin{cases} 1 & if\ \delta_i = 1\ \&\ y_i \geq t_l^\tau \\ 1 & if\ \delta_i = 0\ \&\ y_i > 0.5(t_{l-1}^\tau + t_l^\tau) \\ 0 & otherwise \end{cases} \tag{28}$$

$$surv_f(l,i) = \begin{cases} 1 & if\ \delta_i = 1\ \&\ t_{l-1}^\tau \leq y_i \leq t_l^\tau \\ 0 & otherwise \end{cases} \tag{29}$$

Using these quantities, the likelihood contribution of any censored or uncensored observation across all intervals is given as:

$$L_i = \prod_{l=1}^{\tau} \left(1 - h_l(x_i)\right)^{surv_s(l,i)} h_l(x_i)^{surv_f(l,i)} \tag{30}$$

The log-likelihood is similarly given as:

$$loglik_i = \sum_{l=1}^{\tau} surv_s(l,i)\ln\big(1 - h_l(x_i)\big) + surv_f(l,i)\ln\big(h_l(x_i)\big) \tag{31}$$

Finally, with an ANN predicting conditional survival probability for covariate vector x_i and interval l, denoted by $surv_{pred}(l,i)$, the log-likelihood over all n observations is given as:

$$\begin{aligned} loglik = &\sum_{i=1}^{n}\sum_{l=1}^{\tau} surv_s(l,i)\ln\Big(surv_{pred}(x_i,l)\Big) \\ &+ surv_f(l,i)\ln\Big(1 - surv_{pred}(x_i,l)\Big) \end{aligned} \tag{32}$$

This final log-likelihood serves as the objective function used in training of ANNs with τ output nodes. The contribution of any single observation i to the log-likelihood can be computed using (31) and thus mini-batch processing can be used. Next, the focus turns to how intervals are defined before considering more general decisions in ANN construction.

6.8.2 Defining Time Intervals

In the proposal of NNET Survival, Gensheimer and Narasimhan (2019) investigated the effect of different definitions for time interval widths with simulated data. Widths can be of constant sizes, such as each representing a year or a month, or variable, such as increasing width with follow-up time. In their investigation, Gensheimer and Narasimhan (2019) reported model performance in terms of Harrell's C-index as insensitive to the different interval definitions. Given the lack of sensitivity and no clear recommended definition, time intervals are defined in this project to reflect the problem being modelled. In modelling readmissions, there is disproportionate interest in the risk of readmissions soon after discharge compared to much later, meaning there is a greater need for more granular predictions in this region of the time axis. Readmission events are also more frequent at earlier times.

Time intervals in this project, for a given number of intervals, are defined such that each interval has an approximately equal number of observed events. This approximate

equality should be emphasised, as follow-up time in this work is measured in whole days rather being a truly continuous variable. Despite this, differences in the number of events between intervals are not expected to cause a material difference to subsequent models, given the consistent findings Gensheimer and Narasimhan (2019) across different interval definitions.

The following figures serve to illustrate this approach to defining intervals. When defining 30 intervals using the training data of RH with an approximately equal number of events in each, Figure 4 demonstrates the number of events in each interval against interval ID and interval width. Figure 5 presents interval width against interval IDs, illustrating the reduced density of events as time since discharge increases resulting in wider later intervals.

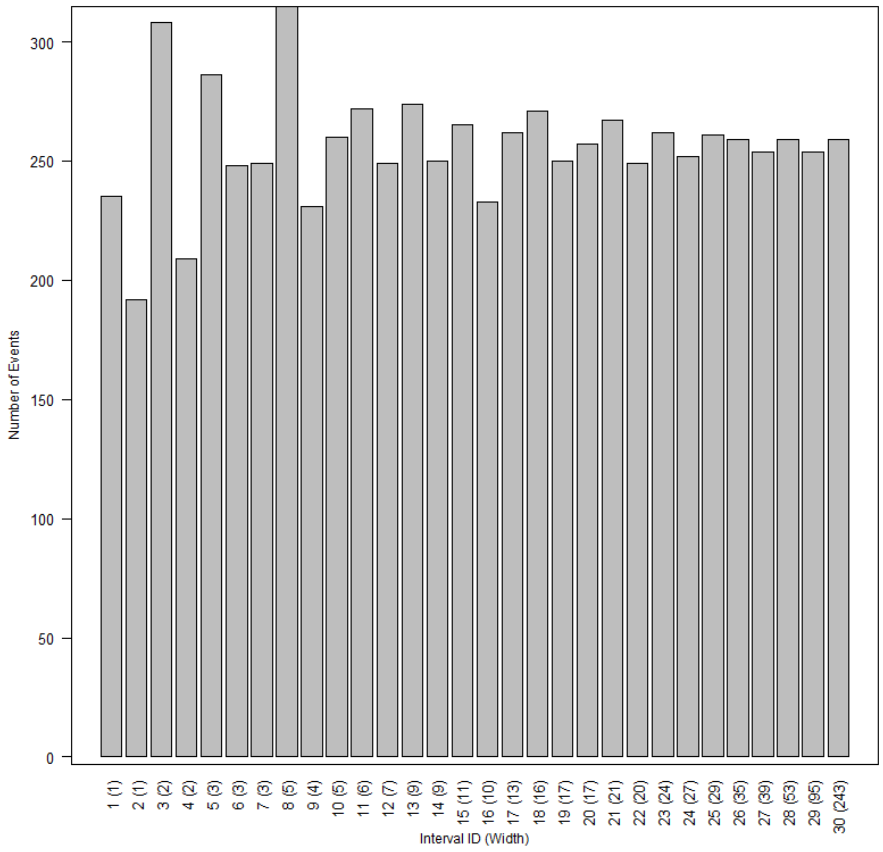

Figure 4. Number of Events in Each Interval

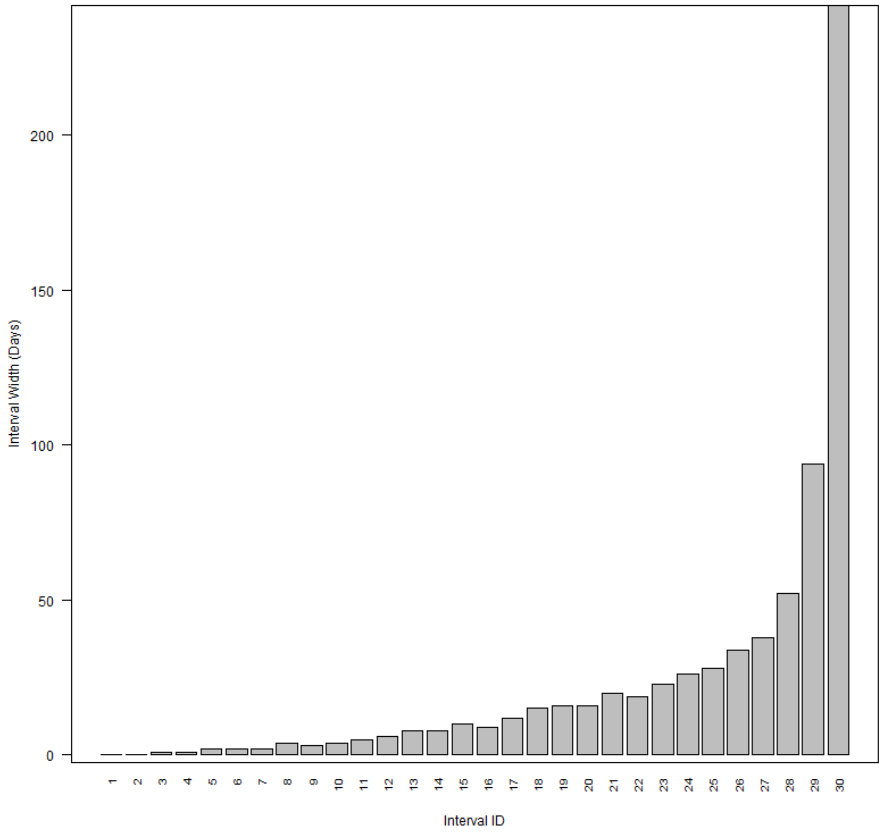

Figure 5. Increasing Interval Width with Later Intervals

The above figures demonstrate several relevant features of the approach taken to interval definitions. Firstly, given the use of integer days for follow-up time, earlier intervals are less even with respect to the number of events included as the numbers do not neatly correspond to the desired proportions. There is limited scope to correct these imbalances given that these early intervals are very narrow, encompassing only a few days. This is a desirable feature of interval definitions for readmission prediction, as it is these earlier intervals where attention is focused. While survival probabilities at any time point can and will be extracted by interpolating between predictions at interval end points, reducing the range over which it is required mitigates the error introduced by interpolation. The latter figure better illustrates how interval width increases dramatically in much later periods, reflecting that once patients are discharged and readmission-free for a long period then they are likely to remain as such.

While the above figures were based on 30 intervals, other values will be considered in this project for this model. The number of intervals influences the complexity of the network as it is equal to the number of output nodes. Specifically, an increased number

of output nodes increases granularity and complexity of the fully connected network. Too few output nodes may mean that the model is not flexible enough to be useful, while too many will cause a model to overfit. To assess the influence of different numbers of intervals in conjunction with other hyperparameters (described below), 20, 30 and 40 were trialled. The lower value was used by Gensheimer and Narasimhan (2019) in their implementation, but a larger quantity of data is available for this work which motivated consideration of 30 and 40 intervals as well.

6.8.3 Decisions in ANN Construction

The remaining decisions to be made for model implementation are those generally associated with ANN construction. Neural networks afford a large degree of customisability through many hyperparameters. For this work, five components (including number of intervals) are considered and varied:

1. The number of hidden layers and nodes
2. The number of epochs in training
3. Mini-batch size
4. Regularisation
5. The number of intervals

The fifth component was the subject of previous paragraphs and is not described further here. The range of values considered for each hyperparameter in the search grid are presented at the end of this section in Table 24.

Number of Hidden Layers and Nodes

The number of hidden nodes and layers in a model corresponds to the model's capacity for complex relationships. As the number of hidden layers and hidden nodes increases, the model can capture increasingly complex relationships, but the likelihood of fitting random noise (overfitting) also increases. The goal is to find a model which is complex enough to capture the true relationships in the data, but not so complex that it also captures enough random noise to offset this benefit. Further, the greater the number of hidden layers in a network, the greater its ability to extract latent features in the data. This is particularly relevant for tasks such as image recognition, where the original data (pixels) are far removed from the actual elements used to recognise images (edges, texture, objects). For this task, models involving more than two hidden layers were not considered

because such deeper models (those with more hidden layers) are primarily valuable for highly abstract tasks such as image recognition.

Epochs and Mini-Batch Size

The second component is the number of epochs used in training the neural network, and is closely tied to the third, which is the batch size used in training. To describe the effect of these hyperparameters, a very high-level description of the network training process is required. The process of training a network involves passing the data forward through the network's layers, evaluating the outputs with an objective function, and then using gradient descent methods to pass backwards through the network to update the weights involved. This process of updating is referred to as gradient descent or backpropagation. In the case that there is no mini-batch processing, then the gradient descent weight updates occur after the entire data set has been passed forwards through the network. That is, there is one update of the weights per epoch, with an epoch involving the entire dataset being passed through the network. This can mean that it takes a very long time to train the model for larger datasets. A faster approach, with an added benefit of a reduced ability to overfit, is to use mini-batch processing. This involves the consideration of data in multiple mini-batches with weight updates occurring after each mini-batch of data is passed forward through the network. For example, if a mini-batch size of 100 is used and the training data consist of 1,000 observations, then the weights will update 10 times for each epoch, corresponding to one update for each batch of 100 observations. There are two key benefits from the mini-batch process. First, it can be much faster to achieve a desired level of performance as a result of more frequent weight updates. Secondly, as different data are used in subsequent update steps, the ability of the network to model random noise across the entire dataset is reduced. A mini-batch which is too small, however, might not provide enough information for the model to properly train.

Regularisation

The final aspect of model development considered is the role of regularisation. Regularisation reduces the variability of a network by adding a penalty term to the objective function which penalises weights. In this work, L2 regularisation is used. This regularisation method imposes a penalty equal to the sum of the square of all weights in the network. For p weights in a network, the penalty can be expressed as:

$$Penalty = \lambda \sum_{j=1}^{p} w_j^2 \quad (33)$$

The magnitude of the penalty is controlled by the penalty parameter λ. For model selection in this project, it is this penalty parameter which was varied. The upper bound on penalty size of $\exp(-4)$ was based on initial results indicating that larger values prevented network training because weights were pulled towards zero.

Other Implementation Details

Several other details of model implementation were not varied but should be specified:

- Activation function – The ReLU activation function was used in the hidden layers of all candidate NNET Survival models, and the sigmoid activation function was used in the output layer to ensure predictions were in the range $[0,1]$. This is consistent with Gensheimer and Narasimhan (2019).
- Data Processing – All data were converted to a numeric format for model training and all covariates were standardised based on the entire training data for each hospital through the usual approach for networks: $x_{i,j}^* = \frac{x_{i,j} - \mu_j}{\sigma_j}$. The same standardisation parameters were used for training and test data.
- Dropout – dropout is a form of regularisation in which nodes in the hidden layer(s) of networks are systematically 'turned off' during training to encourage the network to spread relationships between nodes. Given the use of L2 regularisation, dropout was not considered.
- Interpolation – survival curves across all time points were extracted through linear interpolation of the survival function defined at interval end points.
- Variables – For each hospital, variables with low predictive value excluded from all logistic regression and Cox regression models were not used in ANN models. The variables included in ANNs for GCUH and RH can be found in Appendix A.
- Batch normalisation – Batch normalisation is a procedure in which the outputs of each hidden layer are standardised. This has been found to be useful for very deep models such as convolutional neural networks in computer vision applications but was expected to be less helpful for the relatively shallow networks considered for this model. Batch normalisation was not trialled because of its limited value for

shallow networks, and also to enable greater consideration of other hyperparameters. This was also informed by the limited benefit of batch normalisation for the Cox NNET model, described in Section 6.10, where this procedure was expected to have the greatest potential value.

The hyperparameter values considered for the NNET Survival ANNs in this work are presented in Table 24.

Table 24. Search Grid Hyperparameters (NNET Survival)

Model Type	Parameters Varied	Values Considered
NNET Survival	Hidden layers and nodes	1 layer, 5 nodes
		1 layer, 10 nodes
		1 layer, 15 nodes
		2 layers, 10 and 10 nodes
		2 layers, 15 and 10 nodes
	Epochs	100, 200, 300, …, 1500
	Mini-batch size	128
		256
		512
	Regularisation penalty (L2)	$\exp(-4)$
		$\exp(-5)$
		$\exp(-6)$
	Intervals	20
		30
		40

6.9 Time-Coded ANNs

A second and similar approach to adapting neural networks to survival data is in the form of time-coded ANNs. Like MTP ANNs, time-coded ANNs consider time in discrete intervals. Unlike MTP ANNs, however, time-coded ANNs incorporate time intervals as an input in a network architecture with a single output node. Predicted risk across time points is generated by repeating the covariate vector for each time interval with an additional covariate corresponding to the start of the current interval. The outcome

predicted by the network is the probability of event occurrence in the current interval conditional on having survived to the start of that interval.

6.9.1 Implementing a Time-Coded Model

No notable recent implementations of time-coded ANNs were identified in the review and so a custom implementation was used, though the core principle is based on the work of Biganzoli et al. (2002). To describe the model implementation, the objective function must first be outlined. Linked to the similarity in treatment of time, the initial likelihood formulation for time-coded models is identical to that presented in Section 6.8 and so is not repeated here. Instead, the description for time-coded ANNs starts at the final log-likelihood used in NNET Survival presented in (32) with the associated definitions of the $surv_s$, $surv_f$ and $surv_{pred}$ vectors set out previously. This log-likelihood expression applies to the time-coded ANN considered here as well as NNET Survival. The key practical difference for the objective function in a time-coded ANN is that, for a single input vector, predictions are generated for a single interval l and a single observation i, rather than for all intervals $l = 1, \dots, \tau$ for a single observation i. Accordingly, the log-likelihood specific to a single individual i and interval l is of interest and is obtained by reducing (32):

$$loglik_{i,l} = surv_s(l,i) \ln\Big(surv_{pred}(x_i, l)\Big) + surv_f(l,i) \ln\Big(1 - surv_{pred}(x_i, l)\Big) \tag{34}$$

For this equation, $surv_{pred}(x_i, l)$ can be replaced with $1 - \hat{h}(x_i, t_l^{\tau})$, reflecting that time-coded ANNs as considered in this work take time intervals as inputs and predict conditional hazard rather than conditional survival. Incorporating time intervals as inputs also allows for reformulation of the objective function in the below form:

$$loglik_{i,l} = surv_s(l,i) \ln\,(1 - \hat{h}_{ij}) + \big(1 - surv_s(l,i)\big) \ln\,(\hat{h}_{i,j}) \tag{35}$$

The equivalence of $surv_f(l,i)$ and $1 - surv_s(l,i)$ is linked to the consideration of time intervals as inputs in time-coded ANNs rather than outputs. In MTP ANNs, the objective

function is evaluated for all τ intervals, whereas in a time-coded ANN the objective function is only evaluated for intervals relevant to each observation. To illustrate this, let L_i be the maximum value of l for which $surv_s(i,l) + surv_f(i,l) = 1$. The objective function, evaluated over all observations and relevant intervals, can then be expressed as the full log-likelihood below:

$$loglik = \sum_{i=1}^{n} \sum_{l=1}^{L_i} surv_s(l,i) \ln\left(1 - \hat{h}_{ij}\right) + \left(1 - surv_s(l,i)\right) \ln\left(\hat{h}_{i,j}\right) \tag{36}$$

This expression is useful because minimising the negative log-likelihood for a given observation i and interval j is equivalent to minimising the cross-entropy loss function implemented in the Keras package (Chollet, 2015) for Python:

$$CrossEntropy = -(ylog(p) + (1-y)(1-p) \tag{37}$$

Given this equivalence, cross-entropy serves as the loss function used in the time-coded ANN in this work.

6.9.2 Data Preparation

With the objective function for training the network now established, attention now turns to preparing the data. To implement the time-coded model, the data are first transformed so that each row of the original data (corresponding to a single discharge) is duplicated up until the interval where the event occurred. In the case of censoring, rows are duplicated for all intervals before the one where censoring occurred. The row is also duplicated for the interval associated with censoring if censoring occurred in the second half of the interval. As a simple example, using intervals $(1, 2, 3)$ with end points of $(2.01, 4.01, 6.01)$, the example data in Table 25 is converted to the format in Table 26. The use of the decimal in the end point definition avoids ties between follow-up times and interval end points, made relevant in this work by follow-up time being recorded in whole days.

Table 25. Example Raw Data Format – Time-Coded Models

Discharge ID	Covariates				Time	Event
1	…	…	…	…	6	0
2	…	….	…	…	3	1
3	…	…	…	…	2	0

Table 26. Example Data Format – Time-Coded Models

Discharge ID	Covariates				Interval	Readmitted
1	…	…	…	…	1	0
1	…	….	…	…	2	0
1	…	…	…	…	3	0
2	…	…	…	…	1	0
2	…	….	…	…	2	1
3	…	…	…	…	1	0

After transforming the data, the remaining decisions are those typical of training an ANN. The additional decision specific to the time-coded ANN relates to the intervals used in the data transformation. No relevant guidance beyond problem-specific needs was identified in the relevant literature. Without any literature indicating otherwise, intervals are defined in the same manner as for NNET Survival such that each interval has approximately equal event numbers. Again, this results in the desirable property of greater granularity in earlier periods.

Unlike the MTP ANN, a larger number of intervals does not directly increase the complexity of the time-coded ANN as it does not increase the number of output nodes. More intervals do, however, increase the size of the transformed data, resulting in a larger memory requirement and longer training times. Increased training times reduce the range of models which can be practically considered. For this model, 40 intervals were used to balance granularity against computational requirements. More than 40 intervals would provide only marginal benefit in increasing the granularity of intervals, particularly given that intervals were defined such that earlier periods have shorter intervals. It would also reduce the range of other hyperparameters considered. Table 27 below demonstrates the result of transforming the training data for each hospital to the format required by time-coded models with 40 intervals. It should be noted that the increase in size when

generating predictions is larger, as generating predictions for each interval requires that the covariate vector be replicated for all intervals, rather than only up until the interval associated with readmission or censoring. Thus, when transforming data for the purpose of predictions rather than model training, the transformed data are 40 times larger than the original.

Table 27. Time-Coded Model Data Sizes (40 Intervals)

GCUH Training Data		RH Training Data	
Data	**Rows**	**Data**	**Rows**
Original	32,661	Original	16,783
Time-Coded Training	1,001,592	Time-Coded Training	503,417
Time-Coded Prediction	1,306,440	Time-Coded Prediction	671,320

6.9.3 Decisions in ANN Construction

Having selected the number of intervals, which is not varied, four hyperparameters were considered in network construction:

1. The number of hidden layers and nodes
2. The number of epochs in training
3. Mini-Batch size
4. Regularisation

Each of these hyperparameters and their role in neural networks has been described in Section 6.8.3 and so these descriptions are not repeated here. Instead, the differences between the time-coded ANN and NNET Survival search grids are briefly described.

Firstly, an additional architecture is considered involving 20 nodes in the single hidden layer. As each observation is now replicated, the variance within the data is greatly reduced. This is expected to increase the complexity required in a network to relate this reduced variability to outcomes and may advantage a larger single layer model.

Secondly, a wider range of epochs and batch sizes is considered. This is driven again by the increase in dataset size from replication of observations. The same batch size now makes up a much smaller proportion of the total data, and an even smaller proportion of the effective data resulting from duplication across intervals. As a result, training occurs

more slowly and requires more epochs. Larger mini-batch sizes are also considered to assess the result of faster learning with more information in each batch.

Finally, the number of intervals was not varied to allow greater variation in other hyperparameters. Unlike with NNET Survival, changing the number of intervals does not influence the network architecture and so is less relevant provided a sufficiently large number of intervals are used.

Other details of model implementation mirror those described for NNET Survival in Section 6.8.3 exactly. Data standardisation was done prior to data transformation to the time-coded format described above. The set of hyperparameter values considered for the time-coded ANNs in this work are shown in Table 28.

Table 28. Search Grid Hyperparameters (Time-Coded ANN)

Model Type	Parameters Varied	Values Considered
Time-Coded ANN	Hidden layers and nodes	1 layer, 5 nodes
		1 layer, 10 nodes
		1 layer, 15 nodes
		1 layer, 20 nodes
		2 layers, 10 and 10 nodes
		2 layers, 15 and 10 nodes
	Epochs	100, 200, 300, …, 1500
	Mini-batch size	128
		256
		512
		1,024
		2,048
		4,096
		8,192
	Regularisation penalty (L2)	$\exp(-4)$
		$\exp(-5)$
		$\exp(-6)$

6.10 Hybrid Cox-ANN Model

A final approach for adapting neural networks to survival data considered in this work is through their integration into Cox's Proportional Hazards model (Faraggi & Simon, 1995). This is done through replacing the linear predictor of the Cox model with the neural network and then deriving the network's weights by using the partial likelihood as an objective function. This combining of a neural network with a statistical model through

replacement of the linear predictor is analogous to the approach described for the time-coded ANN of Biganzoli et al. (2002). Repeating expressions from Section 2.2.1.3, the partial likelihood for the Cox model is given by:

$$PL(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} \right)^{\delta_i} \tag{38}$$

Letting $g(x_i, \theta)$ be the output of an artificial neural network with weights θ and input vector x_i, the partial likelihood for the adapted Cox model is given by the following:

$$PL(\theta) = \prod_{i=1}^{n} \left(\frac{\exp\big(g(x_i, \theta)\big)}{\sum_{j \in R(t_i)} \exp\Big(g\big(x_j, \theta\big)\Big)} \right)^{\delta_i} \tag{39}$$

6.10.1 Implementing a Hybrid Cox-ANN Model

The hybrid Cox-ANN approach has been employed several times since it was originally proposed (Faraggi et al., 1997; Mariani et al., 1997), with recent implementations incorporating more modern ANN designs but sharing the same underlying idea (Ching et al., 2018; Katzman et al., 2018). Of these, the model of Ching et al. (2018) available at https://github.com/traversc/cox-nnet and based on the Theano library (Theano Development Team, 2016) is used for this project, referred to hereafter as Cox NNET. The primary modification made to the original code was to allow for the use of batch normalisation in model training, which is discussed below.

One issue relevant to implementing models based on the Cox framework relates to the use of the partial likelihood as an objective function. Unlike other objective functions, evaluating the partial likelihood for a given observation i requires consideration of all other observations at risk at the last follow-up time of observation i. This precludes the use of mini-batches in model training because data outside of a given mini-batch are needed to evaluate the partial likelihood. Related to this, model training is extremely computationally time-consuming and memory-intensive. Gensheimer and Narasimhan (2019) noted that the maximum sample size for which they were able to construct the Cox

NNET model was 31,622 observations. This is because the implementation requires a matrix be saved with dimensions equal to $n \times n$, where n is equal to the sample size. They also found that the training time increased as a function of sample size more quickly for Cox NNET than for two other extensions of ANNs for survival analysis. While this is inherent to a model using the partial likelihood, ignoring approximation methods (Kvamme, Borgan, & Scheel, 2019), the consequence is simply that hyperparameter selection is less rigorous than for more computationally tractable models. This is expected to correspond to real-world implementations facing the same obstacles.

6.10.2 Decisions in ANN Construction

The decisions involved in fitting and selecting a Cox NNET model overlap with those outlined for multiple time point models and time-coded models, but some differences should be noted. Firstly, Cox NNET does not consider time in discrete intervals and so does not require modification to the original data beyond that relevant to networks in general (i.e., input data must be numeric and should be standardised). Secondly, because of the definition of the partial likelihood, mini-batch processing is not possible. Finally, unlike for previous ANN model types, batch normalisation was considered for Cox NNET, motivated by its semi-parametric nature and the additional benefit from faster learning given its long training times.

The final decisions to be made regarding the final Cox NNET model relate to the following:

- The number of hidden layers and nodes
- The number of epochs in training
- Regularisation
- Batch normalisation[5]

The first three aspects of ANN construction have been discussed previously and this discussion is not repeated here. Despite this, certain changes in the values considered do bear explanation. Firstly, slightly less complex architectures are considered than the previous two ANN techniques. This is due to the reduced complexity inherent in the ANN given that only one output node is used, time is not an input, and there is no data

[5] Asymmetry between MTP ANNs, time-coded ANNs, and Cox NNET with respect to consideration of batch normalisation is related to asynchronous model training. As batch normalisation had the most relevance for Cox NNET but did not materially change performance, it was not considered in later model selection for other ANN techniques.

duplication. With respect to the sparser range of epochs considered, this was due to the longer run times of the model and its implementation in Theano making saving intermediate models more difficult. That is, unlike models implemented in the Keras architecture, the implementation of Cox NNET did not have a clear mechanism for saving intermediate models. Consequently, two models differing only in terms of epochs are built entirely separately. One advantage of this relative to previous models is that this provides more information regarding the influence of random starting points which is otherwise difficult to assess for large models. The disadvantage, however, is that generating models for the same range of epoch values is much more time consuming. For the omission of L2 penalties equal to $\exp(-3)$, this was driven by preliminary results indicating that this penalty (and larger penalties) prevented reasonable model training.

For the final aspect of model construction, batch normalisation was considered for two reasons. Firstly, data distribution is more important for semi-parametric techniques such as Cox regression which Cox NNET is derived from. Secondly, batch normalisation can, in some cases, reduce time taken to achieve a given performance level. This was relevant given the long training times encountered in this work and reported elsewhere (Gensheimer & Narasimhan, 2019). When using batch normalisation, the output of each hidden layer is standardised with the usual $\frac{x-\mu}{\sigma}$ transformation before being passed to the subsequent weights and activation function. Batch normalisation was not originally allowed for in the code used by Ching et al. (2018). Thus, it was modified as part of this work. Models constructed with and without batch normalisation to assess the associated effect.

An additional difference to previous models in hyperparameters is in the activation function. To remain faithful to the implementation of Ching et al. (2018), the hyperbolic tangent activation function was used rather than ReLU. Other details of model implementation related to data processing and variables are identical to those described for MTP ANNs in Section 6.8.3.

The set of hyperparameter values considered for Cox NNET in this project are shown in Table 29 below.

Table 29. Search Grid Hyperparameters (Cox NNET)

Model Type	Parameters Varied	Values Considered
Hybrid Cox-ANN (Cox NNET)	Hidden layers and nodes	1 layer, 8 nodes
		1 layer, 14 nodes
		1 layer, 21 nodes
		2 layers, 5 and 5 nodes
		2 layers, 7 and 4 nodes
	Epochs	50
		100
		200
		500
		600
		1000
	Regularisation penalty (L2)	$\exp(-5)$
		$\exp(-6)$
	Batch normalisation	Yes
		No

7 Results and Discussion

A range of results are associated with the data and methodology described in preceding chapters. Of these, the performance of the final models associated with each technique for each hospital is of primary interest. This model performance directly addresses the research questions and is thus the focus of this chapter. Additional results relate to the model selection process, specifically the cross-validated performance of different hyperparameter values and which hyperparameter values were used for constructing the final models of this work. These results are ancillary as they do not directly address the research questions. They are thus provided in the Appendi. The final model settings used for each technique on each hospital and for each research question are presented in Appendix B. Visualisations of model performance across the hyperparameter settings outlined in Chapter 6 are presented in Appendix C. This is done for each technique, each hospital, and each of the model selection performance metrics (AUC and IBS) corresponding to the two research questions. For consistency, visualisations are also included for ANN models despite the large number of hyperparameters greatly limiting their interpretability.

The remainder of this chapter outlines and discusses the results of this work for each research question in terms of final model performance.

7.1 Results for RQ1

RQ1: Can machine learning survival techniques improve upon statistical survival models when predicting 30-day hospital readmissions?

As outlined in Section 5.2.2, the primary measure of performance for RQ1 is discrimination as measured by AUC, with calibration as measured by the Hosmer-Lemeshow test being secondary. Accuracy, sensitivity, and specificity are also reported as supplementary performance measures. Lastly, in line with the goals of the research question and given the inherent stochastic element in model performance, the significance of AUC differences between Cox regression and other models is assessed using the DeLong test (DeLong, DeLong, & Clarke-Pearson, 1988). Table 30 and Table 31 report the test data performance of the final model of each type outlined in Section 5.1 in predicting 30-day readmissions for GCUH and RH, ranked by each model's AUC.

Table 30. Final Model Performance for GCUH (RQ1)

Ranking	Method	AUC	HL (*p*-value)	Accuracy	Sensitivity	Specificity	DeLong Test *p*-value (2-tailed)
1	LR (AIC)	72.8678%	37.5774%	86.0766%	7.6923%	99.0590%	0.9130%
2	LR (BIC)	72.5665%	21.4265%	86.0909%	8.1951%	98.9924%	51.7111%
3	RSF	72.4634%	0.0083%	86.0980%	6.3851%	99.3005%	99.1111%
4	**Cox Regression**	**72.4610%**	**1.6340%**	**85.9551%**	**6.3851%**	**99.1340%**	**100.0000%**
5	CURE	72.3843%	0.0109%	85.7908%	0.0000%	100.0000%	71.2503%
6	NNET Survival	72.3826%	85.6071%	85.9909%	5.3293%	99.3505%	66.0509%
7	RIST	72.2922%	14.7152%	86.0123%	4.1730%	99.5670%	41.6894%
8	Time-Coded ANN	72.2330%	1.2063%	85.8980%	1.6591%	99.8501%	24.0096%
9	Cox NNET	72.1090%	0.0000%	85.7908%	0.0000%	100.0000%	4.7483%
10	BART	71.9202%	39.7077%	85.9266%	4.4746%	99.4171%	1.6606%
11	Survival Tree (Likelihood)	71.3292%	75.8304%	86.0051%	5.6310%	99.3172%	0.0039%
12	Survival Tree (Log Rank)	71.2283%	98.4777%	86.0337%	5.3796%	99.3921%	0.0006%
13	CURT V2	70.8949%	NA	85.8051%	0.8547%	99.8751%	0.0001%
14	CURT V1	50.0000%	NA	85.7908%	0.0000%	100.0000%	0.0000%

Shading represents the four worst models.

Table 31. Final Model Performance for RH (RQ1)

Ranking	Model	AUC	HL (*p*-value)	Accuracy	Sensitivity	Specificity	DeLong Test *p*-value (2-tailed)
1	CURE	71.4995%	0.0253%	83.5952%	0.4219%	100.0000%	3.6628%
2	RIST	71.4843%	9.2968%	83.8315%	5.7384%	99.2344%	2.3700%
3	LR (BIC)	71.3400%	1.8705%	83.9427%	10.6329%	98.4021%	10.1705%
4	LR (AIC)	71.3383%	0.4881%	84.0261%	10.7173%	98.4854%	4.2526%
5	RSF	71.2002%	4.0777%	83.8593%	8.1857%	98.7850%	23.5484%
6	NNET Survival	71.0517%	0.0001%	83.9010%	7.2574%	99.0180%	35.7394%
7	Time-Coded ANN	70.9827%	0.0000%	83.7898%	4.5570%	99.4174%	55.7764%
8	Cox NNET	70.9523%	0.0000%	83.5256%	0.0000%	100.0000%	67.0344%
9	BART	70.8627%	0.0016%	83.7481%	5.8228%	99.1178%	96.9522%
10	**Cox Regression**	**70.8534%**	**0.0065%**	**83.9149%**	**7.5105%**	**98.9847%**	**100.0000%**
11	Survival Tree (Likelihood)	70.0008%	0.7073%	83.8454%	7.6793%	98.8682%	3.8638%
12	CURT V2	69.8891%	2.8862%	83.8454%	14.2616%	97.5699%	2.2501%
13	Survival Tree (Log Rank)	69.6228%	0.0344%	83.7342%	9.8734%	98.3023%	0.2025%
14	CURT V1	68.8419%	NA	83.5256%	0.0000%	100.0000%	0.0008%

Shading represents the four worst models.

Prior to linking these results back to the research question several high-level observations can be made with respect to the discrimination and calibration results. First, the four worst models are shaded for both hospitals and are those based on individual trees, namely the survival tree and CURT models. These four models are notably worse than all other models for both datasets. They are thus afforded little discussion and not considered in the following observations. Secondly, all models achieved similar AUC performance for each hospital. The AUCs fell within ranges of 71.9202% to 72.8678% and 70.8534% to 71.4995% for GCUH and RH respectively. Despite this within-hospital similarity, there are clear between-hospital differences validating the decision to consider the two hospitals separately in this work. More specifically, GCUH appears to represent a less challenging modelling problem characterised by higher discrimination and more frequent calibration, with five of the 10 models (ignoring the shaded models) evaluated being well-calibrated at the 5% level of significance. In contrast, RH is characterised by lower discrimination and only the RIST model is well calibrated at the 5% level of significance. Lastly, the relative ranking of models is not consistent between the two hospitals, which is discussed further below with respect to its bearing on the research question. Having described these high-level features of the results, the focus now turns to the comparison of machine learning and statistical survival models which is the core of the research question.

Considering the GCUH results, the two variations of logistic regression offer the greatest discrimination, are both well calibrated, and the BIC variation has significantly improved discrimination compared to the Cox regression model. This feature of the results is not wholly unexpected, given that logistic regression models represent a classification approach which is also the basis for model assessment, advantaging these models relative to the survival models. Restricting attention to the survival models of primary interest for the research question, Cox regression exhibits superior discrimination to all machine learning models except for RSF model in terms of point estimates, but the difference is significant only for the Cox NNET and BART models. The Cox model is also not calibrated at the 5% level of significance. Overall, while the Cox model has good relative performance in terms of AUC point estimates, the NNET Survival (MTP) and RIST models are not significantly different with respect to AUC and are also calibrated at the 5% level of significance.

Turning attention to the RH results, the best two models in terms of discrimination are the CURE and RIST models. Both exhibit superior and significantly different discrimination compared to the Cox regression model. The RIST model is also the only model that is calibrated at the 5% level of significance. Despite representing classification approaches, the two logistic regression models perform worse in terms of discrimination than the CURE and RIST models and are not calibrated.

The differences in results for GCUH and RH should also be highlighted before further discussion. For GCUH, the statistical survival model was not calibrated at the 5% level of significance and did not have significantly different discrimination to two well-calibrated machine learning survival models. For RH, the statistical survival model was not calibrated and, excluding the shaded models, exhibited the lowest discrimination of the considered models. Two of the machine learning survival models had significantly different and better discrimination, one of which was also calibrated at the 5% level of significance. Where the statistical classification models were best on GCUH, they were outperformed in terms of discrimination and were not well calibrated on RH. In general, machine learning models performed better relative to both the survival and classification statistical models for RH but were not as compelling for GCUH.

Finally, brief comments can be made about the relative performance of categories of machine learning models. Excluding BART, the ensemble models of CURE, RIST, and RSF typically ranked higher in terms of AUC than the NNET Survival, Time-Coded ANN, and Cox NNET models. For both hospitals, these ensembles and ANNs both ranked more highly than the Bayesian ensemble model of BART, which in turn did better than the four models using individual survival trees.

7.1.1 Discussion of Findings for RQ1

The results presented above provide support for the assertion that machine learning survival models can improve upon statistical survival models when predicting 30-day hospital readmissions. This should be qualified, however, to reflect that the degree of improvement ought to be expected to vary between problems, with some settings being more tractable. In particular, the results provide evidence that the greatest potential improvement is in more complex settings such as RH. In this case, the complexity of the hospitals is based on the modelling results. In practice, domain experts are expected to be

best situated to gauge the difficulty of readmission prediction prior to the actual modelling exercise. Though the comparison of only two hospitals serves as a limited basis for this interpretation, it is strengthened by its link to the more general expectation that machine learning models offer the greatest potential improvement on complex problems. This interpretation of the results is based on the more complex problem (RH), characterised by poorer performance in general, also being characterised by much better machine learning model performance relative to statistical models. On this problem, the best performance for discrimination was achieved by machine learning models and the only calibrated model was RIST. This is made more compelling by favourable performance of machine learning models to the statistical classification models, which would be expected to be advantaged given the basis for performance assessment. On the less complex problem (GCUH), characterised by better performance in general, evidence in favour of machine learning survival models over statistical survival models is less pronounced. The statistical classification models are also clearly better in terms of discrimination and are well calibrated.

The findings make several contributions to the research area and have implications for practice.

First, it was demonstrated that machine learning survival models can improve upon previously used statistical survival models for 30-day unplanned readmission prediction. Previous research in this area has considered a limited range of survival techniques and has not provided thorough comparisons of their performance. Here, a range of machine learning survival techniques not previously employed in the identified research were considered, and models were assessed with comparison to the benchmark statistical survival model. This assessment reflected the most common application of readmission prediction models noted in the reviewed literature. The potential for improvement from the machine learning models should motivate their consideration in future studies and applications which opt for the use of survival models, rather than limiting consideration to Cox regression. While a common critique of the improvements of machine learning models is that they come at the cost of interpretability, the machine learning models may be appropriate if the goal is prediction rather than inference. Further, despite similar barriers to interpretation, there has been great recent interest in machine learning classification techniques for readmission prediction. The issue of interpretation is also of reduced relevance with the increased availability of methods such as LIME (Ribeiro et

al., 2016) or SHAP (Lundberg & Lee, 2017) for extracting rules and relationships from machine learning models.

Secondly, following on from the comparison to statistical survival techniques, this study also contributes the first extensive comparison of machine learning survival techniques in predicting hospital readmissions. Other than Cox regression, the only survival technique applied in the reviewed literature was the random survival forest. This study has included the random survival forest as well as eight distinct machine learning survival techniques and provided a comparison on two hospital populations. Though generalisability should be assessed in future research, clear underperformance from individual tree methods was identified as well as encouraging performance from RIST, CURE and RSF on both datasets. In particular, the final RIST model had competitive discrimination and was the only model well-calibrated on both datasets. While the performance of these models is encouraging, the range of comparisons makes reliably determining the reason for the improved performance difficult. Despite this, the results do support the assertion that these models should be subject to greater consideration in future readmission problems involving survival analysis techniques.

Thirdly, this study provides evidence for machine learning survival models being able to achieve performance comparable to the most common classification benchmark in readmission modelling (logistic regression) on the more complex dataset. The above contributions have been framed in terms of the potential value of machine learning techniques and their comparison to statistical techniques where survival models are the focus, despite being evaluated in terms of a binary outcome at a single time point. The set of models compared on the two hospitals included, however, two logistic regression models. Given that the basis of evaluation for all models reflected a classification approach, it would be expected that classification models are advantaged unless the readmission dynamics are complex enough to warrant more complex techniques. On the less complex dataset, both logistic regression models were ranked above all survival techniques. On the more complex dataset, the best two models were both machine learning models, namely CURE and RIST. Unlike the RIST model, both logistic regressions were not well calibrated in this case. The results provide evidence in favour of machine learning techniques over statistical techniques more generally for settings with complex readmission dynamics, as there appears to be benefit in the additional capacity for complexity. They also provide evidence for the potential of machine learning

survival models to achieve comparable performance to the benchmark classification model for classification-style model application. This has implications for contexts where classification performance equivalent to the benchmark is needed and the additional insights of a survival model are also desirable. In these circumstances, machine learning survival techniques should be considered in addition to established classification techniques.

The final contributions are more general in nature and relate to region-specific readmission modelling research and the empirical comparison of machine learning survival techniques. First, when considering the readmission literature, there has been limited research considering an Australian context, especially in comparison to the United States. The need for context-specific research was described in Section 2.1.2.4, and the results presented here represent an addition to the small subset of literature that is specific to Australia. Secondly, considering the literature on machine learning technique for survival data, it was noted in Section 2.2.5 that there have been few comparisons of categories of techniques. That is, while research extending neural networks to survival data typically compares proposed models to other neural network extensions, they are not typically compared to other machine learning techniques for survival analysis such as ensembles like BART or RIST. A similar statement can be made for the literature involving tree ensembles, which have not been frequently compared to extensions of neural networks to survival analysis. Given the lack of comprehensive and cross-category comparisons, the results produced contribute to the literature through provision of a limited (given the single time point evaluation) empirical comparison of a range of available machine learning techniques for survival analysis.

7.2 Results for RQ2

How well can various survival modelling techniques capture aspects of hospital readmission risk over time relevant to managerial decision-making?

Section 5.2.3 described several applications of survival models to support managerial decision-making, highlighted relevant aspects of model performance in these applications, and identified corresponding measures of performance. These measures are:

- Time-dependent concordance: this measure relates to the discriminative ability of a model's predicted risk over time.

- D-Calibration: this measure is a hypothesis test assessing the calibration of entire survival curves produced by a model.
- IBS: this is a commonly employed measure of survival model performance considering both discrimination and calibration.

In presenting the results, an argument could be made for ordering either by time-dependent concordance or IBS. It is analogous to common practice in readmission literature to prioritise AUC, and time-dependent concordance represents an extension of this approach to a survival context. It is also the relevant measure for applications such as dynamic risk ranking where calibration is less important. The IBS measure, on the other hand, was used to select the final models being presented and considers both calibration and discrimination, albeit in a distinct and fixed manner. Rather than select one, both bases for ranking are used to present the results in Table 32 and Table 33 for GCUH and RH, respectively. This enriches the results by facilitating easy comparisons between the rankings. The *p*-value results from the test of D-Calibration are considered only in terms of whether a model is calibrated, as the omnibus nature of the underlying χ^2 test makes ranking these values inappropriate.

Table 32. Final Model Performance for GCUH (RQ2)

Rank	Ordered by Concordance				Ordered by IBS			
	Method	**Time-Dependent Concordance**	**D-Calibration *p*-value (k=10)**	**IBS**	**Method**	**Time-Dependent Concordance**	**D-Calibration *p*-value (k=10)**	**IBS**
1	NNET Survival	72.1235%	3.7668%	0.1243	RSF	70.9374%	73.2305%	0.12050
2	Cox Regression	72.0204%	93.8873%	0.1218	RIST	71.3790%	99.6067%	0.12096
3	Cox NNET	71.6616%	5.3971%	0.1224	Cox Regression	72.0204%	93.8873%	0.12177
4	Time-Coded ANN	71.5640%	47.5766%	0.1233	CURE	71.2976%	66.4773%	0.12191
5	RIST	71.3790%	99.6067%	0.1210	Cox NNET	71.6616%	5.3971%	0.12242
6	CURE	71.2976%	66.4773%	0.1219	Survival Tree (Likelihood)	68.7104%	99.2736%	0.12326
7	BART	71.2200%	72.2484%	0.1239	Time-Coded ANN	71.5640%	47.5766%	0.12331
8	RSF	70.9374%	73.2305%	0.1205	BART	71.2200%	72.2484%	0.12389
9	Survival Tree (Likelihood)	68.7104%	99.2736%	0.1233	CURT V2	68.3680%	97.6678%	0.12392
10	Survival Tree (Log Rank)	68.4682%	77.6016%	0.1239	Survival Tree (Log Rank)	68.4682%	77.6016%	0.12394
11	CURT V2	68.3680%	97.6678%	0.1239	NNET Survival	72.1235%	3.7668%	0.12426
12	CURT V1	0.0000%	99.9787%	0.1491	CURT V1	0.0000%	99.9787%	0.14915

Table 33. Final Model Performance for RH (RQ2)

Rank	Ordered by Concordance				Ordered by IBS			
	Method	**Time-Dependent Concordance**	**D-Calibration *p*-value (k=10)**	**IBS**	**Method**	**Time-Dependent Concordance**	**D-Calibration *p*-value (k=10)**	**IBS**
1	CURE	70.0901%	51.4138%	0.1326	NNET Survival	69.3910%	26.9462%	0.1300
2	RIST	70.0790%	94.5840%	0.1311	RIST	70.0790%	94.5840%	0.1311
3	Cox NNET	69.9858%	12.3572%	0.1322	RSF	69.5290%	99.4790%	0.1312
4	Cox Regression	69.9082%	97.5226%	0.1328	Cox NNET	69.9858%	12.3572%	0.1322
5	Time-Coded ANN	69.8737%	83.3974%	0.1342	CURE	70.0901%	51.4138%	0.1326
6	BART	69.6933%	79.0386%	0.1337	Survival Tree (Likelihood)	65.9155%	87.1188%	0.1328
7	RSF	69.5290%	99.4790%	0.1312	Cox Regression	69.9082%	97.5226%	0.1328
8	NNET Survival	69.3910%	26.9462%	0.1300	CURT V2	66.9108%	99.4329%	0.1330
9	CURT V2	66.9108%	99.4329%	0.1330	BART	69.6933%	79.0386%	0.1337
10	Survival Tree (Likelihood)	65.9155%	87.1188%	0.1328	Time-Coded ANN	69.8737%	83.3974%	0.1342
11	Survival Tree (Log Rank)	65.0441%	71.1968%	0.1345	Survival Tree (Log Rank)	65.0441%	71.1968%	0.1345
12	CURT V1	55.8811%	79.9270%	0.1367	CURT V1	55.8811%	79.9270%	0.1367

These results are briefly described in terms of each of the three measures individually. A summary of the measure results is then provided before the discussion.

Concordance performance can be seen to vary within a tight band for each hospital. When excluding the worst four models for GCUH and RH, the range of concordance values were 1.186% and 0.699% respectively. This is in contrast to the much lower performance seen for the survival tree and CURT models, which exhibited concordance values at least 2.227% and 2.480% lower than all other models for GCUH and RH, respectively. Whereas the worst four models did not change between hospitals, there is slight variation in the best four models between the two problems. Excluding the poorly calibrated model (NNET Survival), the four models on GCUH with highest concordance were Cox regression, Cox NNET, time-coded ANN, and RIST. The four models on RH with highest concordance were CURE, RIST, Cox NNET, and Cox regression. RH also appears to be a more complex problem characterised by lower performance in general with respect to concordance (and IBS as will be mentioned below). As reported for the first research question, machine learning models demonstrated improved performance compared to the statistical model on the more complex problem. Overall, in addition to the difference in typical performance between hospitals, the results show that the worst models are clearly identifiable based on concordance while the best performers are less distinguished.

The calibration aspect of the results is more straightforward. All models were found to be D-calibrated at the 5% level of significance apart from the NNET survival model on GCUH. This is an encouraging finding as it indicates the identified survival modelling techniques can result in suitably calibrated models for the proposed applications.

The results are now considered with respect to IBS and how the relative ranking of models change compared to considering only concordance. When using concordance, it was highlighted that the worst four models were clearly the two survival trees and two CURT models on both GCUH and RH. This is no longer the case based on IBS, with the survival tree using a one-step likelihood splitting function being ranked 6^{th} for both hospitals and the modified CURT being 8^{th} on RH. The worst four performers for GCUH include the two CURT models, the survival tree using a log-rank splitting function, and the poorly calibrated NNET survival model. The worst four performers for RH are the BART, Time-Coded ANN, survival tree using a log-rank splitting function, and unmodified CURT model. While the individual tree models still make up most of the poor performers, this is less pronounced than when considering concordance alone. The best performers in

terms of IBS for GCUH are RSF, RIST, Cox regression, and CURE. The best performers for RH are NNET survival, RIST, RSF, and Cox NNET. As before, there is some between-hospital consistency for the top performers with RIST and RSF being common to both. Also consistent with consideration of concordance-ranked results, the Cox regression model exhibited lower relative performance on the more complex problem.

Finally, comments can be made regarding the performance of categories of machine learning models. The clearest stratification of categories of techniques comes from the generally poor performance of individual survival tree and CURT models, though this was less pronounced when considering IBS compared with concordance. The distinction between the ensemble and ANN models is less consistent. For GCUH, ANN models performed better than ensembles in terms of concordance, but the reverse is seen in terms of IBS. For RH, ensembles did slightly better in general than ANN models in terms of concordance, but performance was more mixed in terms of IBS. Finally, the Bayesian model performed relatively poorly in all instances.

To summarise, nearly all models were found to be D-calibrated on both hospitals, with only one exception. Comparing hospitals, it was noted that RH is more complex than GCUH and is also characterised by less competitive performance of the statistical survival model (Cox regression). In terms of the measure used for model ranking, some variation was observed in both the best and worst models when using concordance versus IBS. Most notable is the variation in the worst performing models, where concordance ranking found individual tree models to be substantially worse than all others on both hospitals but IBS ranking of these four models was less severe, with one being ranked 6^{th} on both hospitals. Finally, while differences related to hospital and measures considered manifested in notable differences in relative model performance, some models demonstrated strong performance in all instances, most notably the RIST model.

7.2.1 Discussion of Findings for RQ2

The above results provide an empirical demonstration of the ability of various survival modelling techniques to capture the aspects of model performance relevant for managerial decision-making. This section addresses the following questions:

- Which models are promising?

- Do machine learning survival techniques improve on statistical survival techniques?
- What are the implications for model consideration from comparing results across hospitals and across ranking measures?
- What are the implications for performance measurement from comparing results across hospitals and ranking measures?

When discussing which of the models considered are particularly promising, it should be noted that the practical differences linked to differences in the used performance measures are not immediately apparent. Even without direct quantification of real-world differences, however,

Considering the consistent aspects of model performance, several modelling techniques appear promising for inclusion in future research given their strong relative performance across both hospitals for all metrics. Most prominent of these is the RIST modelling technique, which resulted in models with good relative performance for both hospitals and both ranking metrics. Also notable are the Cox regression and the Cox NNET models, which were in the top four models in three of the four scenarios. While it is not expected that all relevant instances of future research consider the breadth of models applied here, these three modelling techniques should be prominent among those that are.

Warranting greater discussion, the variability in model rankings as a function of both hospital and basis for ranking have several implications. Focusing first on comparisons between the two hospitals, the Cox regression model had slightly worse relative performance on the more complex problem represented by RH for both ranking metrics. This is consistent with the results presented for the first research question and more general expectations regarding the situations where machine learning techniques are most promising. In line with the interpretation that machine learning techniques have the greatest potential to outperform statistical techniques on more complicated readmission problems, the relative ranking of the Cox NNET and Cox regression models is of interest. Cox NNET represents a machine learning (ANN) extension of the statistical Cox model. This machine learning extension ranked below the statistical model on the less complex problem (GCUH) for both ranking measures, but this was reversed for the more complex problem (RH).

Turning attention to variability in relative model rankings related to both hospitals and performance measures, two aspects are of interest. As mentioned above, the RIST, Cox NNET and Cox regression models performed well consistently. Less consistently, the best models also included CURE, RSF, and the time-coded ANN depending on the scenario. The first aspect of interest is in how the complexity of the problem and the basis for performance measurement affect the recommendation of modelling techniques. As the consistent performance of RIST, Cox NNET and Cox regression has been highlighted above and it is recommended all three be considered in general, the following recommendations relate to scenario-specific decisions. If discrimination is the most important measure, the time-coded ANN should be considered for less complex problems whereas CURE should be considered where there is greater complexity. Where a combination of discrimination and calibration is required, the recommended modelling technique is CURE for less complex problems and RSF otherwise. These recommendations provide an informed starting point in model consideration for both research and practice based on the results of this study. The second aspect of interest is not in the models specific to each scenario but instead in the between-scenario variability itself. While a few models performed consistently well, there was high variability in the best models across scenarios; it is unlikely that a single modelling technique will offer the best performance across settings (such as hospitals or patient populations) and across performance measures in general. For example, RIST models were most consistent in having good relative performance but were never best. The consequence of this is that model development, whether within research or practice, should aim to consider a variety of modelling techniques to better account for variation in which technique is best. This is particularly pertinent in the healthcare setting, where the magnitude of financial and patient welfare costs makes marginal improvement important.

The final discussion point relates to the use of IBS for model selection and model evaluation. It has previously been noted that the IBS equation can be formulated as a sum of a calibration and discrimination component, making it a useful measure given these are the aspects of model performance determined to be relevant for managerial decision-making. It does not, however, explicitly report the contribution of these components. This bears discussion given the changes in the worst-performing models when ranking is based on concordance versus IBS. When considering concordance, the two survival tree models and the two CURT models performed notably worse than all other models. When ranked

based on IBS, these models were no longer substantially poorer than other models. In particular, the survival tree using a splitting function based on a one-step likelihood was ranked sixth on both GCUH and RH. This is relevant and surprising because while almost all models exhibited acceptable D-calibration this model was characterised by notably lower discrimination. This may indicate that, depending on model applications considered, the IBS measure should be modified to adjust the relative balance between calibration and discrimination components. For example, the RH survival tree is ranked sixth in terms of IBS but only generates 20 unique survival curves which may be insufficient for certain applications such as dynamic risk ranking. The emphasis placed on calibration by the unadjusted IBS measure and its use in model selection may also have been a contributor to almost all models being D-calibrated.

7.2.2 Contributions from RQ2

Through this research question several contributions have been made to both practice and the literature. These contributions relate to identification of modelling techniques, relevant aspects of their performance, measures capturing these aspects, and the provision of empirical results.

First, to address this research question a range of machine learning survival modelling techniques were applied. These were identified through reviewing literature for major types of machine learning techniques and reducing the techniques found to those producing probabilistic outputs over time. Through this process, a contribution has been made through the identification of available, applicable, and previously unused modelling techniques in hospital readmission modelling. This contribution is made more notable by the increased consideration of machine learning techniques in the readmission literature under classification approaches. Machine learning techniques under survival approaches have been much rarer in previous research. The identified techniques provide alternatives to the most common survival modelling technique of Cox regression.

Secondly, for those studies taking survival approaches, only one instance was identified in which a survival model was applied in a manner distinct from that of classification models (Hao et al., 2015), only one instance where survival model assessment reflected the characteristics of survival data (Padhukasahasram et al., 2015), and no instances in which both model applications and assessment reflected were specific to survival data.

This study contributes through the proposal of several novel survival model applications to assist managerial decision-making, the determination of the relevant aspects of model performance for such applications, and the identification of appropriate performance measures. In doing so, this study aims to motivate the consideration of survival models outside of classification-centric applications and to provide guidance on the appropriate assessment of these models in future research and in institution-specific implementations.

Thirdly, building on the identification of applicable survival modelling techniques and appropriate performance measures, this study provided an empirical assessment of a wide range of survival modelling techniques on two hospitals. This empirical comparison demonstrated that appropriately calibrated survival curves can be produced in almost all instances; it highlighted the need for considering a variety of modelling techniques, and provided an initial indication as to the most promising techniques, namely RIST, Cox regression, and Cox NNET.

Several more general contributions are also made. These overlap with the contribution of the first research question but are made distinct using survival-specific performance measures. First, the results provided additional evidence indicating that the machine learning survival modelling techniques are most promising in more complex settings. The implications for both future research and practice lie in the need to consider problem complexity to inform the choice of techniques for modelling. Research should also aim to further quantify and establish the generalisability of this finding through comparisons of institutions and patient populations. Secondly, this study adds to the limited literature on readmission modelling in the Australian context. As mentioned previously, Australian readmissions have received relatively little research, with most research considering the USA. The need for region-specific investigations is driven by differences in population dynamics and healthcare systems in general, and specifically between Australia and the USA in this case. This impedes generalisation of results. Thirdly, this study provides an empirical comparison of machine learning survival modelling techniques using measures specific to survival analysis. Previous research into machine learning survival models has typically been characterised by intra-category comparisons, but there are few inter-category ones. For example, research proposing or investigating neural network extensions to survival analysis often involves comparisons to other neural network extensions, but not to ensemble or tree methods. This study provided an empirical comparison of a range of ensemble, individual tree, and neural network extensions to

survival analysis on two problems, adding to the literature with both intra- and inter-category comparisons.

8 Conclusion

This work has systematically reviewed the readmission modelling literature and identified key gaps. It has provided empirical assessment machine learning survival techniques that have not been considered previously in readmission modelling. It has also proposed novel applications specific to survival models and identified appropriate performance measures for such applications. Using the identified measures, an empirical assessment of a wide range of survival techniques on two hospitals was provided. These findings are expected to assist institutions in better managing readmissions through increased consideration of survival techniques and survival-specific applications. This work also provides guidance for healthcare institutions with respect to the available and promising techniques, potential applications, and appropriate performance assessment.

This chapter presents the conclusions from the systematic literature review and the overall conclusions of the resultant research questions. This is followed by suggestions for future research and acknowledgement of the relevant limitations in this work.

8.1 Conclusions of the Systematic Literature Review

The literature review of this work consists of a systematic review of hospital readmission research as well as a review of available statistical and machine learning survival analysis techniques. The primary conclusions of the readmission review are:

- Readmission modelling research has focused on classification models both in general and in increased consideration of machine learning techniques. This has been driven by how readmissions are measured under healthcare policy and by the wide range of applicable classification techniques. There is a lack of research investigating survival approaches in terms of available machine learning techniques, model applications, and model assessment. This is despite survival models providing more flexible outputs than classification models that may be more useful for individual institutions in managing readmissions.
- Since the recommendation for greater usage in the literature review of Artetxe et al. (2018), methods for addressing class imbalance have been more frequently employed and have been associated with improved classification performance. Research taking classification approaches should continue to employ these

methods as an avenue for improving model performance as well as adding to the evidence on their value.

- Comparability of studies in readmission modelling is made difficult by inherent sources of heterogeneity, and this has been exacerbated by the lack of a consistent categorisation of data features for reporting purposes. The research area would benefit from the adoption of a consistent categorisation, such as that employed by Kansagara et al. (2011).

All three conclusions contain recommendations for future research, and the first pertains to the gaps addressed in this research, namely the lack of consideration of machine learning survival techniques and survival model applications. The dominance of research under classification approaches matches the readmission definitions within healthcare policy, but machine learning and statistical survival models may be more useful for institutions aiming to effectively manage discharged patients in terms of interventions, follow-up care, and demand forecasting. Expanding on this, the key gaps related to the limited consideration of survival approaches were the following:

- While machine learning classification techniques have increasingly been considered as an avenue for improving on statistical techniques, this trend has largely ignored machine learning survival techniques. It is unclear whether machine learning survival techniques could improve on statistical survival techniques, in what scenarios, and how they compare with classification techniques.
- Little research has considered the use of survival models for readmission prediction, practical applications, and measures of performance distinct from those used under classification approaches. This is despite the additional information provided by survival models allowing for greater flexibility in practical applications.

8.2 Overall Conclusions

Motivated by the costs of readmissions, healthcare policy, and the gaps identified through the systematic literature review, this work has assessed machine learning and statistical survival models for readmission prediction. This was done when treating readmission as

a binary outcome, as in previous work, and when considering risk over time. At the highest level, the conclusions of this work are the following:

- Many machine learning survival techniques are applicable for readmission modelling but have not been considered in previous readmission research.
- Machine learning survival techniques can improve on statistical survival techniques and be competitive with statistical classification techniques when predicting 30-day unplanned readmissions.
- The relevant aspects of survival model performance for practical application in supporting managerial decision-making are the discrimination and calibration of time-varying risk predictions. Appropriate measures capturing these aspects are time-dependent concordance and D-calibration, neither of which have been applied in the reviewed literature.
- Survival techniques, both machine learning and statistical, can capture relevant aspects of readmission risk sufficiently for a variety of applications supporting managerial decision-making.

These conclusions together make the argument for greater consideration of machine learning survival techniques, practical applications of survival models, and associated performance measures. Classification models serve an important purpose, particularly for standardised readmission measurement, but are less useful for decision-making for individual institutions. In this work, survival techniques were found capable of capturing aspects of readmission risk relevant to a non-exhaustive range of novel applications involving dynamic risk stratification, identifying patient-specific risk periods, and readmission forecasting. The value proposition for survival models in readmission modelling is thus distinct and complementary to the value of classification models. Even in the case that a singular model is desired to avoid user-confusion or conflicting signals, survival models should be considered as well as classification models. Machine learning survival models can, in some scenarios, provide 30-day prediction performance competitive with classification models as well as predicting risk across time.

Conclusions are also made with respect to when machine learning techniques are most appropriate, which are and which are not promising, and the influence of different possible applications. These conclusions are drawn from the comparison of a wider range

of survival techniques than in any prior readmission modelling work using two hospitals with differences in common patient profiles.

- **Machine learning applicability** – Machine learning techniques have the greatest potential for improvement over statistical techniques on complex problems characterised by lower performance in general. For low complexity problems, more interpretable statistical survival techniques may be most appropriate.
- **Promising machine learning techniques** – RIST is the most promising survival technique for readmission modelling, regardless of fixed point or time varying risk prediction. For fixed point prediction, the CURE and RSF ensemble techniques are also promising, as well as Cox NNET for time varying risk prediction.
- **Unpromising machine learning techniques** – Individual survival tree, CURT, and BART techniques are least promising in this context, despite the survival tree and CURT techniques being the most interpretable.
- **Varied Applications and Techniques** – A wide range of techniques should be considered when selecting a model for practical use. The best technique is dependent on setting (e.g., hospital or population of interest) and the relative value of discrimination and calibration determined by intended application.

While the high-level conclusions have implications for the use of survival approaches in readmission modelling, these more granular conclusions have direct implications for the techniques considered under this approach in both practice and research.

Having detailed the conclusions of this work, its secondary contributions are also noted. These relate to readmission research in Australia and the empirical comparison of machine learning survival techniques. Most readmission research has taken place in the US, with other regions having a smaller body of relevant readmission literature. Differences in healthcare systems and populations, and in other regional factors, inhibit generalisability and make region-specific research valuable. This is particularly pertinent for Australia where healthcare policy targeting readmissions is a very recent development. This work adds to the readmission literature relevant to Australia with respect to emergency department presentations and a wide range of modelling techniques.

The contribution of this work in the provision of an empirical comparison of machine learning survival techniques stems from the limited inter-category comparisons in the

related literature. The literature review for machine learning survival techniques noted that newly proposed techniques of a given category (e.g., ANNs or ensembles) were typically compared to other techniques of the same category. For example, in the proposal of the NNET Survival technique by Gensheimer and Narasimhan (2019), the points of comparison were two other ANN survival techniques and Cox regression. Similarly, the RIST technique proposed by R. Zhu and Kosorok (2012) was compared with two other ensembles and Cox regression. This work contributes to machine learning survival literature through the provision of an extensive intra- and inter-category empirical comparison of individual tree, ensemble, and ANN extensions to survival data. For fixed point prediction, this comparison favoured ensemble techniques. For time-varying risk prediction, both ensemble and ANN techniques performed well without any clear distinction between categories. In both research questions, the techniques based on individual trees rather than ensembles performed poorly. Further, the BART ensemble technique also performed poorly, outperforming only the individual tree techniques.

8.3 Future Research

Several areas for future research are now suggested. These relate to research establishing the generalisability of the presented findings, extending comparisons to include machine learning classification techniques, and the further consideration of practical applications specific to survival models.

8.3.1 Patient, Region, and Data Generalisability

The results presented in this work were generated through consideration of patients admitted to emergency departments for two hospitals on the Gold Coast of Australia. Models were trained using administrative data available at discharge. Future research should thus establish the generalisability of these findings with alternative cohort definitions, other regions, and expanded data sources. These are each briefly expanded upon here.

Patient Cohort – This work's findings relate to patients admitted to the Emergency Department, which represents a relatively broad cohort definition. Policy and research have generally opted to define patient cohorts in terms of demographic features (for instance, paediatric patients defined in terms of age) or diagnosis-related group (DRG) (such as heart failure patients). This leads to smaller but more homogeneous cohorts

which may be of particular interest and on which predictive models may offer improved performance. Future research should thus assess whether the encouraging survival model performance extends to other patient cohorts where suitable survival models could be applied to assist decision-makers. Reflecting the emphasis of survival model applications on meeting institution-specific needs, these cohorts should be those for which better modelling of readmission dynamics is most important for institutions. Additionally, this project provided evidence for machine learning techniques having the greatest potential for improvement over statistical techniques for more complex problems. While this comparison was between hospitals, future research considering alternative patient cohorts should extend this comparison to establish whether machine learning techniques are more promising for certain conditions or demographic groups.

Region – These findings relate to two hospitals in a single Australian city. As mentioned previously, differences in readmission dynamics between regions may manifest in differences in model performance. This work has added to the Australian readmission literature, which has received relatively little research. Future research should further extend this work to assess the consistency of positive survival model findings across a broader range of Australian hospitals, where regional differences should be small. This should then be extended to other regions, such as the US, where regional differences are larger.

Data Sources – the focus for this research were the techniques employed and associated performance measurement in readmission prediction, specifically in relation to survival approaches. Relative to these, the data used to train the models were de-emphasised and only administrative costing data available at discharge was included. This strengthens the contributions made in the project by making findings more broadly applicable, but future research should extend this work by replicating it with more comprehensive data sources for two reasons. First, the inclusion of non-standard data has been a recent trend in readmission research, whether by better extracting data from fields in traditional sources or by leveraging novel data sources. There is thus a need to demonstrate these findings are consistent with modern readmission prediction methods. It is hypothesised that the additional information, especially in non-standard formats, would further advantage the machine learning techniques identified. Secondly, several fields not recorded by the time of discharge in the data source considered in this project may have predictive value. Future research should compare model performance under different data scenarios; these

should include performance when using all data and when restricted to data currently available at discharge.

This research would identify potential changes in hospital processes which would better support the use of predictive support tools, either through prompter data recording or integration of multiple data sources. As a notable example, DRG codes were found to have high variable importance in the IHPA's newly proposed pricing model (Independent Hospital Pricing Authority, 2021b). This field is not available, however, until six to eight weeks after discharge in the costing data used in this work.

8.3.2 Extending Comparisons to Machine Learning Classifiers

Logistic regression is the most employed technique in readmission prediction in general and is also the most common point of comparison in readmission research considering machine learning techniques. This made it an appropriate benchmark in this work to reflect the current classification standard of prediction. This work concluded that, for 30-day unplanned readmission prediction, machine learning survival models can achieve performance better than statistical survival models and competitive with logistic regression on more complex problems. This conclusion provides support for the practical use of machine learning survival models for readmission prediction, given they can provide risk over time without compromising fixed time-point performance compared with the benchmark classifier. This is of relevance where there is a preference for a single model predicting both fixed time and time-varying risk, rather than two tools which may produce conflicting messages and cause user-confusion. If, however, the competitive performance of machine learning survival models is achieved in the same scenarios in which machine learning classifiers outperform the benchmark (hypothesised to be more complex problems), then these machine learning classifiers should be the point of comparison. Future research should thus compare the techniques found to be promising, such as RIST and CURE, with machine learning classifiers for fixed time prediction. This would also further add to the evidence regarding the scenarios where machine learning models have the greatest potential to improve on statistical models.

8.3.3 Applications of Survival Models

This work proposed several novel survival model applications. These were then discussed in the general sense to determine the relevant aspects of survival model performance for practical use. Future research should expand on these proposed applications through collaboration with relevant stakeholders in several respects. First, the applications

proposed in this work are not necessarily exhaustive with respect to managerial decision-making, and no consideration was given to the needs of other areas of hospital operations, such as the needs of clinicians. The potential for other applications to meet stakeholders' needs should thus be a focus for future research. Secondly, future research should aim to provide guidance regarding the most effective implementation of those applications proposed in this work which allow for variation. For example, DRR could be based on daily risk of readmission or risk of readmission in some period. Similarly, various criteria could be used to determine ERP, as illustrated in Figure 2 in Section 5.2.3. Determining the details of such applications in practice should be done through collaborative research including hospital stakeholders to ensure the applications reflect their needs. Thirdly, the value of these applications when supporting decision-making from patient outcome and cost perspectives should be assessed.

8.3.4 Customisable Model Selection with IBS

In this work, measures were employed for appropriately capturing discrimination and calibration for the proposed applications of survival models, namely time-dependent concordance and D-calibration. Neither was used for model selection as no appropriate method for combining the two measures was immediately evident, and using either individually would ignore the need for the other. Instead, IBS was used given that it can be decomposed into a linear sum of discrimination and calibration components, albeit with a fixed balance and measured in a distinct manner to the aspect-specific measures. The fixed balance between the components in the IBS measure is relevant when considering that different balances of discrimination and calibration are desirable for different applications and scenarios. As a simple example, the DRR application proposed requires little calibration but high discrimination. The fixed balance of IBS and the need for a customisable balance in both model selection and final model evaluation motivates the need for future research to explore modification of the IBS measure. Specifically, future research should aim to customise the balance between discrimination and calibration components of the IBS and explore the way performance of the final models selected under different balances varies. This research would better enable customisation of models to specific applications.

8.4 Limitations

Finally, limitations of this research are outlined.

- **Generalisability** – Establishing the generalisability of this study's findings will require future research, as the two hospitals considered may not be representative of Australia as a whole. Similarly, the findings are based on patients with an initial presentation to the emergency department and may differ for other populations (for instance, a population with an alternative initial presentation or defined in terms of diagnosis group). Finally, the results may vary between regions (such as the US) as a result of differences in healthcare systems and populations, and other regional factors.
- **Uncaptured readmissions** – Some patients may have had their index admission at GCUH or RH but later been readmitted elsewhere. It is not, however, feasible to estimate the frequency of this occurrence in the absence of data from other institutions without making material assumptions. This limitation is minor as the two public hospitals are the two largest hospitals in the Gold Coast region.
- **Reason for Index Admission** – The reason for admission is often included in readmission prediction research to define patient cohorts and would be expected to improve model performance in this setting. It was not included, however, because it is not recorded in the relevant systems within the two hospitals at the time of discharge. As this work aimed to support decision-making based on information at the time of discharge, the reason for admission was not considered, though future research assessing the predictive value of this field may motivate its being recorded more promptly.
- **Improve Metrics Determining Model Appropriateness** – Other aspects of model predictions not explicitly captured by the identified performance measures may influence their appropriateness for practical applications. For example, some of the tree models were noted as producing a very limited number of unique predictions, which is undesirable for risk stratification applications. As another example, some models produced hazard functions characterised by periodic jumps (such as survival trees), which may pose difficulties when attempting to use these functions to identify when readmission risk stabilises and may necessitate interpolation or smoothing.

- **Practical Implementation of DRR and ERP Applications** – These two applications were discussed in the general sense, but the details of their implementation were not described. DRR could be based on risk of readmission in the next day, week, or some other period. Other summaries of readmission risk may also be relevant. Similarly, a variety of options is available for determining ERP, with three example options illustrated in Figure 2 in Section 5.2.3. To avoid limiting consideration to specific implementations, relevant aspects of model performance were determined based on what aspects are desirable for any implementation. Future research should substantiate these applications with recommendations for how they would be implemented in practice to maximise value to institutions, as suggested in Section 8.3.3.
- **Granularity of Model Selection** – Model selection in this research was more thorough than any prior readmission work reviewed. It may, however, have been less thorough than it would be in practice. This is because of the range of techniques considered, the computational requirement of BART, and the algorithm used for Cox-NNET. With respect to the range of techniques employed, a less extensive range would have allowed for model selection to be more thorough with respect to hyperparameters. The range of techniques considered was a strength of this work, however, and it is argued that improvement from considering more hyperparameter values would be minor given the level of granularity used in the search grids and the allowance for initial results to inform ranges considered. With respect to the BART technique, this limitation is more relevant because training time and memory requirements made consideration of hyperparameters difficult, though previous research has demonstrated that varying the default hyperparameters is of limited value. Finally, the implementation of the Cox-NNET technique used in this work was associated with long training times and memory requirements, as documented in previous work (Gensheimer & Narasimhan, 2019). A more efficient implementation has since been proposed (D. Wang, Jing, He, & Garmire, 2021); this would have reduced training times and memory requirements and in turn allowed for consideration of additional or more detailed hyperparameters. Future research should use this newer implementation for generating Cox-NNET models.
- **COVID** – Finally, the data used in this project related to a pre-COVID environment which may differ to some extent from the current and future

environment in terms of population dynamics and health processes. Australia was comparatively effective in managing the pandemic, however, which may reduce the magnitude of COVID's long term influence.

9 References

Aalen, O. (1978). Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics, 6*(4), 701-726. Retrieved from http://www.jstor.org/stable/2958850

Aalen, O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine, 8*(8), 907-925. doi:doi:10.1002/sim.4780080803

Aalen, O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine, 12*(17), 1569-1588. doi:doi:10.1002/sim.4780121705

Alajmani, S., & Elazhary, H. (2019). Hospital readmission prediction using machine learning techniques: A comparative study. *International Journal of Advanced Computer Science and Applications, 10*(4), 212-220. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065823405&partnerID=40&md5=2c0e54e05837861895eb5bafda57e819

Almardini, M., & Raś, Z. W. (2017) A supervised model for predicting the risk of mortality and hospital readmissions for newly admitted patients. In*: Vol. 10352 LNAI. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 29-36).

Almussallam, B., Joyce, M., Marcello, P. W., Roberts, P. L., Francone, T. D., Read, T. E., . . . Ricciardi, R. (2016). What factors predict hospital readmission after colorectal surgery? *American Surgeon, 82*(5), 433-438. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020168595&partnerID=40&md5=40057b95669732cd7dcfc5c71f080cfb

Anderson, G. F., & Steinberg, E. P. (1985). Predicting hospital readmissions in the Medicare population. *Inquiry, 22*(3), 251-258. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0022116686&partnerID=40&md5=0d816d7e89faf6298321fad8f1dadbcd

Andres, A., Montano-Loza, A., Greiner, R., Uhlich, M., Jin, P., Hoehn, B., . . . Kneteman, N. M. (2018). A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PLoS One, 13*(3), e0193523. doi:10.1371/journal.pone.0193523

Antolini, L., Boracchi, P., & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Stat Med, 24*(24), 3927-3944. doi:10.1002/sim.2427

Ardura-Garcia, C., Stolbrink, M., Zaidi, S., Cooper, P. J., & Blakey, J. D. (2018). Predictors of repeated acute hospital attendance for asthma in children: A systematic review and meta-analysis. *Pediatric Pulmonology, 53*(9), 1179-1192. doi:10.1002/ppul.24068

Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer Methods and Programs in Biomedicine, 164*, 49-64. doi:10.1016/j.cmpb.2018.06.006

Atkins, E. R., Geelhoed, E. A., Knuiman, M., & Briffa, T. G. (2014). One third of hospital costs for atherothrombotic disease are attributable to readmissions: a linked data analysis. *BMC Health Services Research, 14*(1), 338. doi:10.1186/1472-6963-14-338

Australian Commission on Safety and Quality in Healthcare. (2019). Avoidable Hospital Readmissions. Retrieved from https://www.safetyandquality.gov.au/our-work/indicators/avoidable-hospital-readmissions/

Bae, S., Dey, A. K., & Low, C. A. (2016). *Using passively collected sedentary behavior to predict hospital readmission.* Paper presented at the UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing.

Baechle, C., Agarwal, A., Behara, R., Zhu, X. Q., & Ieee. (2017). Latent Topic Ensemble Learning for Hospital Readmission Cost Reduction. In *2017 International Joint Conference on Neural Networks* (pp. 4594-4601). New York: IEEE.

Baesens, B., Gestel, T. V., Stepanova, M., Poel, D. V. d., & Vanthienen, J. (2005). Neural Network Survival Analysis for Personal Loan Data. *The Journal of the Operational Research Society, 56*(9), 1089-1098. Retrieved from http://www.jstor.org/stable/4102202

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association, 101*(473), 119-137. doi:10.1198/016214505000000628

Bair, E., & Tibshirani, R. (2004). Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLOS Biology, 2*(4), e108. doi:10.1371/journal.pbio.0020108

Barnes, E. L., Kochar, B., Long, M. D., Kappelman, M. D., Martin, C. F., Korzenik, J. R., & Crockett, S. D. (2017). Modifiable Risk Factors for Hospital Readmission among Patients with Inflammatory Bowel Disease in a Nationwide Database. *Inflammatory Bowel Diseases, 23*(6), 875-881. doi:10.1097/MIB.0000000000001121

Benuzillo, J., Caine, W., Evans, R. S., Roberts, C., Lappe, D., & Doty, J. (2018). Predicting readmission risk shortly after admission for CABG surgery. *Journal of Cardiac Surgery, 33*(4), 163-170. doi:10.1111/jocs.13565

Bernabeu-Mora, R., García-Guillamón, G., Valera-Novella, E., Giménez-Giménez, L. M., Escolar-Reina, P., & Medina-Mirapeix, F. (2017). Frailty is a predictive factor of readmission within 90 days of hospitalization for acute exacerbations of chronic obstructive pulmonary disease: A longitudinal study. *Therapeutic Advances in Respiratory Disease, 11*(10), 383-392. doi:10.1177/1753465817726314

Biganzoli, E., Boracchi, P., Ambrogi, F., & Marubini, E. (2006). Artificial neural network for the joint modelling of discrete cause-specific hazards. *Artificial Intelligence in Medicine, 37*(2), 119-130. doi:https://doi.org/10.1016/j.artmed.2006.01.004

Biganzoli, E., Boracchi, P., Coradini, D., Grazia Daidone, M., & Marubini, E. (2003). Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors. *Annals of Oncology, 14*(10), 1484-1493. doi:10.1093/annonc/mdg422

Biganzoli, E., Boracchi, P., Mariani, L., & Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine, 17*(10), 1169-1186. doi:10.1002/(SICI)1097-0258(19980530)17:10<1169::AID-SIM796>3.0.CO2-D

Biganzoli, E., Boracchi, P., & Marubini, E. (2002). A general framework for neural network models on censored survival data. *Neural Networks, 15*(2), 209-218. doi:https://doi.org/10.1016/S0893-6080(01)00131-9

Billings, J., Dixon, J., Mijanovich, T., & Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ (Clinical research ed.), 333*(7563), 327-327. doi:10.1136/bmj.38870.657917.AE

Binder, H., & Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics, 9*, 14-14. doi:10.1186/1471-2105-9-14

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., & Do, K. A. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics, 27*(3), 359-367. doi:10.1093/bioinformatics/btq660

Bondell, H. D., & Reich, B. J. (2008). Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics, 64*(1), 115-123. doi:doi:10.1111/j.1541-0420.2007.00843.x

Boracchi, P., Biganzoli, E., & Marubini, E. (2001). Modelling cause-specific hazards with radial basis function artificial neural networks: application to 2233 breast cancer patients. *Statistics in Medicine, 20*(24), 3677-3694. doi:10.1002/sim.1112

Borer, S. M., Kokkirala, A., O'Sullivan, D. M., & Silverman, D. I. (2011). Systolic Strain Abnormalities to Predict Hospital Readmission in Patients With Heart Failure and Normal Ejection Fraction. *Cardiology Research, 2*(6), 274-281. doi:10.4021/cr104w

Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*: Wadsworth International Group.

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics, 30*(1), 89-99. doi:10.2307/2529620

Brown, S. F., Branford, A. J., & Moran, W. (1997). On the use of artificial neural networks for the analysis of survival data. *IEEE transactions on neural networks, 8*(5), 1071-1077. doi:10.1109/72.623209

Buckley, J., & James, I. (1979). Linear Regression with Censored Data. *Biometrika, 66*(3), 429-436. doi:10.2307/2335161

Centers for Medicare and Medicaid Services. (2020). Hospital Readmissions Reduction Program (HRRP). Retrieved from https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program

Chandra, A., Rahman, P. A., Sneve, A., McCoy, R. G., Thorsteinsdottir, B., Chaudhry, R., . . . Takahashi, P. Y. (2019). Risk of 30-Day Hospital Readmission Among Patients Discharged to Skilled Nursing Facilities: Development and Validation of a Risk-Prediction Model. *Journal of the American Medical Directors Association, 20*(4), 444-450.e442. doi:10.1016/j.jamda.2019.01.137

Chaturvedi, N., de Menezes, R. X., & Goeman, J. J. (2014). Fused lasso algorithm for Cox′ proportional hazards and binomial logit models with application to copy number profiles. *Biometrical Journal, 56*(3), 477-492. doi:10.1002/bimj.201200241

Cheng, B. T., & Silverberg, J. I. (2019). Predictors of hospital readmission in US children and adults with atopic dermatitis. *Annals of Allergy, Asthma and Immunology, 123*(1), 64-69.e62. doi:10.1016/j.anai.2019.04.016

Cheung, B. L. P., & Dahl, D. (2018). *Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks.* Paper presented at the 2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018.

Chi, C.-L., Street, W. N., & Wolberg, W. H. (2007). Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets. *AMIA Annual Symposium Proceedings, 2007*, 130-134. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813661/

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology, 14*(4), e1006076. doi:10.1371/journal.pcbi.1006076

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: BAYESIAN ADDITIVE REGRESSION TREES. *The Annals of Applied Statistics, 4*(1), 266-298. doi:10.2307/27801587

Chollet, F. (2015). Keras. Retrieved from https://keras.io

Cholleti, S., Post, A., Gao, J., Lin, X., Bornstein, W., Cantrell, D., & Saltz, J. (2012). Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2012*, 103-111. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540449/pdf/amia_2012_symp_0103.pdf

Chopra, C., Sinha, S., Jaroli, S., Shukla, A., & Maheshwari, S. (2017). *Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients.* Paper presented at the ACM International Conference Proceeding Series.

Colosimo, E., Ferreira, F. v., Oliveira, M., & Sousa, C. (2002). Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. *Journal of Statistical Computation and Simulation, 72*(4), 299-308. doi:10.1080/00949650212847

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. doi:10.1007/BF00994018

Cotter, P. E., Bhalla, V. K., Wallis, S. J., & Biram, R. W. S. (2012). Predicting readmissions: poor performance of the LACE index in an older UK population. *Age and Ageing, 41*(6), 784-789. doi:10.1093/ageing/afs073

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187-220. Retrieved from http://www.jstor.org/stable/2985181

Cui, Y., Metge, C., Ye, X., Moffatt, M., Oppenheimer, L., & Forget, E. L. (2015). Development and validation of a predictive model for all-cause hospital readmissions in Winnipeg, Canada. *Journal of Health Services Research & Policy, 20*(2), 83-91. doi:10.1177/1355819614565498

Cunha Ferré, M. F., Gallo Acosta, C. M., Dawidowski, A. R., Senillosa, M. B., Scozzafava, S. M., & Saimovici, J. M. (2019). 72-hour hospital readmission of older people after hospital discharge with home care services. *Home Health Care Services Quarterly, 38*(3), 153-161. doi:10.1080/01621424.2019.1616024

Cutler, S. J., & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *Journal of Chronic Diseases, 8*(6), 699-712. doi:https://doi.org/10.1016/0021-9681(58)90126-7

da Silva, K. R., Albertini, C. M. M., Crevelari, E. S., de Carvalho, E. I. J., Fiorelli, A. I., Filho, M. M., & Costa, R. (2016). Complications after surgical procedures in patients with cardiac implantable electronic devices: Results of a prospective registry. *Arquivos Brasileiros de Cardiologia, 107*(3), 245-256. doi:10.5935/abc.20160129

De Laurentiis, M., De Placido, S., Bianco, A. R., Clark, G. M., & Ravdin, P. M. (1999). A Prognostic Model That Makes Quantitative Estimates of Probability of Relapse for Breast Cancer Patients. *Clinical Cancer Research, 5*(12), 4133. Retrieved from http://clincancerres.aacrjournals.org/content/5/12/4133.abstract

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics, 44*(3), 837-845. doi:10.2307/2531595

Deschepper, M., Eeckloo, K., Vogelaers, D., & Waegeman, W. (2019). A hospital wide predictive model for unplanned readmission using hierarchical ICD data. *Computer Methods and Programs in Biomedicine, 173*, 177-183. doi:10.1016/j.cmpb.2019.02.007

Deschodt, M., Flamaing, J., Wellens, N., Boonen, S., Moons, P., & Milisen, K. (2012). Comparison of Three Screening Tools to Predict Hospital Readmission and Mortality in Older Adults. *Journal of the American Geriatrics Society, 60*, S198-S198. Retrieved from <Go to ISI>://WOS:000302464800575

Donzé, J., Aujesky, D., Williams, D., & Schnipper, J. L. (2013). Potentially Avoidable 30-Day Hospital Readmissions in Medical Patients: Derivation and Validation of a Prediction Model. *JAMA Internal Medicine, 173*(8), 632-638. doi:10.1001/jamainternmed.2013.3023

dos Santos, M. R., Sayegh, A. L. C., Groehs, R. V. R., Fonseca, G., Trombetta, I. C., Barretto, A. C. P., . . . Alves, M. J. N. N. (2015). Testosterone deficiency increases hospital readmission and mortality rates in male patients with heart failure. *Arquivos Brasileiros de Cardiologia, 105*(3), 256-264. doi:10.5935/abc.20150078

Du, P., Ma, S., & Liang, H. (2010). PENALIZED VARIABLE SELECTION PROCEDURE FOR COX MODELS WITH SEMIPARAMETRIC RELATIVE RISK. *The Annals of Statistics, 38*(4), 2092-2117. Retrieved from http://www.jstor.org/stable/20744484

Duggal, R., Shukla, S., Chandra, S., Shukla, B., & Khatri, S. K. (2016a). Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India. *International Journal of Diabetes in Developing Countries, 36*(4), 469-476. doi:10.1007/s13410-016-0495-4

Duggal, R., Shukla, S., Chandra, S., Shukla, B., & Khatri, S. K. (2016b). Predictive risk modelling for early hospital readmission of patients with diabetes in India. *International Journal of Diabetes in Developing Countries, 36*(4), 519-528. doi:10.1007/s13410-016-0511-8

Duncan, I., & Huynh, N. (2018). A Predictive Model for Readmissions Among Medicare Patients in a California Hospital. *Population Health Management, 21*(4), 317-322. doi:10.1089/pop.2017.0150

Dupuis-Lozeron, E., Soccal, P. M., Janssens, J. P., Similowski, T., & Adler, D. (2018). Severe dyspnea is an independent predictor of readmission or death in COPD patients surviving acute hypercapnic respiratory failure in the ICU. *Frontiers in Medicine, 5*(MAY). doi:10.3389/fmed.2018.00163

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association, 72*(359), 557-565. doi:10.2307/2286217

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407-499. doi:10.1214/009053604000000067

Eleuteri, A., Tagliaferri, R., Milano, L., De Placido, S., & De Laurentiis, M. (2003). A novel neural network-based survival analysis model. *Neural Networks, 16*(5), 855-864. doi:https://doi.org/10.1016/S0893-6080(03)00098-4

Eleuteri, A., & Taktak, A. F. G. (2012). *Support vector machines for survival regression*. Paper presented at the 8th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. https://link.springer.com/chapter/10.1007%2F978-3-642-35686-5_15

Evers, L., & Messow, C.-M. (2008). Sparse kernel methods for high-dimensional survival data. *Bioinformatics, 24*, 1632-1638. doi:10.1093/bioinformatics/btn253

Fan, J. (1997). Comments on «Wavelets in statistics: A review» by A. Antoniadis. *Journal of the Italian Statistical Society, 6*(2), 131. doi:10.1007/bf03178906

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association, 96*(456), 1348-1360. Retrieved from http://www.jstor.org/stable/3085904

Fan, J., & Li, R. (2002). Variable Selection for Cox's proportional Hazards Model and Frailty Model. *Ann. Statist., 30*(1), 74-99. doi:10.1214/aos/1015362185

Fan, J., Su, X.-G., Levine, R. A., Nunn, M. E., & LeBlanc, M. (2006). Trees for Correlated Survival Data by Goodness of Split, With Applications to Tooth Prognosis. *Journal of the American Statistical Association, 101*(475), 959-967. doi:10.1198/016214506000000438

Faraggi, D., & Simon, R. (1995). A neural network model for survival data. *Statistics in Medicine, 14*(1), 73-82. doi:10.1002/sim.4780140108

Faraggi, D., & Simon, R. (1998). Bayesian Variable Selection Method for Censored Survival Data. *Biometrics, 54*(4), 1475-1485. doi:10.2307/2533672

Faraggi, D., Simon, R., Yaskil, E., & Kramar, A. (1997). Bayesian Neural Network Models for Censored Data. *Biometrical Journal, 39*(5), 519-532. doi:10.1002/bimj.4710390502

Farrell, S., Mitnitski, A., Rockwood, K., & Rutenberg, A. (2020). Generating synthetic aging trajectories with a weighted network model using cross-sectional data. *Scientific Reports, 10*(1), 19833. doi:10.1038/s41598-020-76827-3

Fathi, R., Bacchetti, P., Haan, M. N., Houston, T. K., Patel, K., & Ritchie, C. S. (2017). Life-Space Assessment Predicts Hospital Readmission in Home-Limited Adults. *Journal of the American Geriatrics Society, 65*(5), 1004-1011. doi:10.1111/jgs.14739

Fischer, C., Lingsma, H. F., Marang-van de Mheen, P. J., Kringos, D. S., Klazinga, N. S., & Steyerberg, E. W. (2014). Is the readmission rate a valid quality indicator? A review of the evidence. *PLoS One, 9*(11), e112282. doi:10.1371/journal.pone.0112282

Franco, L., Jerez, J., & Alba, E. (2005). *Artificial neural networks and prognosis in medicine. Survival analysis in breast cancer patients*.

Friebel, R., Hauck, K., Aylin, P., & Steventon, A. (2018). National trends in emergency readmission rates: a longitudinal analysis of administrative data for England between 2006 and 2016. *BMJ Open, 8*(3), e020325. doi:10.1136/bmjopen-2017-020325

Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics, 56*, 229-238. doi:https://doi.org/10.1016/j.jbi.2015.05.016

Gandy, A., & Jensen, U. (2005). On Goodness-of-Fit Tests for Aalen's Additive Risk Model. *Scandinavian Journal of Statistics, 32*(3), 425-445. doi:doi:10.1111/j.1467-9469.2005.00457.x

Garg, S. K., Sarvepalli, S., Goyal, H., Kandlakunta, H., & Sanaka, M. (2018). Nationwide Trends in Hospital Readmissions, Predictors, and Healthcare Utilization in Patients With Pancreatic Cancer. *Pancreas, 47*(10), 1387-1387. Retrieved from <Go to ISI>://WOS:000449304600099

Gensheimer, M. F., & Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ, 7*, e6257-e6257. doi:10.7717/peerj.6257

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom J, 48*(6), 1029-1040. doi:10.1002/bimj.200610301

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3-42. doi:10.1007/s10994-006-6226-1

Goldberg, Y., & Kosorok, M. (2017). Support vector regression for right censored data. *Electron. J. Stat., 11*(1), 532-569. doi:10.1214/17-EJS1231

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine, 18*(17-18), 2529-2545. doi:https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

Graña, M., Lopez-Guede, J. M., Irazusta, J., Labayen, I., & Besga, A. (2019). Modelling hospital readmissions under frailty conditions for healthy aging. *Expert Systems*. doi:10.1111/exsy.12437

Grzyb, M., Zhang, A., Good, C., Khalil, K., Guo, B., Tian, L., . . . Gu, Q. (2017). *Multi-task cox proportional hazard model for predicting risk of unplanned hospital readmission.* Paper presented at the 2017 Systems and Information Engineering Design Symposium, SIEDS 2017.

Gui, J., & Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics, 21*(13), 3001-3008. doi:10.1093/bioinformatics/bti422

Haider, H., Hoehn, B., Davis, S., & Greiner, R. (2020). Effective Ways to Build and Evaluate Individual Survival Distributions. *Journal of Machine Learning Research, 21*, 1-63.

Hammoudeh, A., Al-Naymat, G., Ghannam, I., & Obied, N. (2018). *Predicting hospital readmission among diabetics using deep learning.* Paper presented at the Procedia Computer Science.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36. doi:10.1148/radiology.143.1.7063747

Hao, S., Wang, Y., Jin, B., Shin, A. Y., Zhu, C., Huang, M., . . . Ling, X. B. (2015). Development, Validation and Deployment of a Real Time 30 Day Hospital Readmission Risk Assessment Tool in the Maine Healthcare Information Exchange. *PLoS One, 10*(10), e0140271-e0140271. doi:10.1371/journal.pone.0140271

Harrell, F. E., Jr., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med, 15*(4), 361-387. doi:10.1002/(sici)1097-0258(19960229)15:4<361::Aid-sim168>3.0.Co;2-4

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction* (Second edition, corrected 12th printing. ed.). New York: Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12*(1), 55-67. doi:10.1080/00401706.1970.10488634

Honda, T., & Härdle, W. K. (2014). Variable selection in Cox regression models with varying coefficients. *Journal of Statistical Planning and Inference, 148*, 67-81. doi:https://doi.org/10.1016/j.jspi.2013.12.002

Honda, T., & Yabe, R. (2017). Variable selection and structure identification for varying coefficient Cox models. *Journal of Multivariate Analysis, 161*, 103-122. doi:https://doi.org/10.1016/j.jmva.2017.07.007

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed. / David W. Hosmer, Stanley Lemeshow, Rodney X. Sturdivant. ed.). Hoboken, N.J: Wiley.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics, 7*(3), 355-373. doi:10.1093/biostatistics/kxj011

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674.

Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine, 23*(1), 77-91. doi:10.1002/sim.1593

Howell, S., Coory, M., Martin, J., & Duckett, S. (2009). Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Services Research, 9*, 96-96. doi:10.1186/1472-6963-9-96

Huang, J., & Harrington, D. (2002). Penalized Partial Likelihood Regression for Right-Censored Data with Bootstrap Selection of the Penalty Parameter. *Biometrics, 58*(4), 781-791. doi:10.1111/j.0006-341X.2002.00781.x

Huang, J., & Harrington, D. (2005). Iterative Partial Least Squares with Right-Censored Data Analysis: A Comparison to Other Dimension Reduction Techniques. *Biometrics, 61*(1), 17-24. doi:10.1111/j.0006-341X.2005.040304.x

Independent Hospital Pricing Authority. (2021a). *National Efficient Price Determination 2021-22*. Retrieved from https://www.ihpa.gov.au/sites/default/files/publications/national_efficient_price_determination_2021-22_0.pdf

Independent Hospital Pricing Authority. (2021b). *Pricing and funding for safety and quality: Risk adjusted models for avoidable hospital readmissions*. Retrieved from https://www.ihpa.gov.au/sites/default/files/publications/pricing_and_funding_for_safety_and_quality_-_avoidable_hospital_readmissions_2021-22_0.pdf

Ishwaran, H., & Kogalur, U. B. (2010). Consistency of random survival forests. *Statistics & Probability Letters, 80*(13), 1056-1064. doi:https://doi.org/10.1016/j.spl.2010.02.020

Ishwaran, H., & Kogalur, U. B. (2021). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). Retrieved from https://cran.r-project.org/package=randomForestSRC

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random Survival Forests. *The Annals of Applied Statistics, 2*(3), 841-860. Retrieved from http://www.jstor.org/stable/30245111

Jelinek, S., & Yunyongying, P. (2016). Predicting Hospital Readmission. *American family physician, 94*(4), 307-309.

Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *N Engl J Med, 360*(14), 1418-1428. doi:10.1056/NEJMsa0803563

Jerez-Aragonés, J. M., Gómez-Ruiz, J. A., Ramos-Jiménez, G., Muñoz-Pérez, J., & Alba-Conejo, E. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine, 27*(1), 45-63. doi:https://doi.org/10.1016/S0933-3657(02)00086-6

Jiang, S., Chin, K. S., Qu, G., & Tsui, K. L. (2018). An integrated machine learning framework for hospital readmission prediction. *Knowledge-Based Systems, 146*, 73-90. doi:10.1016/j.knosys.2018.01.027

Jin, Z., Lin, D. Y., & Ying, Z. (2006). On Least-Squares Regression with Censored Data. *Biometrika, 93*(1), 147-161. Retrieved from http://www.jstor.org/stable/20441267

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(2), 351-370. doi:doi:10.1111/j.1467-9868.2008.00639.x

Johnson, B. A. (2009). On lasso for censored data. *Electron. J. Statist., 3*, 485-506. doi:10.1214/08-EJS322

Jovanovic, M., Radovanovic, S., Vukicevic, M., Van Poucke, S., & Delibasic, B. (2016). Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. *Artificial Intelligence in Medicine, 72*, 12-21. doi:10.1016/j.artmed.2016.07.003

Kalagara, S., Eltorai, A. E. M., Durand, W. M., Mason DePasse, J., & Daniels, A. H. (2019). Machine learning modeling for predicting hospital readmission following lumbar laminectomy. *Journal of Neurosurgery: Spine, 30*(3), 344-352. doi:10.3171/2018.8.SPINE1869

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: A systematic review. *JAMA - Journal of the American Medical Association, 306*(15), 1688-1698. doi:10.1001/jama.2011.1515

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*(282), 457-481. doi:10.2307/2281868

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology, 18*(1), 18-24. doi:10.1186/s12874-018-0482-1

Khan, F. M., & Bayer-Zubek, V. (2008). *Support vector regression for censored data (SVRc): A novel tool for survival analysis*. Paper presented at the IEEE International Conference on Data Mining. https://ieeexplore.ieee.org/document/4781192/

Kim, D.-H., & Jeong, H.-C. (2006). Weighted LS-SVM Regression for Right Censored Data. *Communications for Statistical Applications and Methods, 13*(3), 765-776. doi:10.5351/CKSS.2006.13.3.765

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis A Self-Learning Text, Third Edition* (3rd ed. 2012. ed.). New York, NY: Springer New York : Imprint: Springer.

Koul, H., Susarla, V., & Van Ryzin, J. (1981). Regression Analysis with Randomly Right-Censored Data. *The Annals of Statistics, 9*(6), 1276-1288. Retrieved from http://www.jstor.org/stable/2240417

Koulouridis, I., Price, L. L., Madias, N. E., & Jaber, B. L. (2015). Hospital-acquired acute kidney injury and hospital readmission: A cohort study. *American Journal of Kidney Diseases, 65*(2), 275-282. doi:10.1053/j.ajkd.2014.08.024

Kretowska, M. (2004). Dipolar regression trees in survival analysis. *Biocybernetics and Biomedical Engineering, 24*(3), 25-33.

Kripalani, S., Theobald, C. N., Anctil, B., & Vasilevskis, E. E. (2014). Reducing hospital readmission rates: current strategies and future directions. *Annual review of medicine, 65*, 471-485. doi:10.1146/annurev-med-022613-090415

Kristensen, S. R., Bech, M., & Quentin, W. (2015). A roadmap for comparing readmission policies with application to Denmark, England, Germany and the United States. *Health Policy, 119*(3), 264-273. doi:https://doi.org/10.1016/j.healthpol.2014.12.009

Krompass, D., Esteban, C., Tresp, V., Sedlmayr, M., & Ganslandt, T. (2015). Exploiting Latent Embeddings of Nominal Clinical Data for Predicting Hospital Readmission. *Kunstliche Intelligenz, 29*(2), 153-159. doi:10.1007/s13218-014-0344-x

Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research, 20*(129), 1-30.

Lagoe, R. J., Noetscher, C. M., & Murphy, M. P. (2001). Hospital readmission: Predicting the risk. *Journal of Nursing Care Quality, 15*(4), 69-83. doi:10.1097/00001786-200107000-00008

Lapuerta, P., Azen, S., & Labree, L. (1995). *Use of Neural Networks in Predicting the Risk of Coronary Artery Disease* (Vol. 28).

Laurentiis, M., & Ravdin, P. (1994). Survival analysis of censored data: Neural network analysis detection of complex interactions between variables. *Breast Cancer Research and Treatment, 32*(1), 113-118. doi:10.1007/BF00666212

Leblanc, M., & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics, 48*(2), 411. doi:10.2307/2532300

Li, H., & Gui, J. (2004). Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics, 20*(suppl_1), i208-i215. doi:10.1093/bioinformatics/bth900

Li, H., & Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 65.

Lian, H., Lai, P., & Liang, H. (2013). Partially Linear Structure Selection in Cox Models with Varying Coefficients. *Biometrics, 69*(2), 348-357. doi:doi:10.1111/biom.12024

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Lin, D. Y., & Ying, Z. (1994). Semiparametric Analysis of the Additive Risk Model. *Biometrika, 81*(1), 61-71. doi:10.2307/2337050

Lin, D. Y., & Ying, Z. (1995). Semiparametric Analysis of General Additive-Multiplicative Hazard Models for Counting Processes. *The Annals of Statistics, 23*(5), 1712-1734. Retrieved from http://www.jstor.org/stable/2242542

Lisboa, P. J. G., Etchells, T. A., Jarman, I. H., Arsene, C. T. C., Hane Aung, M. S., Eleuteri, A., . . . Biganzoli, E. (2009). Partial logistic artificial neural network for competing risks regularized with automatic relevance determination. *IEEE transactions on neural networks, 20*(9), 1403. doi:10.1109/TNN.2009.2023654

Lisboa, P. J. G., Etchells, T. A., Jarman, I. H., Hane Aung, M. S., Chabaud, S., Bachelot, T., . . . Négrier, S. (2008). Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer. *Neural Networks, 21*(2), 414-426. doi:https://doi.org/10.1016/j.neunet.2007.12.034

Lisboa, P. J. G., Wong, H., Harris, P., & Swindell, R. (2003). A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine, 28*(1), 1-25. doi:10.1016/S0933-3657(03)00033-2

Liu, X. (2012). *Survival analysis : models and applications*. Chichester, West Sussex: Wiley/Higher Education Press.

Long, J. D., & Mills, J. A. (2018). Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC Medical Research Methodology, 18*(1), 138. doi:10.1186/s12874-018-0592-9

Low, L. L., Liu, N., Ong, M. E. H., Ng, E. Y., Ho, A. F. W., Thumboo, J., & Lee, K. H. (2017). Performance of the LACE index to identify elderly patients at high risk for hospital readmission in Singapore. *Medicine, 96*(19), e6728-e6728. doi:10.1097/MD.0000000000006728

Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. Paper presented at the Advances in Neural Information Processing Systems.

Ma, S., & Huang, J. (2007). Additive risk survival model with microarray data. *BMC Bioinformatics, 8*, 192-192. doi:10.1186/1471-2105-8-192

Malov, S. V., & O'Brien, S. J. (2018). Life table estimator revisited. *Communications in Statistics - Theory and Methods, 47*(9), 2126-2133. doi:10.1080/03610926.2017.1335418

Mariani, L., Coradini, D., Biganzoli, E., Boracchi, P., Marubini, E., Pilotti, S., . . . Rilke, F. (1997). Prognostic factors for metachronous contralateral breast cancer: A comparison of the linear Cox regression model and its artificial neural network extension. *Breast Cancer Research and Treatment, 44*(2), 167-178. doi:10.1023/A:1005765403093

Martinussen, T., & Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika, 89*(2), 283-298. doi:10.1093/biomet/89.2.283

McKeague, I. W., & Sasieni, P. D. (1994). A Partly Parametric Additive Risk Model. *Biometrika, 81*(3), 501-514. doi:10.2307/2337122

Mehta, H. B., Sieloff, E., Veeranki, S. P., Sura, S. D., Riall, T. S., Senagore, A. J., & Goodwin, J. S. (2017). Risk Prediction Models for Hospital Readmission In Surgery: A Systematic Review. *Journal of the American College of Surgeons, 225*(4), E113-E113. doi:10.1016/j.jamcollsurg.2017.07.832

Miller, W. D., Nguyen, K., Vangala, S., & Dowling, E. (2018). Clinicians can independently predict 30-day hospital readmissions as well as the LACE index. *BMC Health Services Research, 18*(1). doi:10.1186/s12913-018-2833-3

Millien, J. E., Townsend, M., Goldberg, J., & Fuhrman, G. M. (2017). An analysis of factors that predict hospital readmission after surgery for perforated appendicitis. *American Surgeon, 83*(9), 991-995. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85030478742&partnerID=40&md5=81605f21bc35a99b2445d3f7278475f6

Molinaro, A. M., Dudoit, S., & van Der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis, 90*(1), 154-177. doi:10.1016/j.jmva.2004.02.003

Morgan, D. J., Bame, B., Zimand, P., Dooley, P., Thom, K. A., Harris, A. D., . . . Liang, Y. (2019). Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. *JAMA network open, 2*(3), e190348. doi:10.1001/jamanetworkopen.2019.0348

Morrissey, E. F. R., McElnay, J. C., Scott, M., & McConnell, B. J. (2003). Influence of drugs, demographics and medical history on hospital readmission of elderly patients: A predictive model. *Clinical Drug Investigation, 23*(2), 119-128. doi:10.2165/00044011-200323020-00005

Mosquera, C., Vohra, N. A., Fitzgerald, T. L., & Zervos, E. E. (2016). Discharge with pancreatic fistula after pancreaticoduodenectomy independently predicts hospital readmission. *American Surgeon, 82*(8), 698-703. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020163161&partnerID=40&md5=ba00b77e90d0217a8aa971bfb7c85532

Nan, B., Lin, X., Lisabeth, L. D., & Harlow, S. D. (2005). A Varying-Coefficient Cox Model for the Effect of Age at a Marker Event on Age at Menopause. *Biometrics, 61*(2), 576-583. doi:doi:10.1111/j.1541-0420.2005.030905.x

Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems* (2nd ed.): Pearson education.

Nelson, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics, 14*(4), 945-966. doi:10.2307/1267144

Neumann, A., Holstein, J., Chatellier, G., & Lepage, E. (2004). A regression shrinkage method tailored to qualitative regressors and clustered data. *Statistics in Medicine, 23*(7), 1147-1157. doi:10.1002/sim.1663

Ohno-Machado, L. (1997). A comparison of Cox proportional hazards and artificial neural network models for medical prognosis. *Computers in Biology and Medicine, 27*(1), 55-65. doi:10.1016/S0010-4825(96)00036-4

Ottenbacher, K. J., Smith, P. M., Illig, S. B., Linn, R. T., Fiedler, R. C., & Granger, C. V. (2001). Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *Journal of Clinical Epidemiology, 54*(11), 1159-1165. doi:10.1016/S0895-4356(01)00395-X

Pack, Q. R., Priya, A., Lagu, T., Pekow, P. S., Engelman, R., Kent, D. M., & Lindenauer, P. K. (2016). Development and Validation of a Predictive Model for Short- and Medium-Term Hospital Readmission Following Heart Valve Surgery. *Journal of the American Heart Association, 5*(9). doi:10.1161/JAHA.116.003544

Padhukasahasram, B., Reddy, C. K., Li, Y., & Lanfear, D. E. (2015). Joint impact of clinical and behavioral variables on the risk of unplanned readmission and death after a heart failure hospitalization. *PLoS One, 10*(6), e0129553-e0129553. doi:10.1371/journal.pone.0129553

Parker, G., & Hadzi-Pavlovic, D. (1995). The Capacity of a Measure of Disability (the LSP) to Predict Hospital Readmission in Those with Schizophrenia. *Psychological Medicine, 25*(1), 157-163. doi:10.1017/S0033291700028178

Pavon, J. M., Zhao, Y., McConnell, E., & Hastings, S. N. (2014). Identifying risk of readmission in hospitalized elderly adults through inpatient medication exposure. *Journal of the American Geriatrics Society, 62*(6), 1116-1121. doi:10.1111/jgs.12829

Pederson, J. L., Majumdar, S. R., Forhan, M., Johnson, J. A., McAlister, F. A., Padwal, R., . . . for the, P. I. (2016). Current depressive symptoms but not history of depression predict hospital readmission or death after discharge from medical wards: A multisite prospective

cohort study. *General Hospital Psychiatry, 39*, 80-85. doi:10.1016/j.genhosppsych.2015.12.001

Perez, A., Chagin, K., Milinovich, A., Ji, X., Pavlescak, J., Kattan, M., . . . Starling, R. (2017). THE HOSPITAL READMISSION MODEL POORLY PREDICTS 30-DAY REHOSPITALIZATION IN HEART FAILURE. *Journal of the American College of Cardiology, 69*(11), 690-690. Retrieved from <Go to ISI>://WOS:000397342301212

Pham, H. N., Chatterjee, A., Narasimhan, B., Lee, C. W., Jha, D. K., Fai Wong, E. Y., . . . Chua, M. C. H. (2019). *Predicting hospital readmission patterns of diabetic patients using ensemble model and cluster analysis.* Paper presented at the Proceedings of 2019 International Conference on System Science and Engineering, ICSSE 2019.

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *Journal Machine Learning Research, 18*(1), 6673–6690.

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R: Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Radespiel-Troger, M., Gefeller, O., Rabenstein, T., & Hothorn, T. (2006). Association between split selection instability and predictive error in survival trees. *Methods Inf. Med., 45*(5), 548-556.

Radespiel-Tröger, M., Rabenstein, T., Schneider, H. T., & Lausen, B. (2003). Comparison of tree-based methods for prognostic stratification of survival data. *Artificial Intelligence in Medicine, 28*(3), 323-341. doi:https://doi.org/10.1016/S0933-3657(03)00060-5

Radovanović, S., Delibašić, B., Jovanović, M., Vukićević, M., & Suknović, M. (2019). A Framework for Integrating Domain Knowledge in Logistic Regression with Application to Hospital Readmission Prediction. *International Journal on Artificial Intelligence Tools, 28*(6). doi:10.1142/S0218213019600066

Radovanović, S., Vukićević, M., Kovačević, A., Štiglic, G., & Obradovic, Z. (2015). Domain knowledge Based Hierarchical Feature Selection for 30-Day Hospital Readmission Prediction. *Artificial Intelligence in Medicine, 9105*, 96-100. doi:10.1007/978-3-319-19551-3_11

Raines, B. T., Ponce, B. A., Reed, R. D., Richman, J. S., & Hawn, M. T. (2015). Hospital Acquired Conditions Are the Strongest Predictor for Early Readmission: An Analysis of 26,710 Arthroplasties. *Journal of Arthroplasty, 30*(8), 1299-1307. doi:10.1016/j.arth.2015.02.024

Rana, S., Tran, T., Luo, W., Phung, D., Kennedy, R. L., & Venkatesh, S. (2014). Predicting unplanned readmission after myocardial infarction from routinely collected administrative hospital data. *Australian Health Review, 38*(4), 377-382. doi:10.1071/AH14059

Ravdin, P., & Clark, G. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment, 22*(3), 285-293. doi:10.1007/BF01840841

Reddy, B. K., & Delen, D. (2018). Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. *Computers in Biology and Medicine, 101*, 199-209. doi:10.1016/j.compbiomed.2018.08.029

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939778

Ripley, R. M., Harris, A. L., & Tarassenko, L. (1998). Neural network models for breast cancer prognosis. *Neural Computing & Applications, 7*(4), 367-375. doi:10.1007/bf01428127

Ritov, Y. (1990). Estimation in a Linear Regression Model with Censored Data. *Ann. Statist., 18*(1), 303-328. doi:10.1214/aos/1176347502

Rubin, D. J., Handorf, E. A., Golden, S. H., Nelson, D. B., McDonnell, M. E., & Zhao, H. (2016). Development and validation of a novel tool to predict hospital readmission risk among patients with diabetes. *Endocrine Practice, 22*(10), 1204-1215. doi:10.4158/E161391.OR

Ruiz, B., García, M., Aguirre, U., & Aguirre, C. (2008). Factors predicting hospital readmissions related to adverse drug reactions. *European Journal of Clinical Pharmacology, 64*(7), 715-722. doi:10.1007/s00228-008-0473-y

Sabourin, C. B., & Funk, M. (1999). Readmission of patients after coronary artery bypass graft surgery. *Heart and Lung: Journal of Acute and Critical Care, 28*(4), 243-250. doi:10.1016/S0147-9563(99)70070-1

Scheike, T. H., & Zhang, M.-J. (2002). An Additive–Multiplicative Cox–Aalen Regression Model. *Scandinavian Journal of Statistics, 29*(1), 75-88. doi:https://doi.org/10.1111/1467-9469.00065

Scott, I. A., Shohag, H., & Ahmed, M. (2014). Quality of care factors associated with unplanned readmissions of older medical patients: a case-control study. *Intern Med J, 44*(2), 161-170. doi:10.1111/imj.12334

Shafer, A. (2019). Hospitalization Patterns over 30 Years Across a Statewide System of Public Mental Health Hospitals: Readmission Predictors, Optimal Follow-Up Period, Readmission Clusters and Individuals with Statistically Significant High Healthcare Utilization. *Psychiatric Quarterly, 90*(2), 263-273. doi:10.1007/s11126-019-9626-7

Shebeshi, D. S., Dolja-Gore, X., & Byles, J. (2020). Unplanned Readmission within 28 Days of Hospital Discharge in a Longitudinal Population-Based Cohort of Older Australian Women. *International journal of environmental research and public health, 17*(9), 3136. doi:10.3390/ijerph17093136

Shiao, H.-T., & Cherkassky, V. (2013). *SVM-Based Approaches for Predictive Modelling of Survival Data.* Paper presented at the The 2013 International Conference on Data Mining.

Shim, J., & Hwang, C. (2009). *Support vector censored quantile regression under random censoring* (Vol. 53).

Shimokawa, A., Kawasaki, Y., & Miyaoka, E. (2015). Comparison of splitting methods on survival tree. *The international journal of biostatistics, 11*(1), 175. doi:10.1515/ijb-2014-0029

Shivaswamy, P. K., Chu, W., & Jansche, M. (2007, 28-31 Oct. 2007). *A Support Vector Approach to Censored Targets.* Paper presented at the Seventh IEEE International Conference on Data Mining (ICDM 2007).

Shyu, Y., Chen, M., & Lee, H. (2002). Caregiver's needs predict hospital readmission for elderly patients. *Gerontologist, 42*, 11-11. Retrieved from <Go to ISI>://WOS:000179541400042

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software, 39*(5). doi:10.18637/jss.v039.i05

Song, R., Lu, W., Ma, S., & Jessie Jeng, X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika, 101*(4), 799-814. doi:10.1093/biomet/asu047

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software, 97*(1), 1-66. doi:10.18637/jss.v097.i01

Sparapani, R. A., Logan, B. R., McCulloch, R. E., & Laud, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine, 35*(16), 2741-2753. doi:10.1002/sim.6893

Steer, J., Gibson, J., & Bourke, S. (2012). Predicting hospital readmission in patients discharged following acute exacerbations of COPD (AECOPD). *European Respiratory Journal, 40*. Retrieved from <Go to ISI>://WOS:000449650904357

Steingrimsson, J. A., Diao, L., Molinaro, A., & Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in Medicine, 35*(20), 3595-3612. doi:doi:10.1002/sim.6949

Steingrimsson, J. A., Diao, L., & Strawderman, R. L. (2019). Censoring Unbiased Regression Trees and Ensembles. *Journal of the American Statistical Association, 114*(525), 370-383. doi:10.1080/01621459.2017.1407775

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., . . . Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for

traditional and novel measures. *Epidemiology, 21*(1), 128-138. doi:10.1097/EDE.0b013e3181c30fb2

Street, W. (1998). A Neural Network Model for Prognostic Prediction. *Proceedings of the Fifteenth International Conference on Machine Learning*, 540-546.

Su, X., & Fan, J. (2004). Multivariate Survival Trees: A Maximum Likelihood Approach Based on Frailty Models. *Biometrics, 60*(1), 93-99. doi:10.1111/j.0006-341X.2004.00139.x

Sundararaman, A., Valady Ramanathan, S., & Thati, R. (2018). Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance. *Big Data Research, 13*, 65-75. doi:10.1016/j.bdr.2018.05.004

Tabata, M., Shimizu, R., Kamekawa, D., Kato, M., Kamiya, K., Akiyama, A., . . . Masuda, T. (2014). Six-minute walk distance is an independent predictor of hospital readmission in patients with chronic heart failure. *International Heart Journal, 55*(4), 331-336. doi:10.1536/ihj.13-224

Taktak, A. F. G., Eleuteri, A., Aung, M. S. H., Lisboa, P. J. G., Desjardins, L., & Damato, B. E. (2009). Survival analysis in cancer using a partial logistic neural network model with Bayesian regularisation framework: a validation study. *International Journal of Knowledge Engineering and Soft Data Paradigms, 1*(3), 277. doi:10.1504/IJKESDP.2009.028819

Tan, P. C., Jacob, R., Quek, K. F., & Omar, S. Z. (2006). Readmission risk and metabolic, biochemical, haematological and clinical indicators of severity in hyperemesis gravidarum. *Australian and New Zealand Journal of Obstetrics and Gynaecology, 46*(5), 446-450. doi:10.1111/j.1479-828X.2006.00632.x

Theano Development Team. (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints, abs/1605.02688*. Retrieved from http://arxiv.org/abs/1605.02688

Therneau, T. (2020). A Package for Survival Analysis in R. Retrieved from https://CRAN.R-project.org/package=survival

Therneau, T., & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. Retrieved from https://CRAN.R-project.org/package=rpart

Thirlwell, S., Donovan, K. A., Turney, M., Emnett, C. E., Lamoreaux, A., & Portman, D. G. (2016). Predicting hospital readmissions in the oncology population. *Journal of Clinical Oncology, 34*(26). doi:10.1200/jco.2016.34.26_suppl.177

Tian, L., Zucker, D., & Wei, L. J. (2005). On the Cox Model With Time-Varying Regression Coefficients. *Journal of the American Statistical Association, 100*(469), 172-183. doi:10.1198/016214504000000845

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288. Retrieved from http://www.jstor.org/stable/2346178

Tibshirani, R. (1997). The Lasso Method for Variable Selection in The Cox Model. *Statistics in Medicine, 16*(4), 385-395. doi:doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(1), 91-108. doi:doi:10.1111/j.1467-9868.2005.00490.x

Troyano, A. M. R., Giner, J., Sousa, D., Merino, J. L., Alonso, A., Feliu, A., . . . Sibila, O. (2018). Predicting hospital readmissions in severe COPD patients using an electronic-nose. *European Respiratory Journal, 52*. doi:10.1183/13993003.congress-2018.PA3854

Tsiatis, A. A. (1990). Estimating Regression Parameters Using Linear Rank Tests for Censored Data. *Ann. Statist., 18*(1), 354-372. doi:10.1214/aos/1176347504

Turgeman, L., & May, J. H. (2016). A mixed-ensemble model for hospital readmission. *Artificial Intelligence in Medicine, 72*, 72-82. doi:10.1016/j.artmed.2016.08.005

Tyson, M., Patton, M., Salevitz, D., Chen, C., & Castle, E. (2014). CHANGE IN CO2 PREDICTS HOSPITAL READMISSION FOR FAILURE TO THRIVE AFTER RADICAL CYSTECTOMY: A SERIES OF OVER 600 PATIENTS. *Journal of Urology, 191*(4), E498-E498. doi:10.1016/j.juro.2014.02.1132

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine, 30*(10), 1105-1117. doi:10.1002/sim.4154

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. (2007). *Support Vector Machines for Survival Analysis.* Paper presented at the Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare.

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. (2008). *Survival SVM: A practical scalable algorithm*. Paper presented at the European Symposium on Artificial Neural Networks.

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. (2010). Additive survival least-squares support vector machines. *Statistics in Medicine, 29*(2), 296-308. doi:10.1002/sim.3743

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. K. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine, 53*(2), 107-118. doi:https://doi.org/10.1016/j.artmed.2011.06.006

van Walraven, C., Bennett, C., Jennings, A., Austin, P. C., & Forster, A. J. (2011). Proportion of hospital readmissions deemed avoidable: a systematic review. *Cmaj, 183*(7), E391-402. doi:10.1503/cmaj.101860

van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., . . . Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal, 182*(6), 551. doi:10.1503/cmaj.091117

Vashi, A. A., Fox, J. P., Carr, B. G., D'Onofrio, G., Pines, J. M., Ross, J. S., & Gross, C. P. (2013). Use of hospital-based acute care among patients recently discharged from the hospital. *Jama, 309*(4), 364-371. doi:10.1001/jama.2012.216219

Verweij, P. J. M., & Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine, 13*(23-24), 2427-2436. doi:doi:10.1002/sim.4780132307

Vinzamuri, B., & Reddy, C. K. (2013). *Cox regression with correlation based regularization for electronic health records*. Paper presented at the International Conference on Data Mining. https://ieeexplore.ieee.org/document/6729560/

Vukićević, M., Radovanović, S., Kovačević, A., Štiglic, G., & Obradovic, Z. (2015) Improving Hospital Readmission Prediction Using Domain Knowledge Based Virtual Examples. In*: Vol. 224. Lecture Notes in Business Information Processing* (pp. 695-706).

Wang, D., Jing, Z., He, K., & Garmire, L. X. (2021). Cox-nnet v2.0: improved neural-network based survival prediction extended to large-scale EMR data. *Bioinformatics*. doi:10.1093/bioinformatics/btab046

Wang, H., Cui, Z., Chen, Y., Avidan, M., Abdallah, A. B., & Kronzer, A. (2018). Predicting Hospital Readmission via Cost-Sensitive Deep Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15*(6), 1968-1978. doi:10.1109/TCBB.2018.2827029

Wang, H., Johnson, C., Robinson, R. D., Nejtek, V. A., Schrader, C. D., Leuck, J., . . . Zenarosa, N. R. (2016). Roles of disease severity and post-discharge outpatient visits as predictors of hospital readmissions. *BMC Health Services Research, 16*(1), 1-10. doi:10.1186/s12913-016-1814-7

Wang, H., Robinson, R. D., Johnson, C., Zenarosa, N. R., Jayswal, R. D., Keithley, J., & Delaney, K. A. (2014). Using the LACE index to predict hospital readmissions in congestive heart failure patients. *BMC Cardiovascular Disorders, 14*(1). doi:10.1186/1471-2261-14-97

Wang, S., Nan, B., Zhou, N., & Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika, 96*(2), 307-322. doi:10.1093/biomet/asp016

Wang, S., Nan, B., Zhu, J., & Beer, D. G. (2008). Doubly Penalized Buckley–James Method for Survival Data with High-Dimensional Covariates. *64*(1), 132-140. doi:doi:10.1111/j.1541-0420.2007.00877.x

Warchol, S. J., Monestime, J. P., Mayer, R. W., & Chien, W.-W. (2019). Strategies to Reduce Hospital Readmission Rates in a Non-Medicaid-Expansion State. *Perspectives in health*

information management, 16(Summer), 1a-1a. Retrieved from https://pubmed.ncbi.nlm.nih.gov/31423116

Weinland, C., Braun, B., Mühle, C., Kornhuber, J., & Lenz, B. (2017). Cloninger Type 2 Score and Lesch Typology Predict Hospital Readmission of Female and Male Alcohol-Dependent Inpatients During a 24-Month Follow-Up. *Alcoholism: Clinical and Experimental Research, 41*(10), 1760-1767. doi:10.1111/acer.13468

Weinreich, M., Nguyen, O. K., Wang, D., Mayo, H., Mortensen, E. M., Halm, E. A., & Makam, A. N. (2016). Predicting the risk of readmission in pneumonia: a systematic review of model performance. *Annals of the American Thoracic Society, 13*(9), 1607-1614. doi:10.1513/AnnalsATS.201602-135SR

Wolff, P., Grana, M., Ríos, S. A., & Yarza, M. B. (2019). Machine Learning Readmission Risk Modeling: A Pediatric Case Study. *BioMed Research International, 2019*. doi:10.1155/2019/8532892

Wong, F. K. Y., Chan, M. F., Chow, S., Chang, K., Chung, L., Lee, W. M., & Lee, R. (2010). What accounts for hospital readmission? *Journal of Clinical Nursing, 19*(23-24), 3334-3346. doi:10.1111/j.1365-2702.2010.03366.x

Wong, F. K. Y., Chow, S., Chung, L., Chang, K., Chan, T., Lee, W. M., & Lee, R. (2008). Can home visits help reduce hospital readmissions? Randomized controlled trial. *Journal of Advanced Nursing, 62*(5), 585-595. doi:10.1111/j.1365-2648.2008.04631.x

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *2017, 77*(1), 17. doi:10.18637/jss.v077.i01

Xiao, C., Ma, T., Dieng, A. B., Blei, D. M., & Wang, F. (2018). Readmission prediction via deep contextual embedding of clinical concepts. *PLoS One, 13*(4). doi:10.1371/journal.pone.0195024

Xu, R., & Adak, S. (2002). Survival Analysis with Time-Varying Regression Effects Using a Tree-Based Approach. *Biometrics, 58*(2), 305-315. doi:10.1111/j.0006-341X.2002.00305.x

Yan, F., Lin, X., Li, R., & Huang, X. (2018). Functional principal components analysis on moving time windows of longitudinal data: dynamic prediction of times to event. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 67*(4), 961-978. doi:https://doi.org/10.1111/rssc.12264

Yan, J., & Huang, J. (2012). Model Selection for Cox Models with Time-Varying Coefficients. *Biometrics, 68*(2), 419-428. doi:doi:10.1111/j.1541-0420.2011.01692.x

Yang, M., Hu, X., Wang, H., Zhang, L., Hao, Q., & Dong, B. (2017). Sarcopenia predicts readmission and mortality in elderly patients in acute care wards: a prospective study. *Journal of Cachexia, Sarcopenia and Muscle, 8*(2), 251-258. doi:10.1002/jcsm.12163

Ying, Z. (1993). A Large Sample Study of Rank Estimation for Censored Regression Data. *Ann. Statist., 21*(1), 76-99. doi:10.1214/aos/1176349016

Yoo, J. W., Jabeen, S., Bajwa, T., Jr., Kim, S. J., Leander, D., Hasan, L., . . . Khan, A. (2015). Hospital readmission of skilled nursing facility residents: A systematic review. *Research in Gerontological Nursing, 8*(3), 148-156. doi:10.3928/19404921-20150129-01

Yu, S., Farooq, F., van Esbroeck, A., Fung, G., Anand, V., & Krishnapuram, B. (2015). Predicting readmission risk with institution-specific prediction models. *Artificial Intelligence in Medicine, 65*(2), 89-96. doi:https://doi.org/10.1016/j.artmed.2015.08.005

Zhang, H. H., & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika, 94*(3), 691-703. doi:10.1093/biomet/asm037

Zhang, J., Lam, S. S., & Poranki, S. (2013). A classification model for hospital readmission using combined neural networks. In (pp. 1088-1097): Institute of Industrial Engineers.

Zhang, J., Yoon, S. W., Khasawneh, M. T., Srihari, K., & Poranki, S. (2013). *Hospital readmission prediction using swarm intelligence-based support vector machines.* Paper presented at the IIE Annual Conference and Expo 2013.

Zhang, S., Wang, L., & Lian, H. (2014). Estimation by polynomial splines with variable selection in additive Cox models. *Statistics, 48*(1), 67-80. doi:10.1080/02331888.2012.748770

Zhao, P., & Yoo, I. (2017). *A self-Adaptive 30-day diabetic readmission prediction model based on incremental learning.* Paper presented at the Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017.

Zheng, S., Hanchate, A., & Shwartz, M. (2019). One-year costs of medical admissions with and without a 30-day readmission and enhanced risk adjustment. *BMC Health Services Research, 19*(1), 155. doi:10.1186/s12913-019-3983-7

Zhu, K., Lou, Z., Zhou, J., Ballester, N., Kong, N., & Parikh, P. (2015). Predicting 30-day Hospital Readmission with Publicly Available Administrative Database. A Conditional Logistic Regression Modeling Approach. *Methods Inf Med, 54*(6), 560-567. doi:10.3414/me14-02-0017

Zhu, R. (2013). *Tree-based methods for survival analysis and high-dimensional data.* (Dissertation). University of North Carolina at Chapel Hill, Available from UNC-Chapel Hill Carolina Digital Repository database.

Zhu, R., & Kosorok, M. R. (2012). Recursively Imputed Survival Trees. *Journal of the American Statistical Association, 107*(497), 331-340. doi:10.1080/01621459.2011.637468

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association, 101*(476), 1418-1429. doi:10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320. doi:doi:10.1111/j.1467-9868.2005.00503.x

10 Appendices

Appendix A Variables used in ANN Training

Table 34. Variables used in ANNs

Variable	GCUH	RH
AdmitWardCode1	✓	✓
Age	✓	✓
ED_NumPresPrevYear	✓	✓
ED_NumPresSincePrevAdm	×	✓
ED_NumPresSincePrevAdmALL	×	✓
GenderCode	✓	✓
iGC	✓	✓
Inpat_NumAdmPrevYearALL	✓	✓
Inpat_PrevAdmLOSPrevYear	✓	✓
Inpat_PrevAdmLOSPrevYearALL	×	✓
Inpat_TimeSincePrevAdmALL	✓	✓
Inpat_TotalAdmInICU	×	✓
Inpat_TotalAdmInICUALL	×	✓
Inpat_TotalTimeAdmPrevYear	✓	×
Inpat_TotalTimeAdmPrevYearALL	✓	×
LOSCalc	✓	×
Outp_NumApptPrevYear	✓	✓
Outp_NumApptSincePrevAdm	✓	×
Outp_NumApptSincePrevAdmALL	✓	×

Appendix B Final Model Settings

This appendix tabulates the final settings for models constructed in this research for each hospital and research question.

- Appendix B-1 Cox Regression
- Appendix B-2 Logistic Regression (AIC)
- Appendix B-3 Logistic Regression (BIC)
- Appendix B-4 Survival Tree (One-Step Likelihood)
- Appendix B-5 Survival Tree (Log Rank Statistic)
- Appendix B-6 CURT (V1)
- Appendix B-7 CURT (V2)
- Appendix B-8 Random Survival Forest
- Appendix B-9 CURE
- Appendix B-10 RIST
- Appendix B-11 BART
- Appendix B-12 NNET Survival
- Appendix B-13 Time-Coded ANN
- Appendix B-14 Hybrid Cox-ANN

Appendix B-1 Cox Regression

Table 35. Terms in final Cox Regression models

Technique	Final Terms Included	
	GCUH	**RH**
Cox regression	Age LOSCalc iGC AdmitWardCode1 Inpat_PrevAdmLOSPrevYear Inpat_NumAdmPrevYearALL Inpat_TimeSincePrevAdmALL Inpat_TotalTimeAdmPrevYearALL ED_NumPresPrevYear Outp_NumApptPrevYear LOSCalc^2 Outp_NumApptPrevYear^2 ED_NumPresPrevYear^2 ED_NumPresPrevYear^3 Inpat_TimeSincePrevAdmALL^2 Age:Inpat_NumAdmPrevYearALL Inpat_NumAdmPrevYearALL:Outp_NumApptPrevYear Age:iGC LOSCalc:Inpat_TimeSincePrevAdmALL Age:Inpat_PrevAdmLOSPrevYear iGC:Inpat_NumAdmPrevYearALL Age:LOSCalc Inpat_PrevAdmLOSPrevYear:Inpat_TotalTimeAdmPrevYearALL LOSCalc:ED_NumPresPrevYear	Age iGC AdmitWardCode1 Inpat_PrevAdmLOSPrevYear Inpat_NumAdmPrevYearALL Inpat_TimeSincePrevAdmALL ED_NumPresPrevYear Outp_NumApptPrevYear Outp_NumApptPrevYear^2 ED_NumPresPrevYear^2 ED_NumPresPrevYear^3 Age:Inpat_NumAdmPrevYearALL Inpat_PrevAdmLOSPrevYear:ED_NumPresPrevYear Age:Inpat_PrevAdmLOSPrevYear Inpat_PrevAdmLOSPrevYear:Outp_NumApptPrevYear iGC:Outp_NumApptPrevYear Inpat_PrevAdmLOSPrevYear:Inpat_TimeSincePrevAdmALL Inpat_TimeSincePrevAdmALL:Outp_NumApptPrevYear Age:Inpat_TimeSincePrevAdmALL Age:iGC

Appendix B-2 Logistic Regression (AIC)

Table 36. Terms in final Logistic Regression (AIC) models

Technique	Final Terms Included	
	GCUH	**RH**
Logistic regression	Age	Age
	LOSCalc	iGC
	iGC	GenderCode
	GenderCode	AdmitWardCode1
	AdmitWardCode1	Inpat_TotalAdmInICU
	Inpat_TotalTimeAdmPrevYear	Inpat_NumAdmPrevYearALL
	Inpat_NumAdmPrevYearALL	Inpat_TotalAdmInICUALL
	Inpat_TimeSincePrevAdmALL	Inpat_TimeSincePrevAdmALL
	ED_NumPresPrevYear	Inpat_PrevAdmLOSPrevYearALL
	Outp_NumApptPrevYear	ED_NumPresPrevYear
	Outp_NumApptSincePrevAdm	ED_NumPresSincePrevAdm
	Outp_NumApptSincePrevAdmALL	ED_NumPresSincePrevAdmALL
	LOSCalc^2	Outp_NumApptPrevYear
	ED_NumPresPrevYear^2	Outp_NumApptPrevYear^3
	ED_NumPresPrevYear^3	Inpat_TimeSincePrevAdmALL^3
	LOSCalc^3	Inpat_PrevAdmLOSPrevYearALL^3
	Inpat_TimeSincePrevAdmALL^2	Age:Inpat_NumAdmPrevYearALL
	Inpat_TotalTimeAdmPrevYear^2	Inpat_PrevAdmLOSPrevYearALL:Outp_NumApptPrevYear
	Age:Inpat_NumAdmPrevYearALL	ED_NumPresPrevYear:ED_NumPresSincePrevAdm
	Inpat_NumAdmPrevYearALL:Outp_NumApptPrevYear	Inpat_PrevAdmLOSPrevYearALL:ED_NumPresPrevYear
	LOSCalc:Inpat_TimeSincePrevAdmALL	iGC:ED_NumPresSincePrevAdm
	Age:iGC	GenderCode:Inpat_PrevAdmLOSPrevYearALL
	ED_NumPresPrevYear:Outp_NumApptSincePrevAdm	Inpat_TotalAdmInICUALL:ED_NumPresSincePrevAdmALL
	iGC:Inpat_NumAdmPrevYearALL	Inpat_TotalAdmInICU:ED_NumPresSincePrevAdm
	LOSCalc:Inpat_TotalTimeAdmPrevYear	Inpat_TotalAdmInICUALL:Inpat_TimeSincePrevAdmALL

Table 36 continued

Technique	Final Terms Included	
	GCUH	**RH**
Logistic regression	Inpat_TimeSincePrevAdmALL:Outp_NumApptSincePrevAdmALL	Inpat_NumAdmPrevYearALL:Inpat_PrevAdmLOSPrevYearALL
	AdmitWardCode1:Outp_NumApptSincePrevAdm	Inpat_TimeSincePrevAdmALL:Inpat_PrevAdmLOSPrevYearALL
	Age:LOSCalc	Inpat_NumAdmPrevYearALL:ED_NumPresPrevYear
	AdmitWardCode1:ED_NumPresPrevYear	AdmitWardCode1:Outp_NumApptPrevYear
	Outp_NumApptSincePrevAdm:Outp_NumApptSincePrevAdmALL	Inpat_TotalAdmInICUALL:ED_NumPresSincePrevAdm
	Age:Inpat_TotalTimeAdmPrevYear	Age:iGC
	GenderCode:Inpat_TotalTimeAdmPrevYear	iGC:Outp_NumApptPrevYear
	LOSCalc:Inpat_NumAdmPrevYearALL	
	Inpat_TotalTimeAdmPrevYear:Inpat_TimeSincePrevAdmALL	

Appendix B-3 Logistic Regression (BIC)

Table 37. Terms in final Logistic Regression (BIC) models

Technique	Final Terms Included	
	GCUH	**RH**
Logistic regression	Age LOSCalc iGC Inpat_NumAdmPrevYearALL Inpat_TimeSincePrevAdmALL ED_NumPresPrevYear Outp_NumApptPrevYear LOSCalc^2) Age:Inpat_NumAdmPrevYearALL Inpat_NumAdmPrevYearALL:Outp_NumApptPrevYear LOSCalc:Inpat_TimeSincePrevAdmALL Age:iGC	Age AdmitWardCode1 Inpat_NumAdmPrevYearALL Inpat_TimeSincePrevAdmALL ED_NumPresPrevYear Age:Inpat_NumAdmPrevYearALL

Appendix B-4 Survival Tree (One Step Likelihood)

Table 38. Parameter values for final Survival Tree (One Step Likelihood) models

Model Type	Parameters	RQ1		RQ2	
		GCUH	**RH**	**GCUH**	**RH**
Survival Tree – One Step Likelihood	Cost-complexity parameter	0.0003	0.00065	0.0004	0.001

Appendix B-5 Survival Tree (Log Rank Statistic)

Table 39. Parameter values for final Survival Tree (Log Rank Statistic) models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
Survival Tree – Log Rank Statistic	Node depth	5	6	7	6

Appendix B-6 CURT (V1)

Table 40. Parameter values for final CURT (V1) models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
CURT	Conditional survival function	Survival Tree	Survival Tree	Survival Tree	Survival Tree

Appendix B-7 CURT (V2)

Table 41. Parameter values for final CURT (V2) models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
CURT	Conditional survival function	Survival Tree	Survival Tree	Survival Tree	Survival Tree
	Cost-complexity parameter	0.000095	0.0008	0.000035	0.00045

Appendix B-8 Random Survival Forest

Table 42. Parameter values for final Random Survival Forest models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
Random Survival Forest	Number of trees	500	1000	1000	750
	Covariates considered at each split	2	2	3	3
	Terminal node size	15	15	15	15

Appendix B-9 CURE

Table 43. Parameter values for final CURE models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
CURE	Conditional survival function	Random Survival Forest	Survival Tree	Random Survival Forest	Survival Tree
	Number of trees	500	1000	750	500
	Covariates considered at each split	3	4	5	6
	Terminal node size	20	20	20	20

Appendix B-10 RIST

Table 44. Parameter values for final RIST models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
RIST	Number of trees	70	40	60	60
	Covariates considered at each split	3	3	7	5
	Terminal node size	20	30	20	20
	Imputation cycles	3	1	2	1

Appendix B-11 BART

Table 45. Parameter values for final BART models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
BART	Number of trees	50	50	50	50
	Draws from the posterior	200	500	200	500
	Burn in sample	250	250	250	250
	Thinning	10	10	10	10

Appendix B-12 NNET Survival

Table 46. Parameter values for final NNET Survival models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
NNET Survival	Hidden layers and nodes	2 layers, 15 and 10 nodes	2 layers, 15 and 10 nodes	2 layers, 15 and 10 nodes	2 layers, 15 and 10 nodes
	Epochs	1000	300	600	1100
	Mini-batch size	256	256	256	128
	Regularisation penalty (L2)	$\exp(-6)$	$\exp(-4)$	$\exp(-5)$	$\exp(-5)$
	Intervals	20	20	20	40

Appendix B-13 Time-Coded ANN

Table 47. Parameter values for final Time-Coded ANN models

Model Type	Parameters	RQ1 GCUH	RH	RQ2 GCUH	RH
Time-Coded ANN	Hidden layers and nodes	1 layer, 15 nodes	1 layer, 15 nodes	1 layer, 10 nodes	1 layer, 20 nodes
	Epochs	400	1400	700	1300
	Mini-batch size	8192	8192	2048	2048
	Regularisation penalty (L2)	$\exp(-6)$	$\exp(-6)$	$\exp(-6)$	$\exp(-6)$

Appendix B-14 Hybrid Cox-ANN

Table 48. Parameter values for final Cox NNET models

Model Type	Parameters	RQ1		RQ2	
		GCUH	**RH**	**GCUH**	**RH**
Hybrid Cox-ANN (Cox NNET)	Hidden layers and nodes	1 layer, 14 nodes	1 layer, 8 nodes	1 layer, 14 nodes	2 layers, 7 and 4 nodes
	Epochs	1000	1000	1000	1000
	Regularisation penalty (L2)	$\exp(-5)$	$\exp(-5)$	$\exp(-6)$	$\exp(-6)$
	Batch normalisation	Yes	No	No	No

Appendix C Model Selection Figures

This appendix presents figures illustrating model performance for each hospital and research question as hyperparameters were varied.

- Appendix C-1 Survival Tree (One-Step Likelihood)
- Appendix C-2 Survival Tree (Log Rank Statistic)
- Appendix C-3 CURT (V1)
- Appendix C-4 CURT (V2)
- Appendix C-5 Random Survival Forest
- Appendix C-6 CURE
- Appendix C-7 RIST
- Appendix C-8 NNET Survival
- Appendix C-9 Time-Coded ANN
- Appendix C-10 Hybrid Cox-ANN

Appendix C-1 Survival Tree (One Step Likelihood)

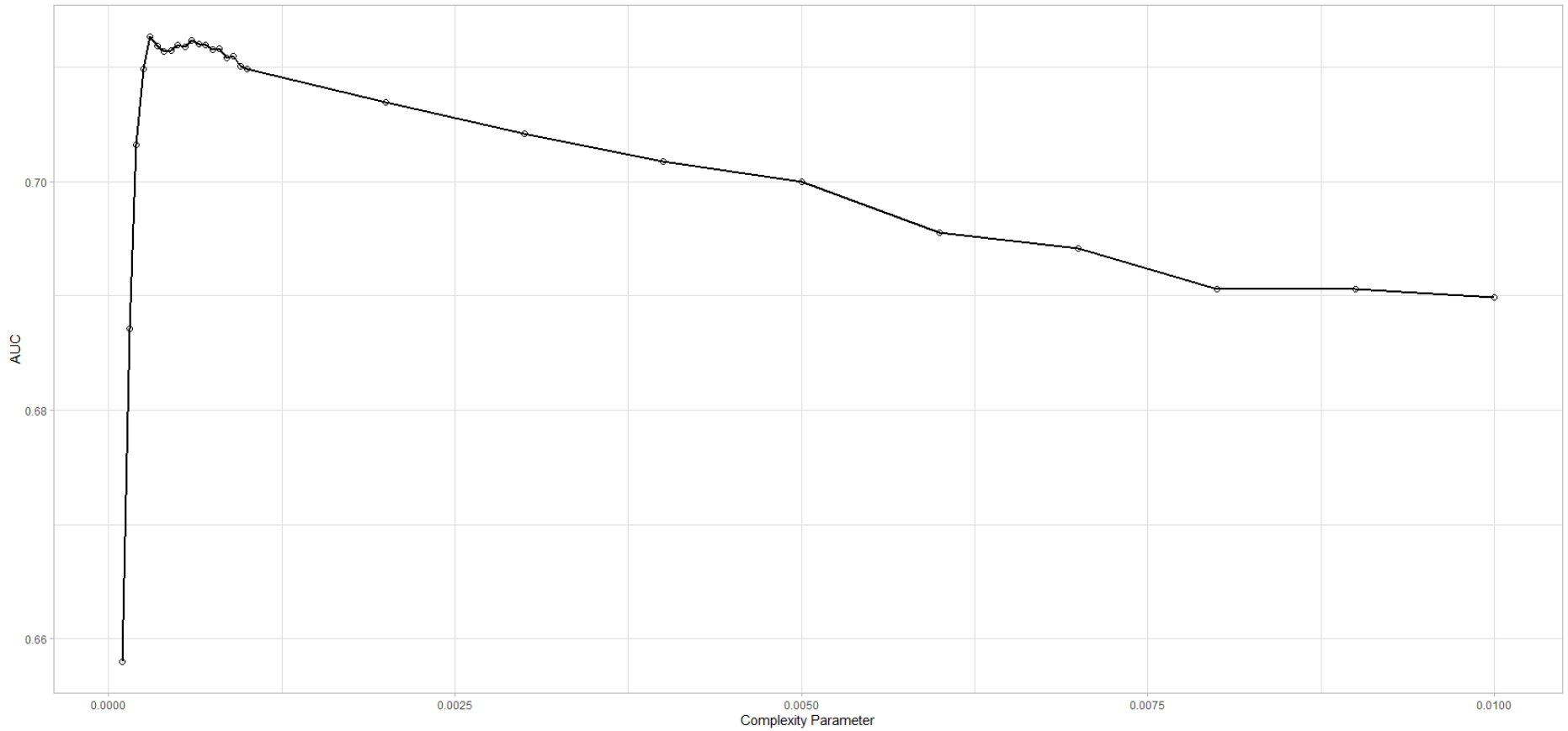

Figure 6. Survival Tree (One Step Likelihood) - RQ1 GCUH Model Selection

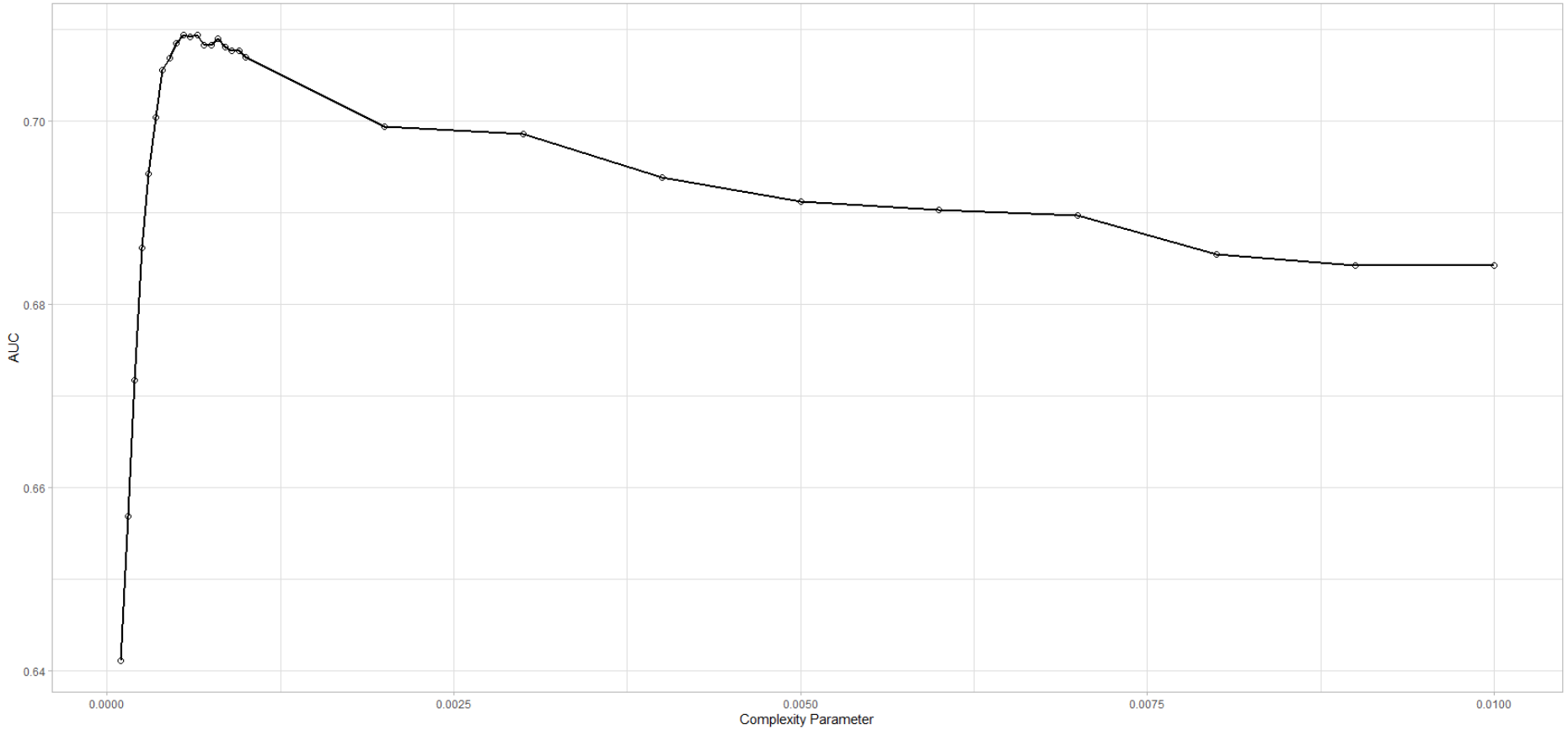

Figure 7. Survival Tree (One Step Likelihood) - RQ1 RH Model Selection

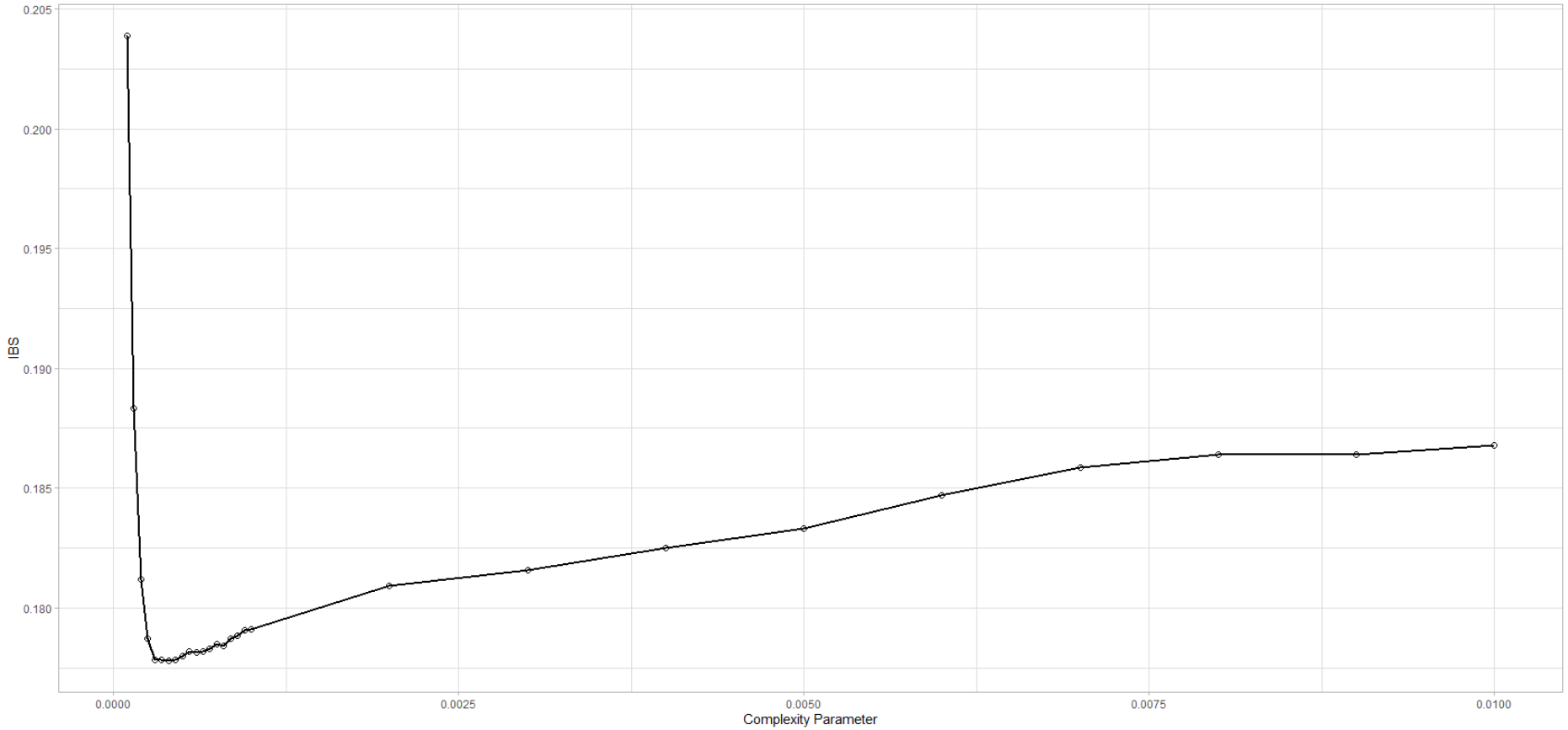

Figure 8. Survival Tree (One Step Likelihood) - RQ2 GCUH Model Selection

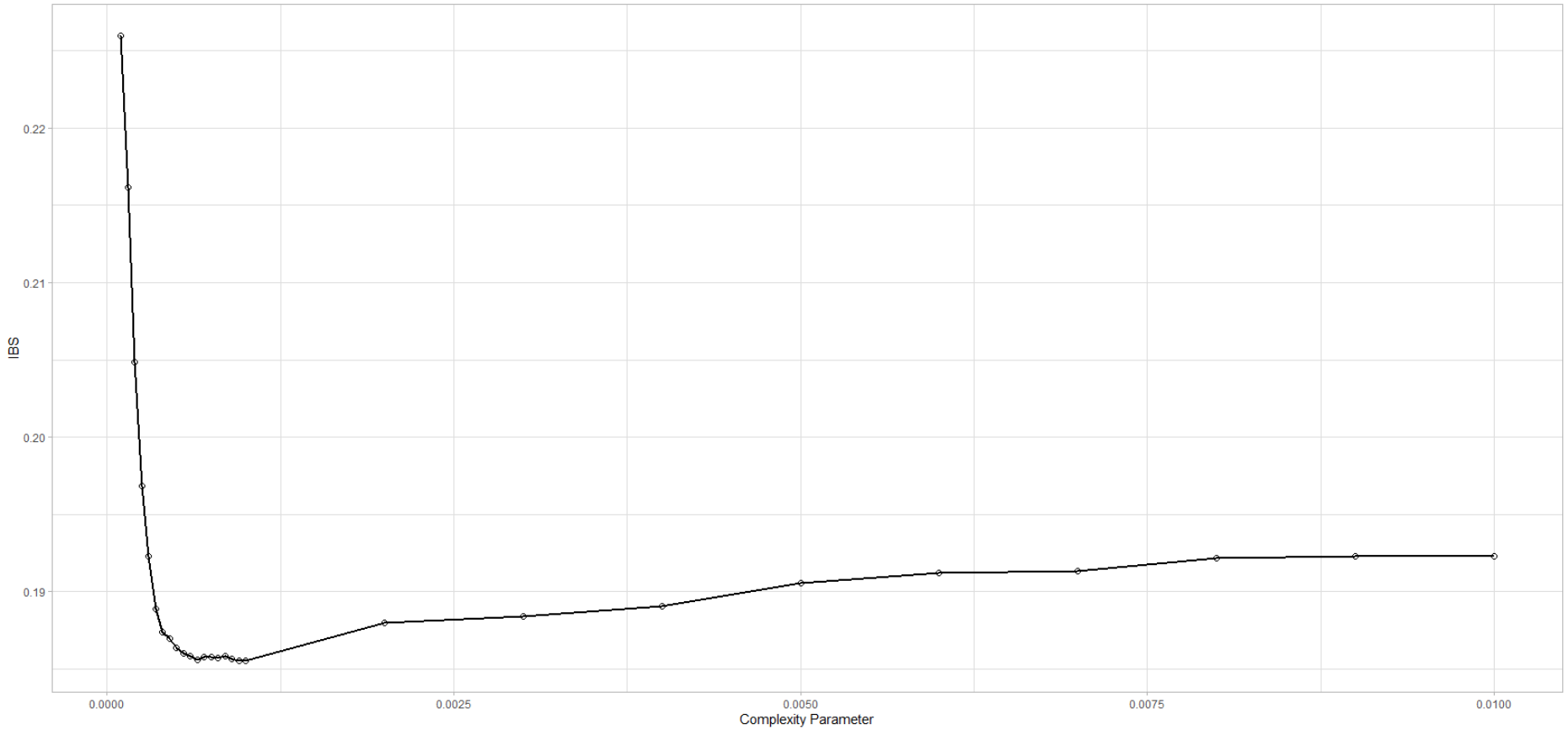

Figure 9. Survival Tree (One Step Likelihood) - RQ2 RH Model Selection

Appendix C-2 Survival Tree (Log Rank Statistic)

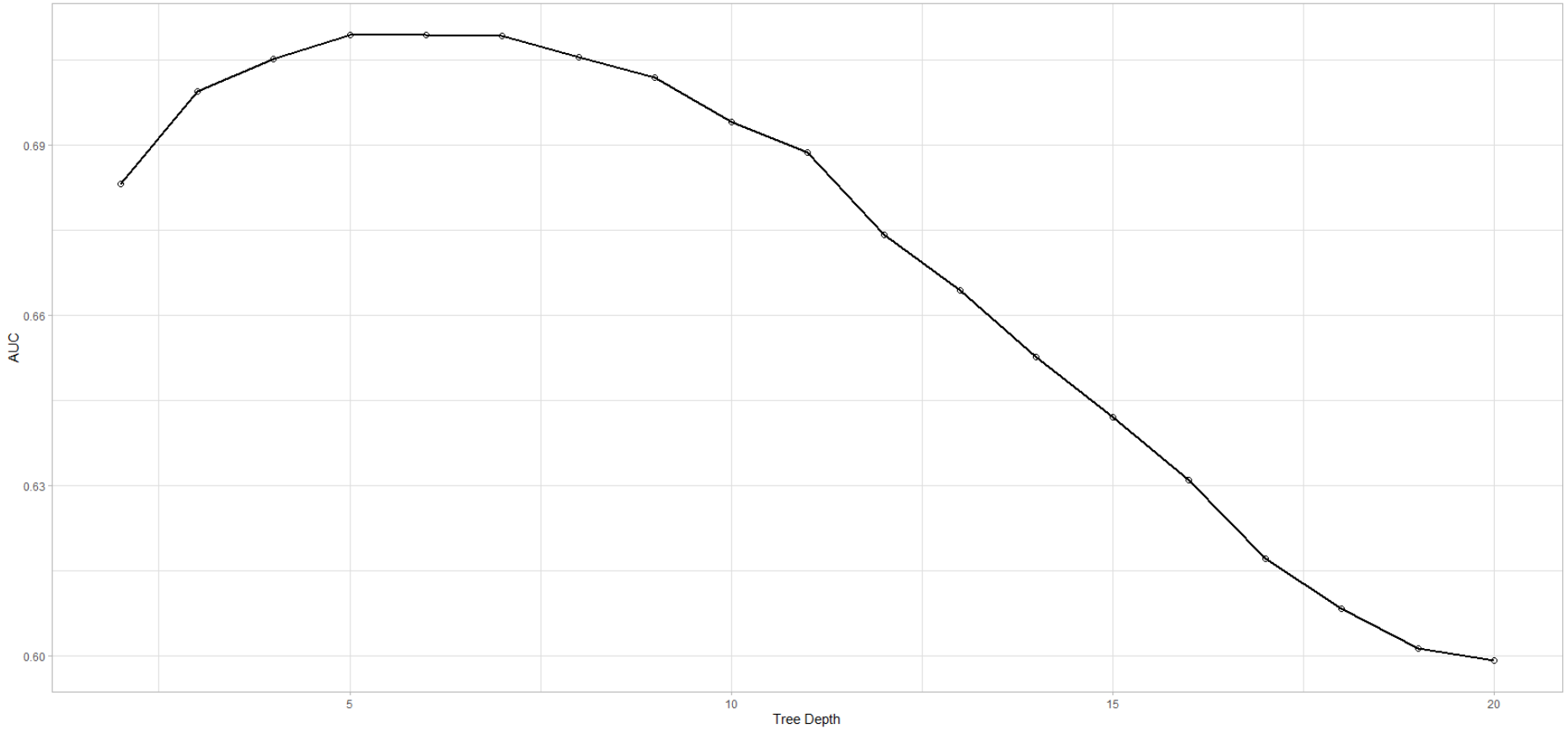

Figure 10. Survival Tree (Log Rank Statistic) - RQ1 GCUH Model Selection

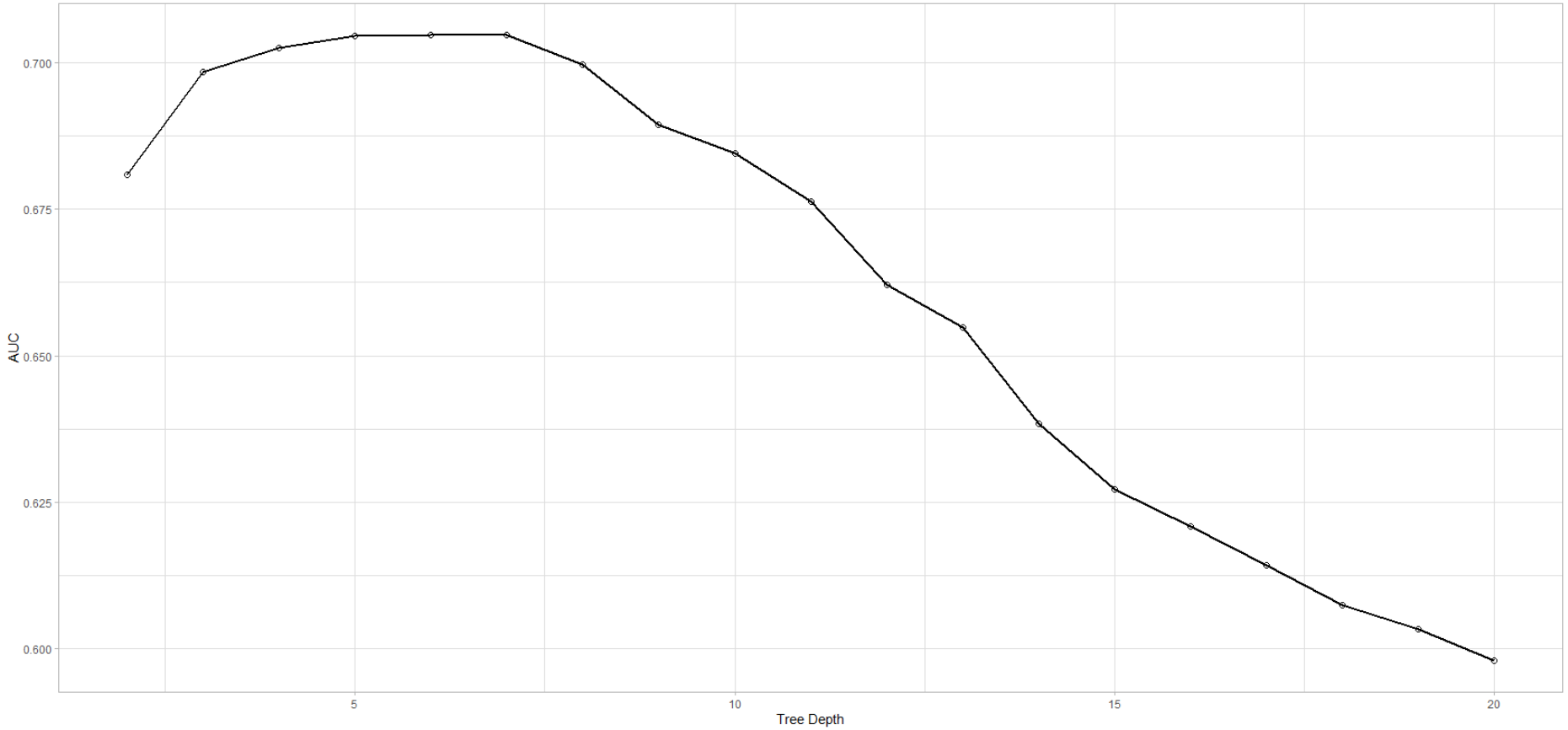

Figure 11. Survival Tree (Log Rank Statistic) - RQ1 RH Model Selection

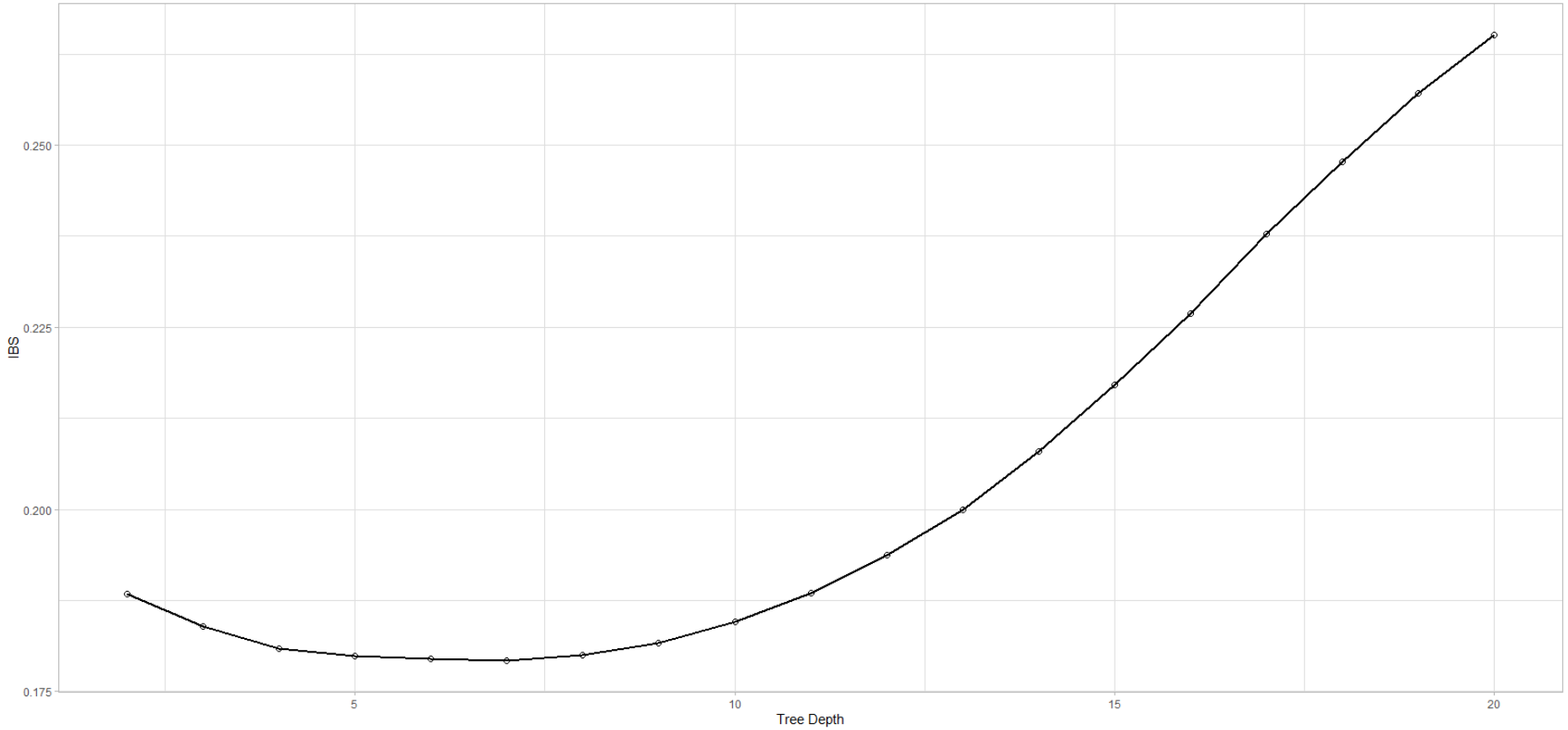

Figure 12. Survival Tree (Log Rank Statistic) - RQ2 GCUH Model Selection

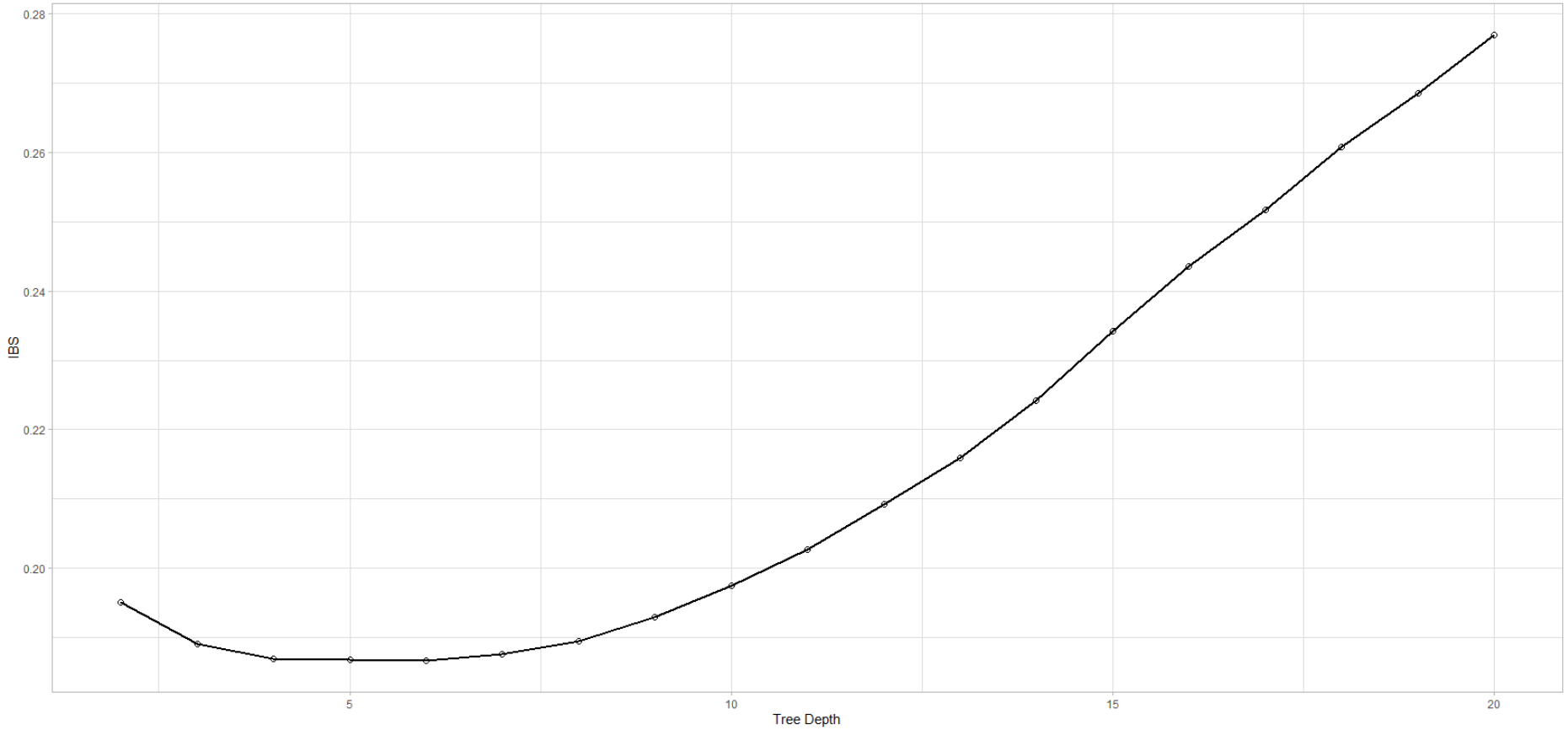

Figure 13. Survival Tree (Log Rank Statistic) - RQ2 RH Model Selection

Appendix C-3 CURT (V1)

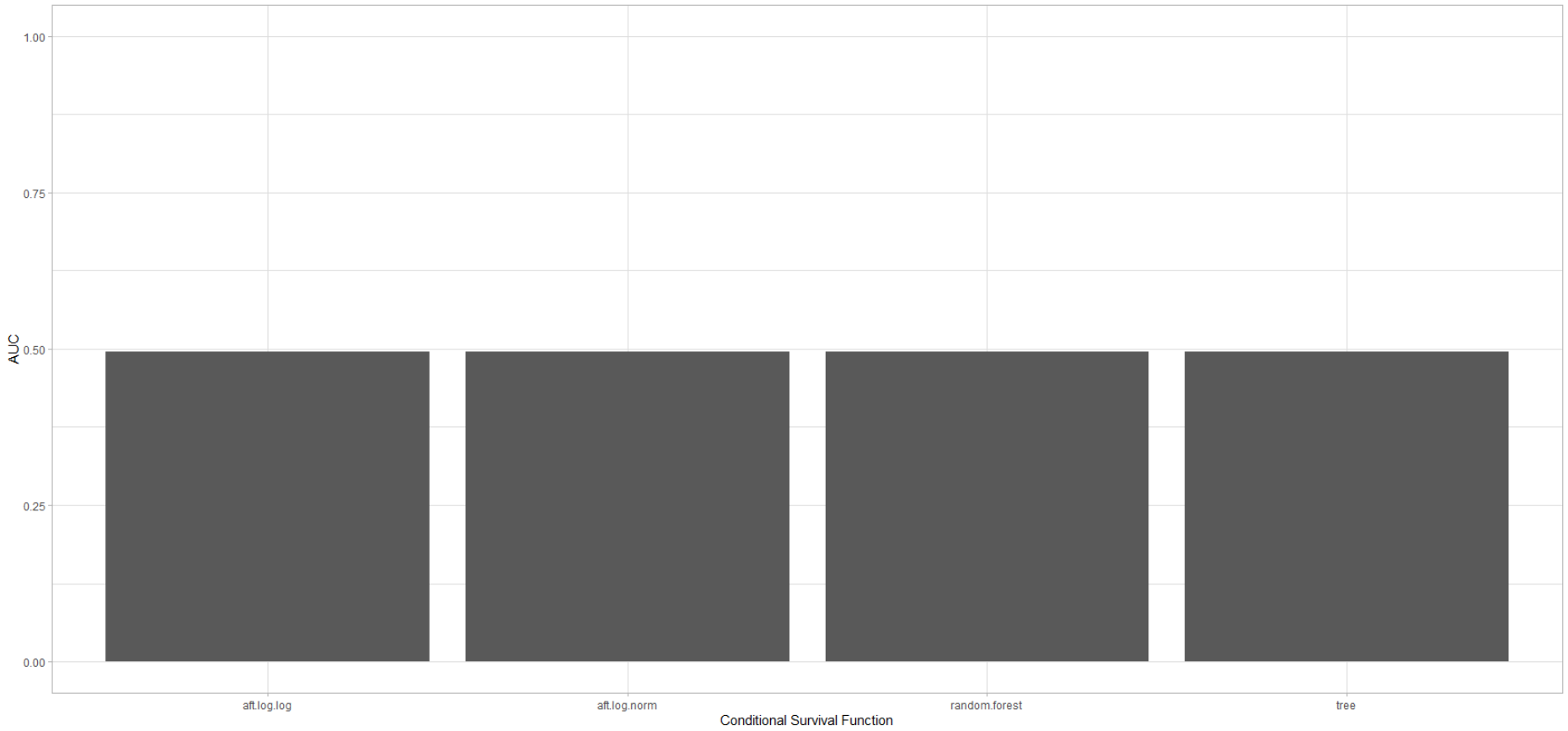

Figure 14. CURT (V1) - RQ1 GCUH Model Selection

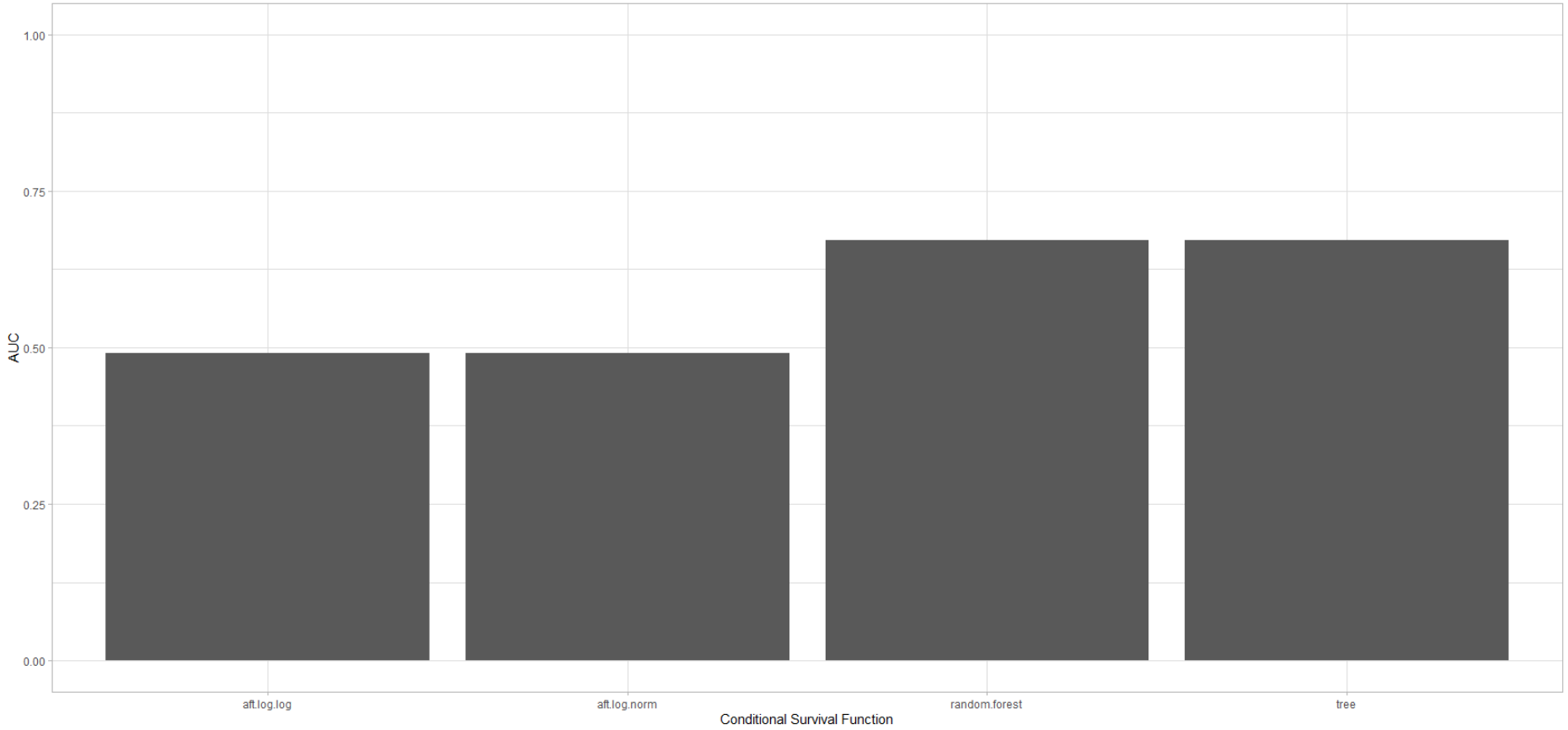

Figure 15. CURT (V1) - RQ1 RH Model Selection

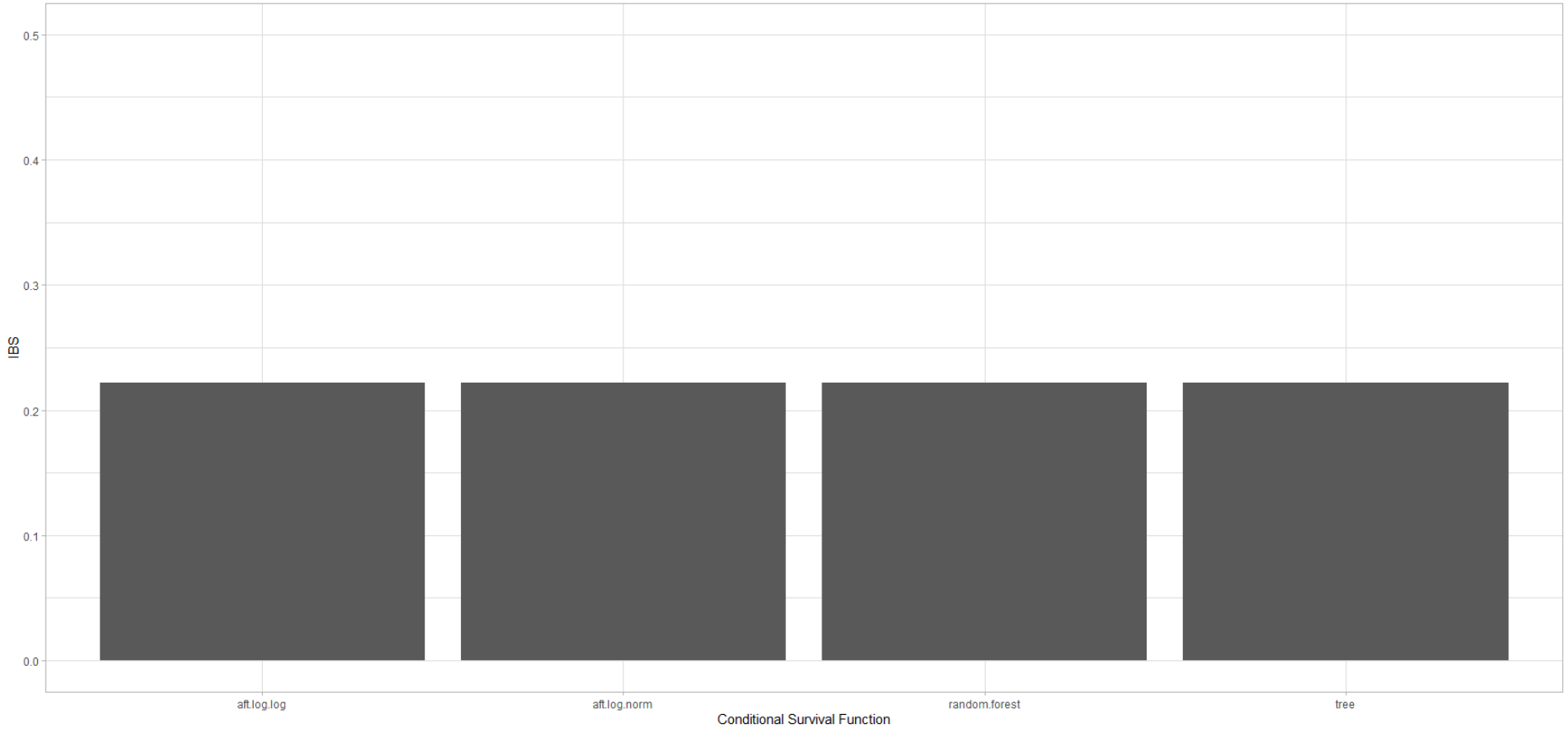

Figure 16. CURT (V1) - RQ2 GCUH Model Selection

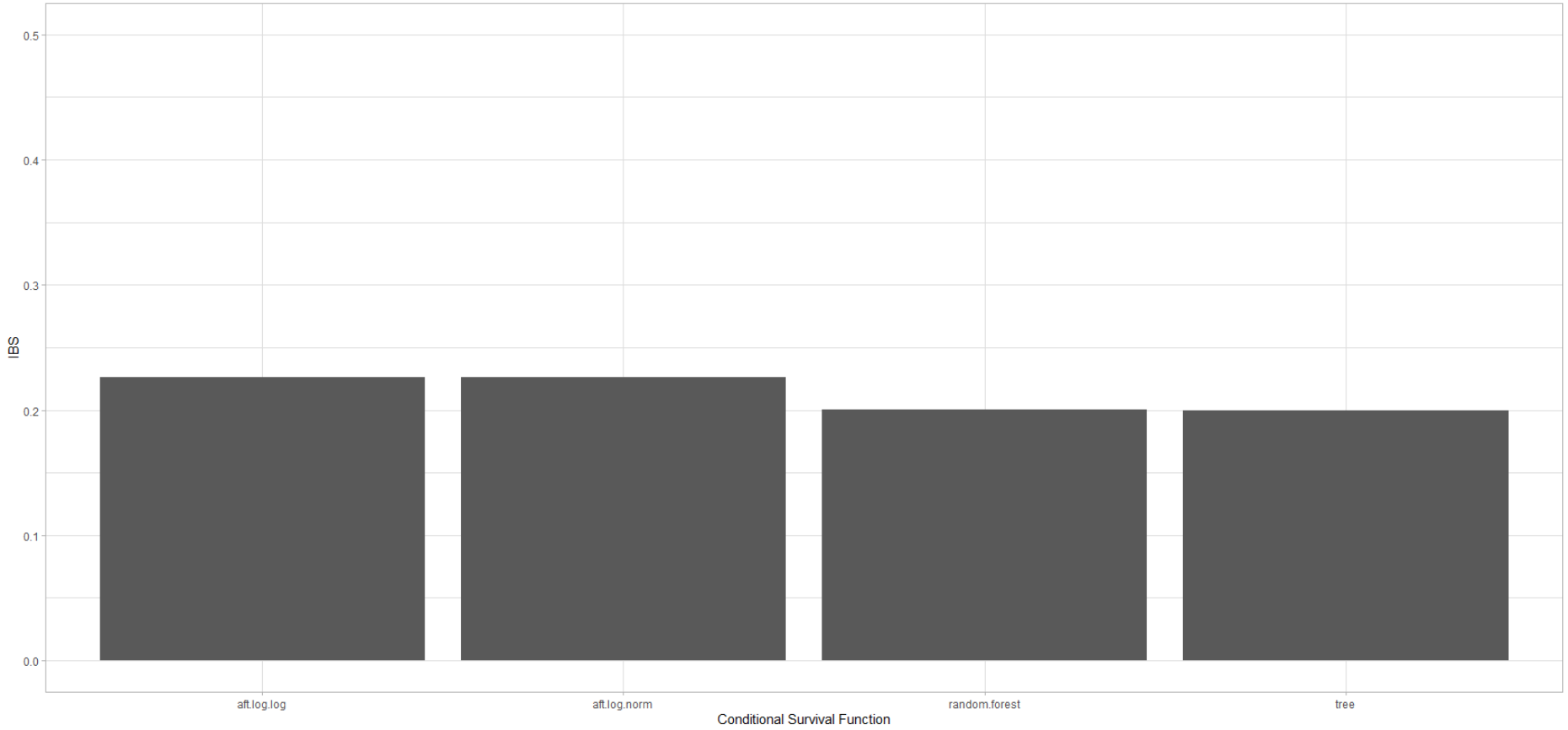

Figure 17. CURT (V1) - RQ2 RH Model Selection

Appendix C-4 CURT (V2)

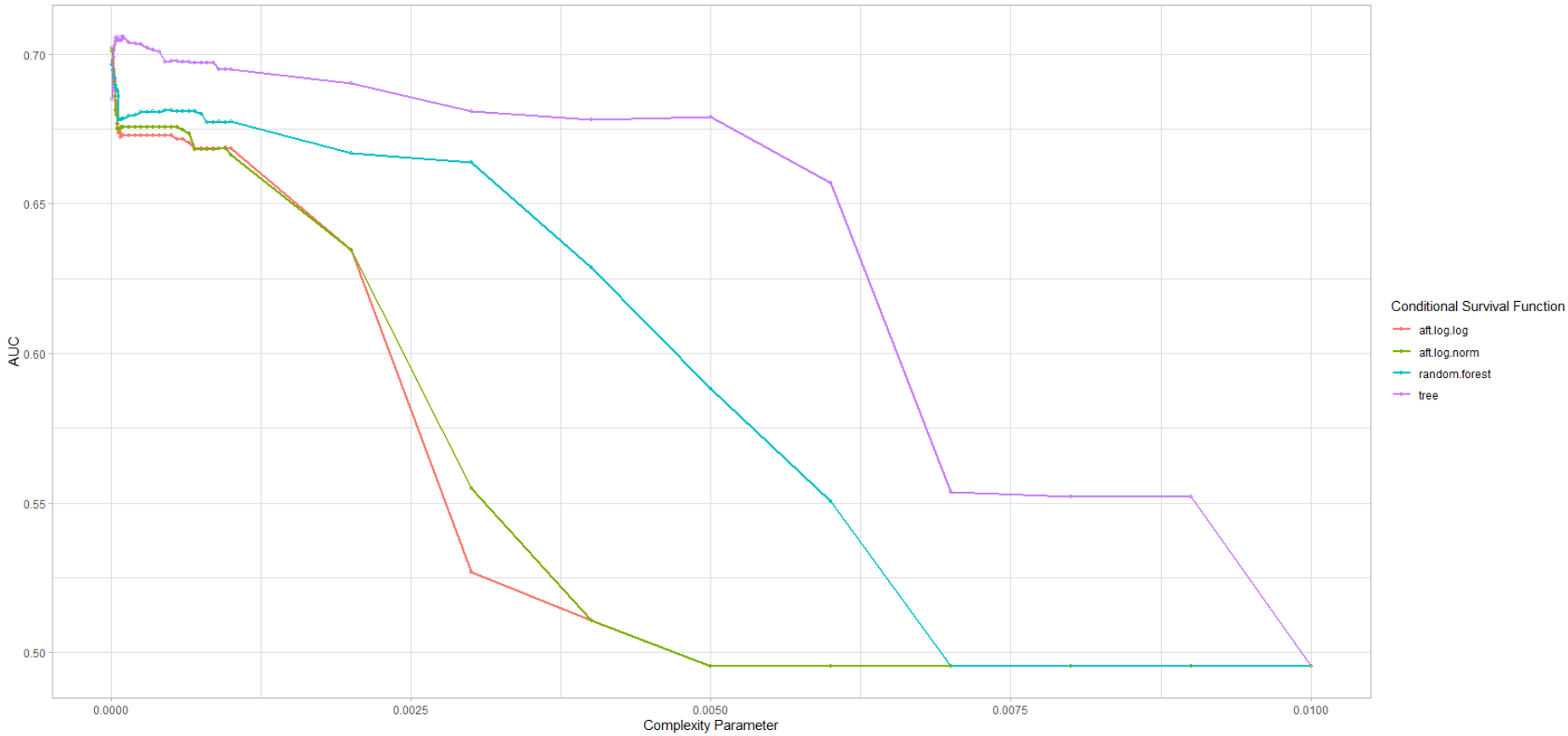

Figure 18. CURT (V2) - RQ1 GCUH Model Selection

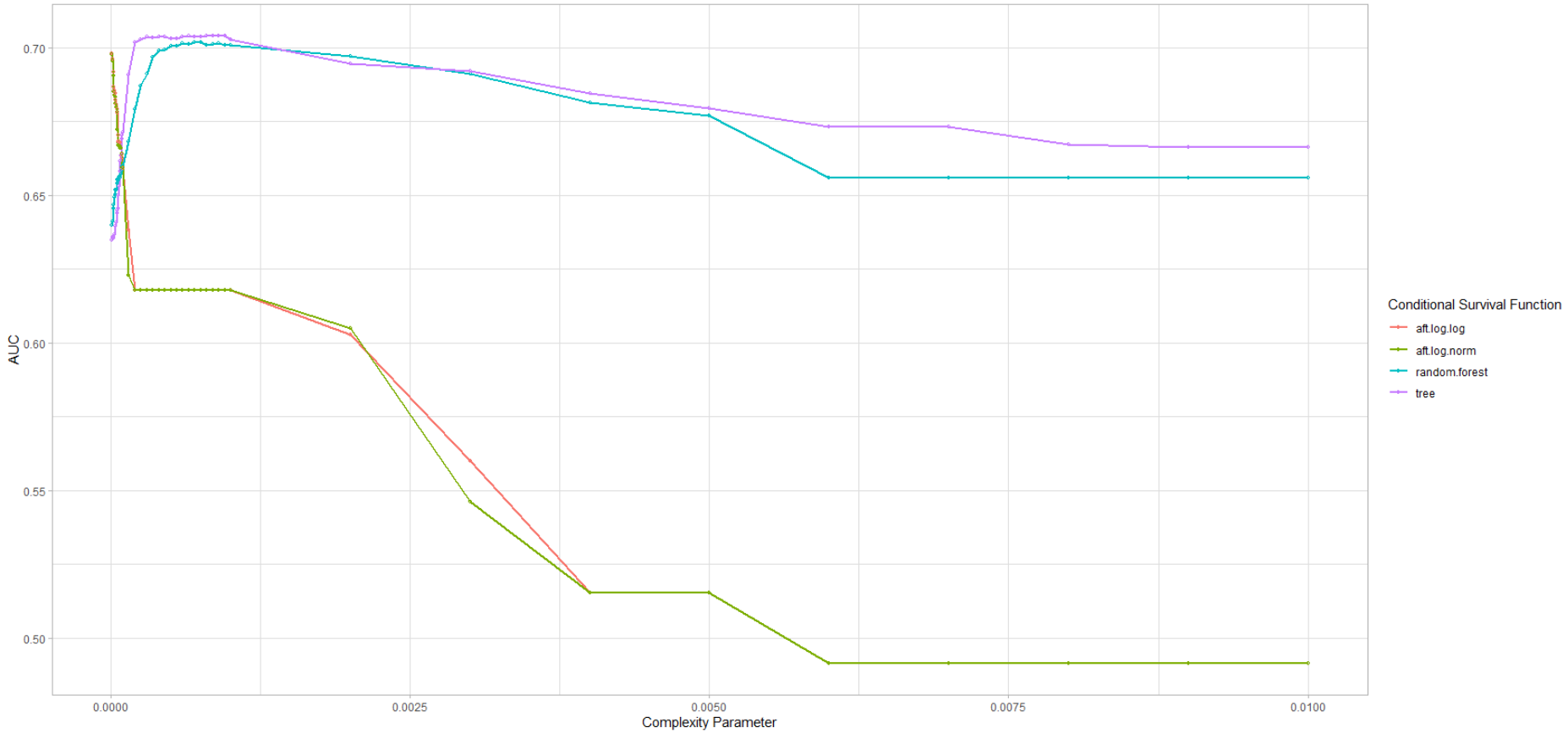

Figure 19. CURT (V2) - RQ1 RH Model Selection

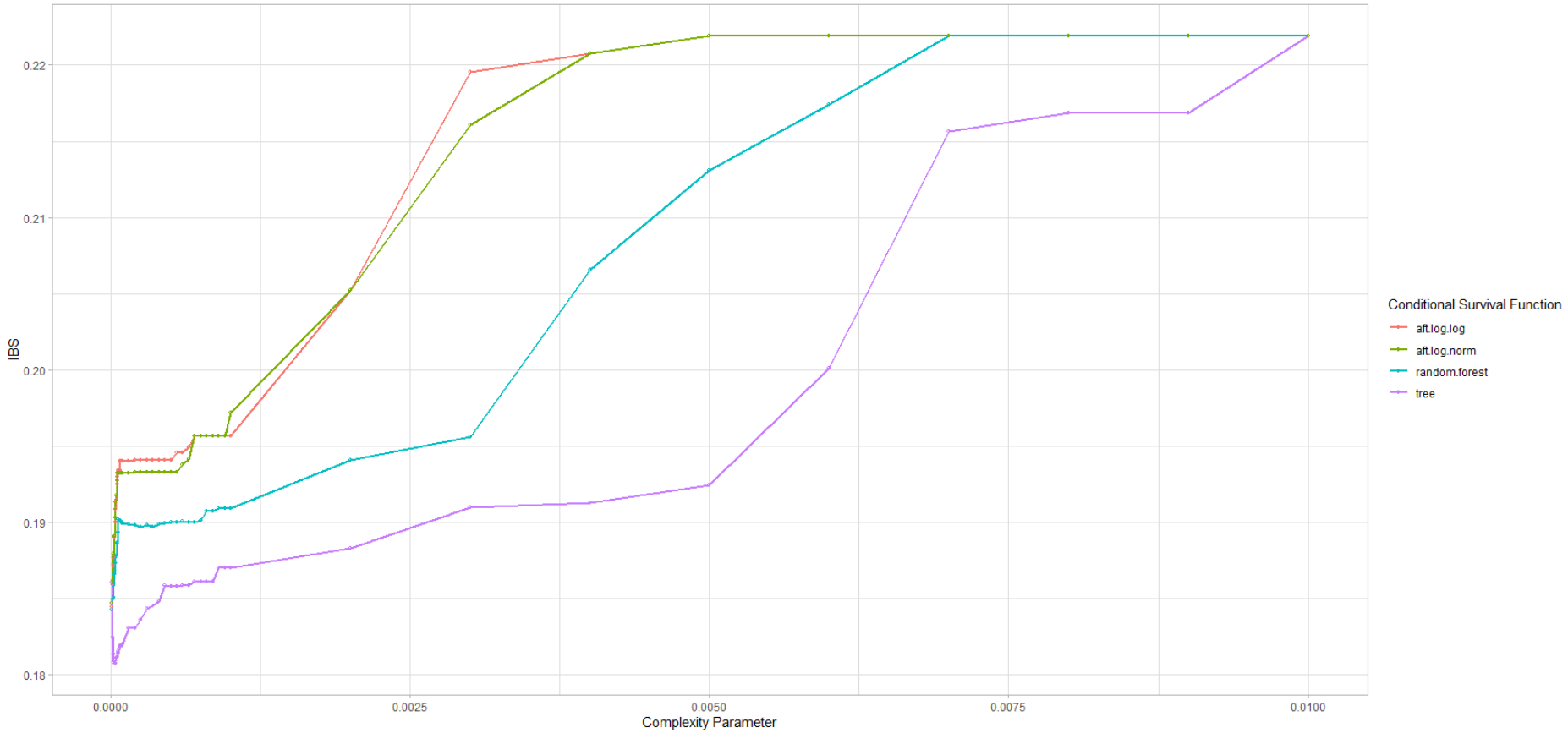

Figure 20. CURT (V2) - RQ2 GCUH Model Selection

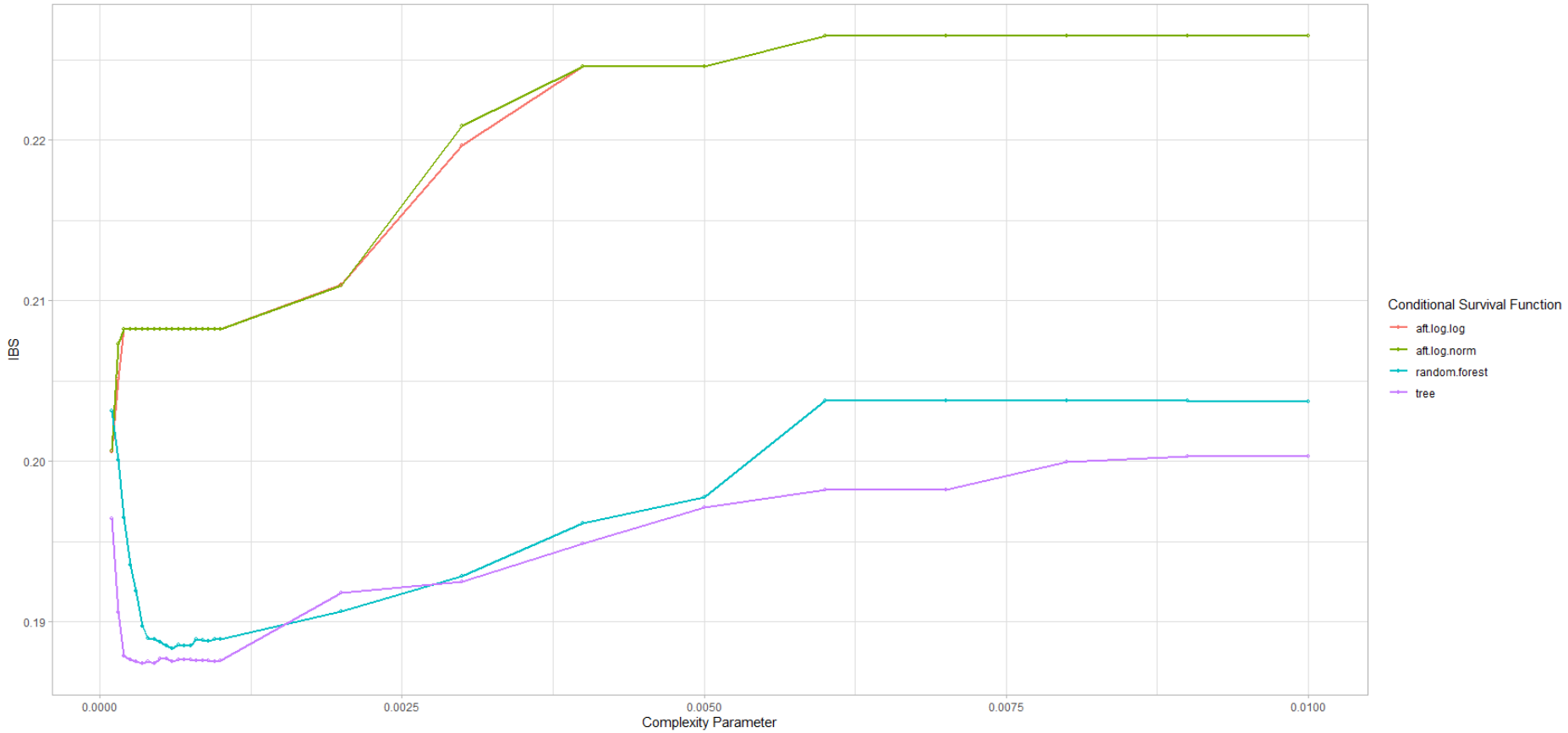

Figure 21. CURT (V2) - RQ2 RH Model Selection

Appendix C-5 Random Survival Forest

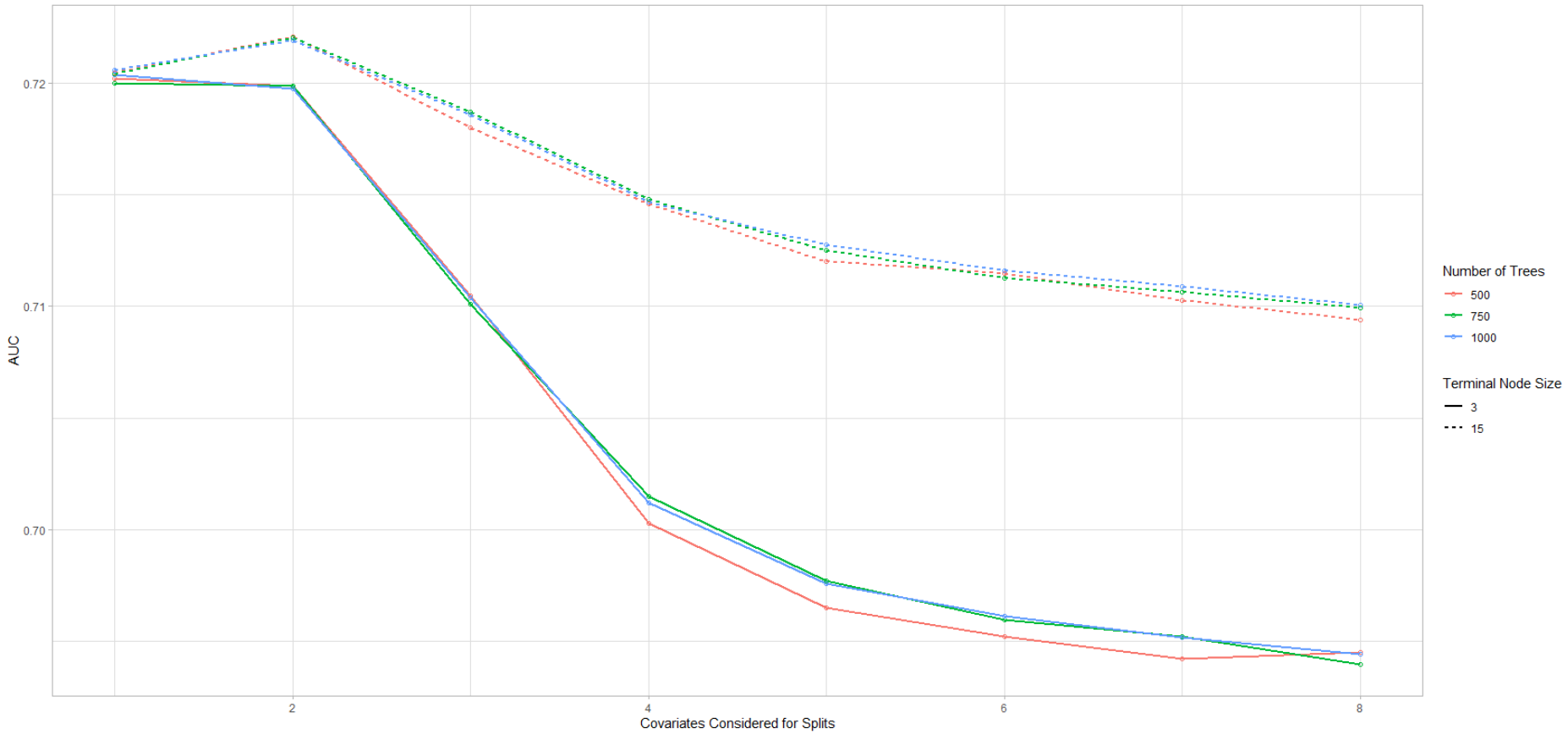

Figure 22. RSF - RQ1 GCUH Model Selection

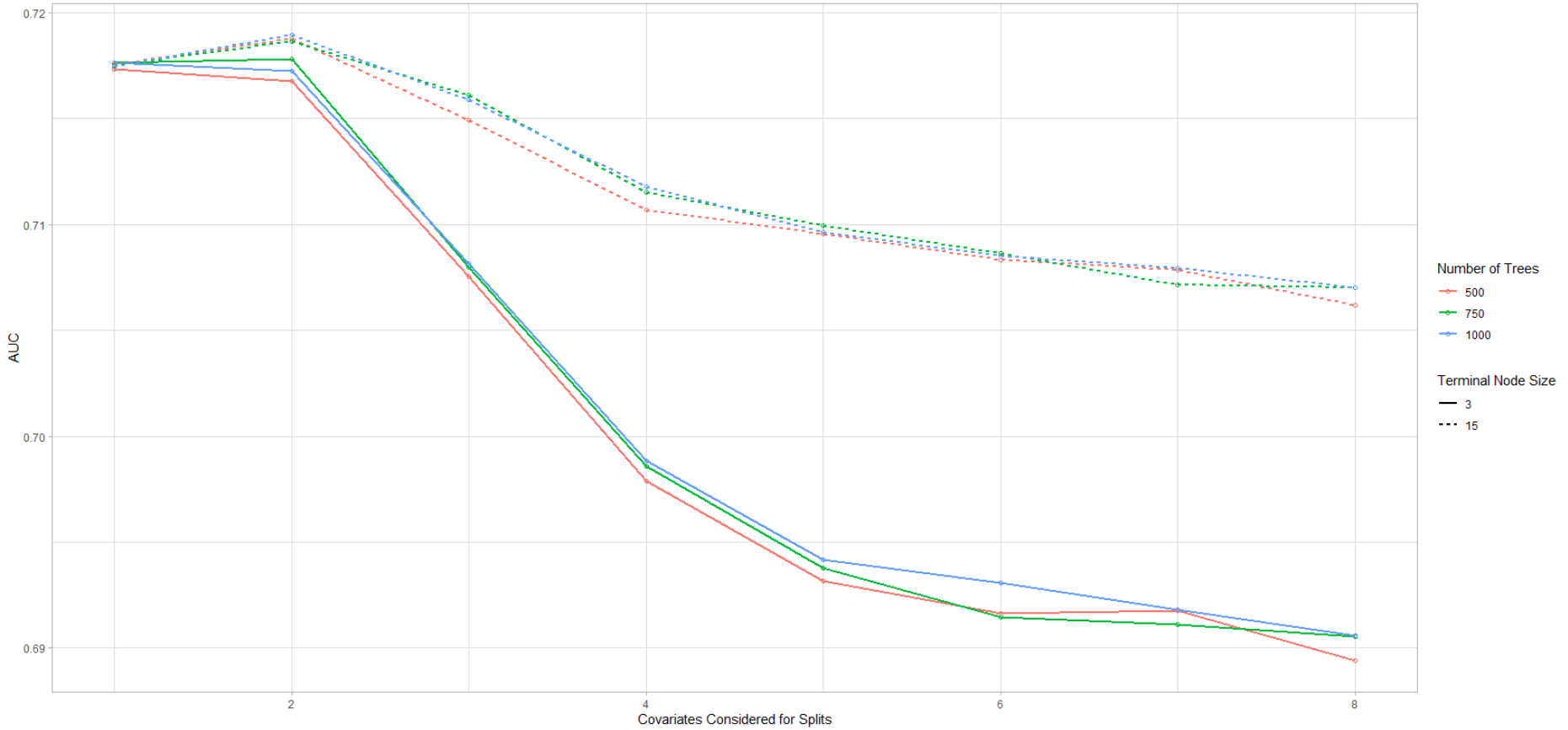

Figure 23. RSF - RQ1 RH Model Selection

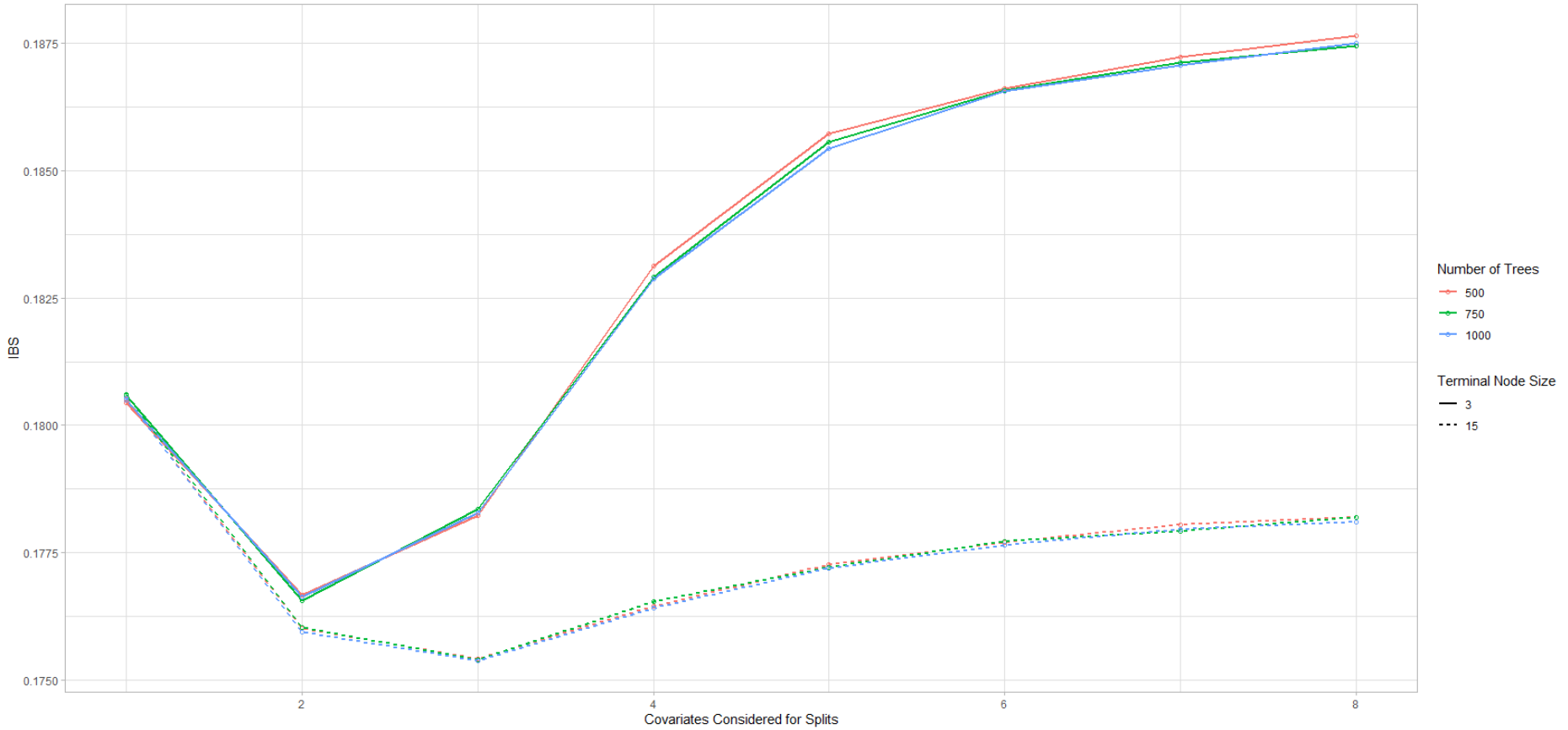

Figure 24. RSF - RQ2 GCUH Model Selection

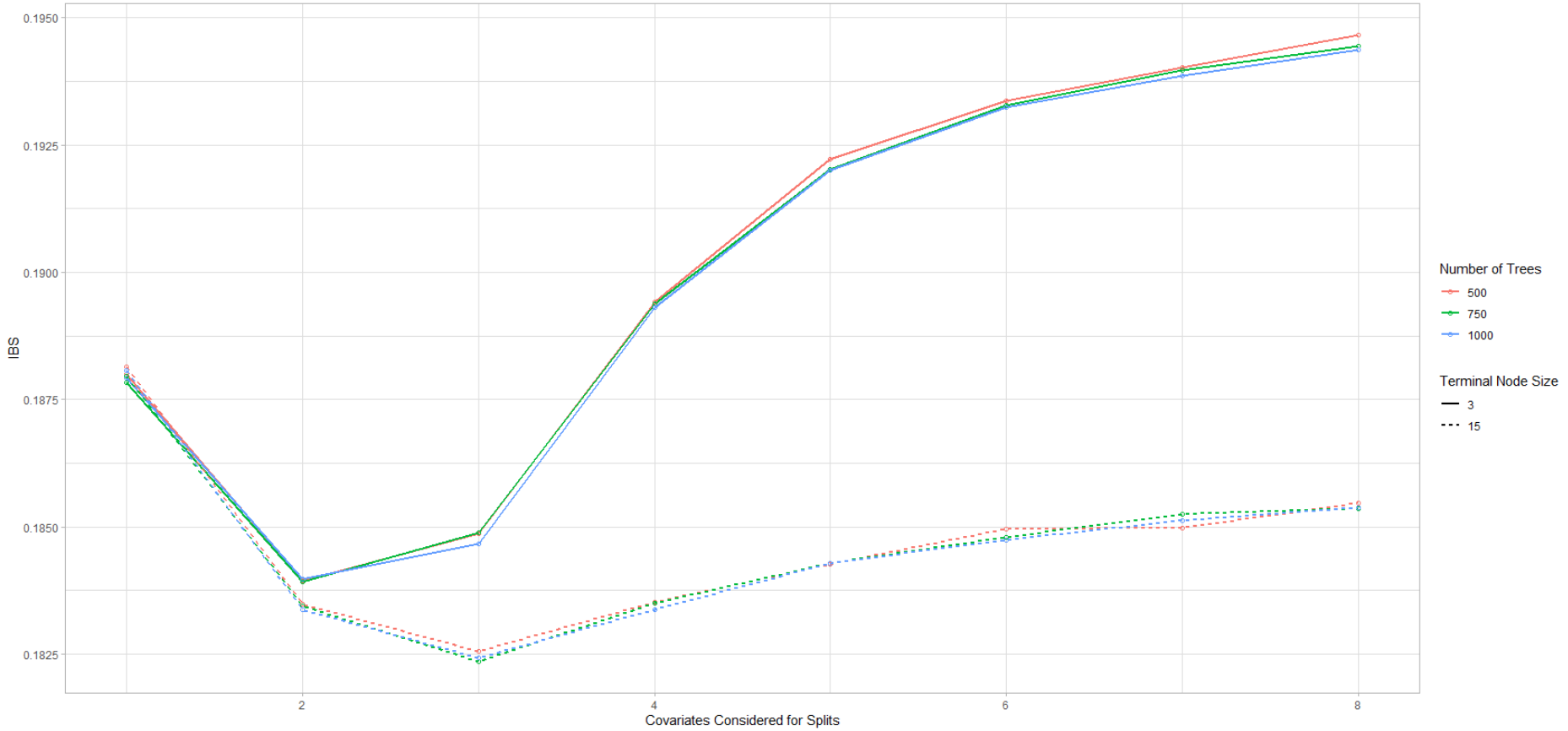

Figure 25. RSF - RQ2 RH Model Selection

Appendix C-6 CURE

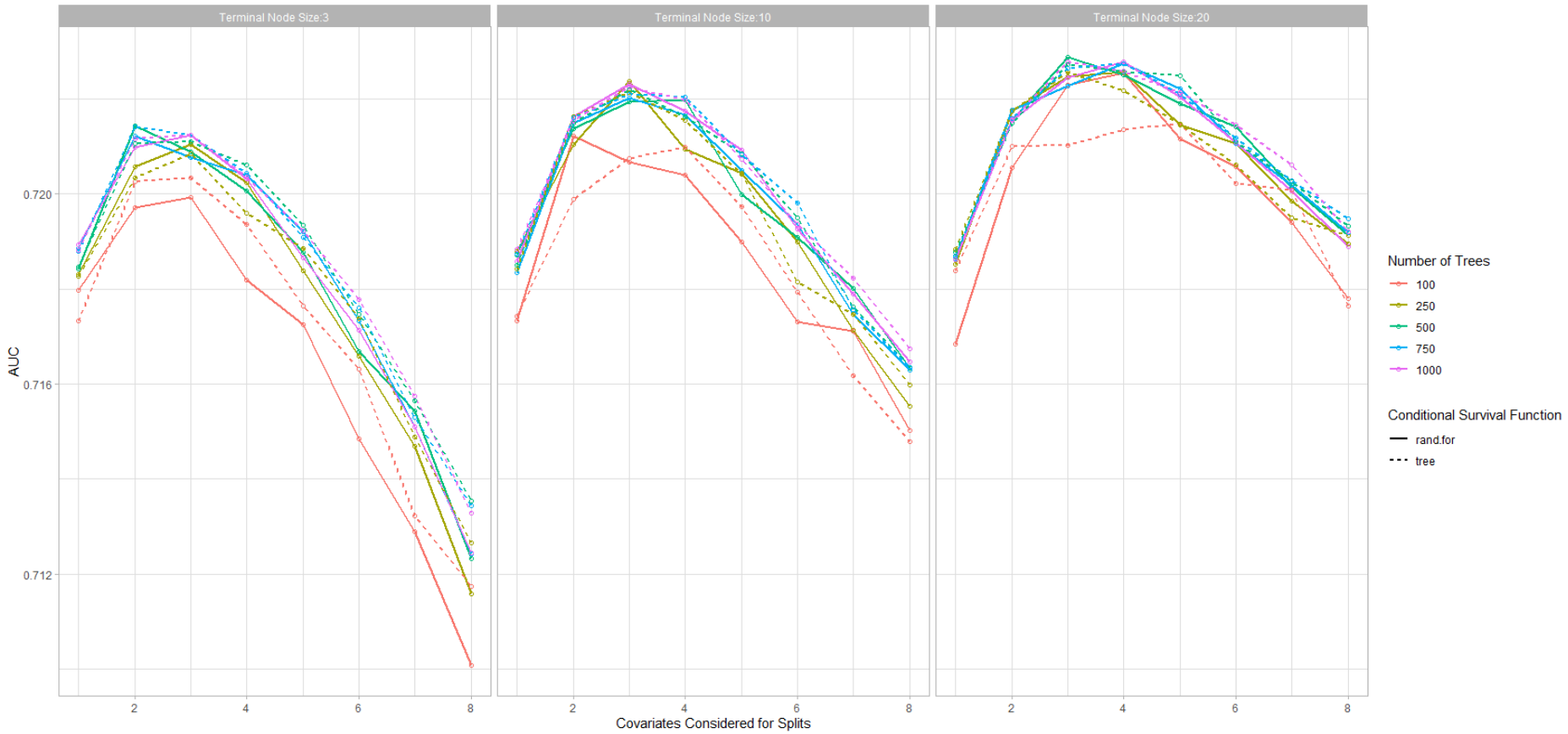

Figure 26. CURE - RQ1 GCUH Model Selection

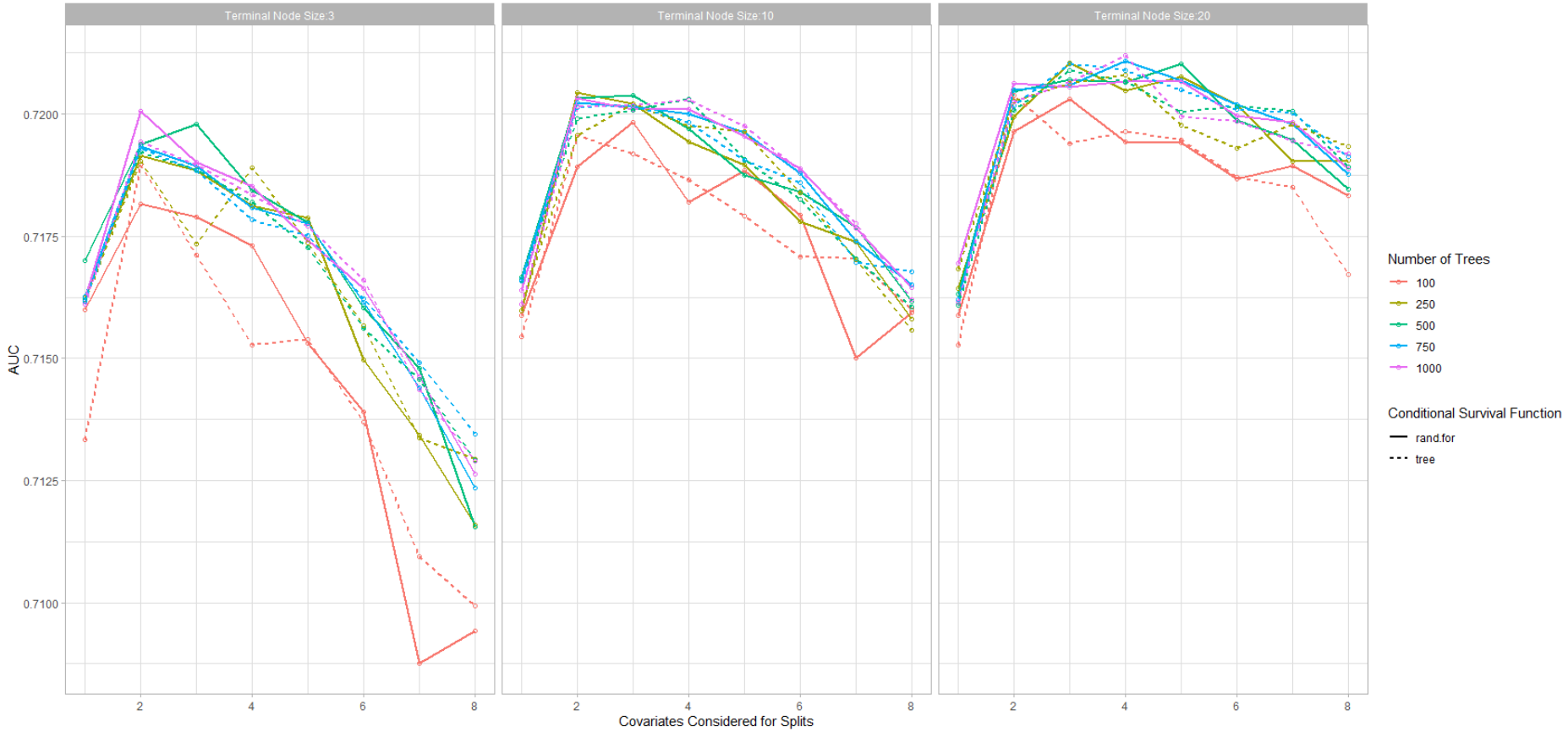

Figure 27. CURE - RQ1 RH Model Selection

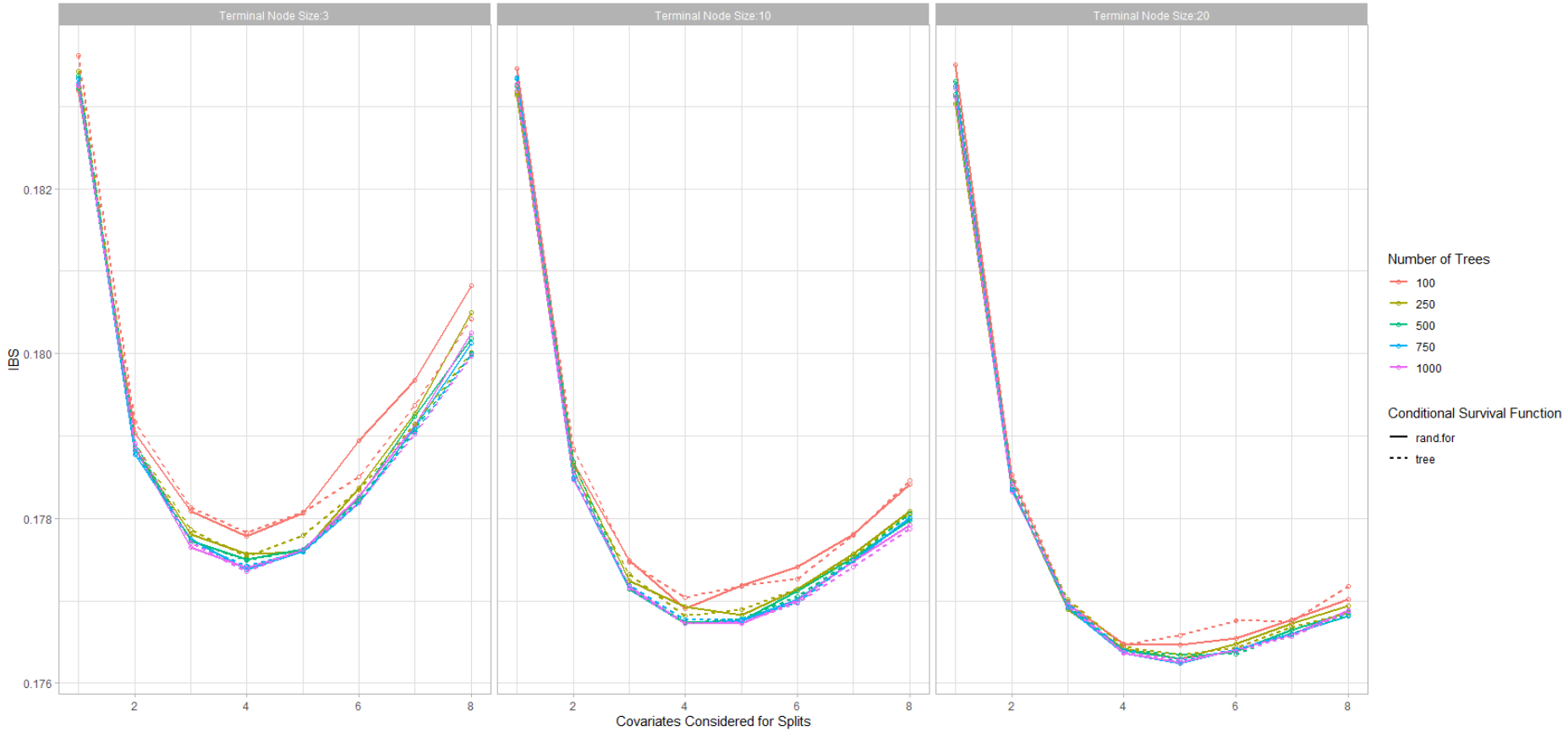

Figure 28. CURE - RQ2 GCUH Model Selection

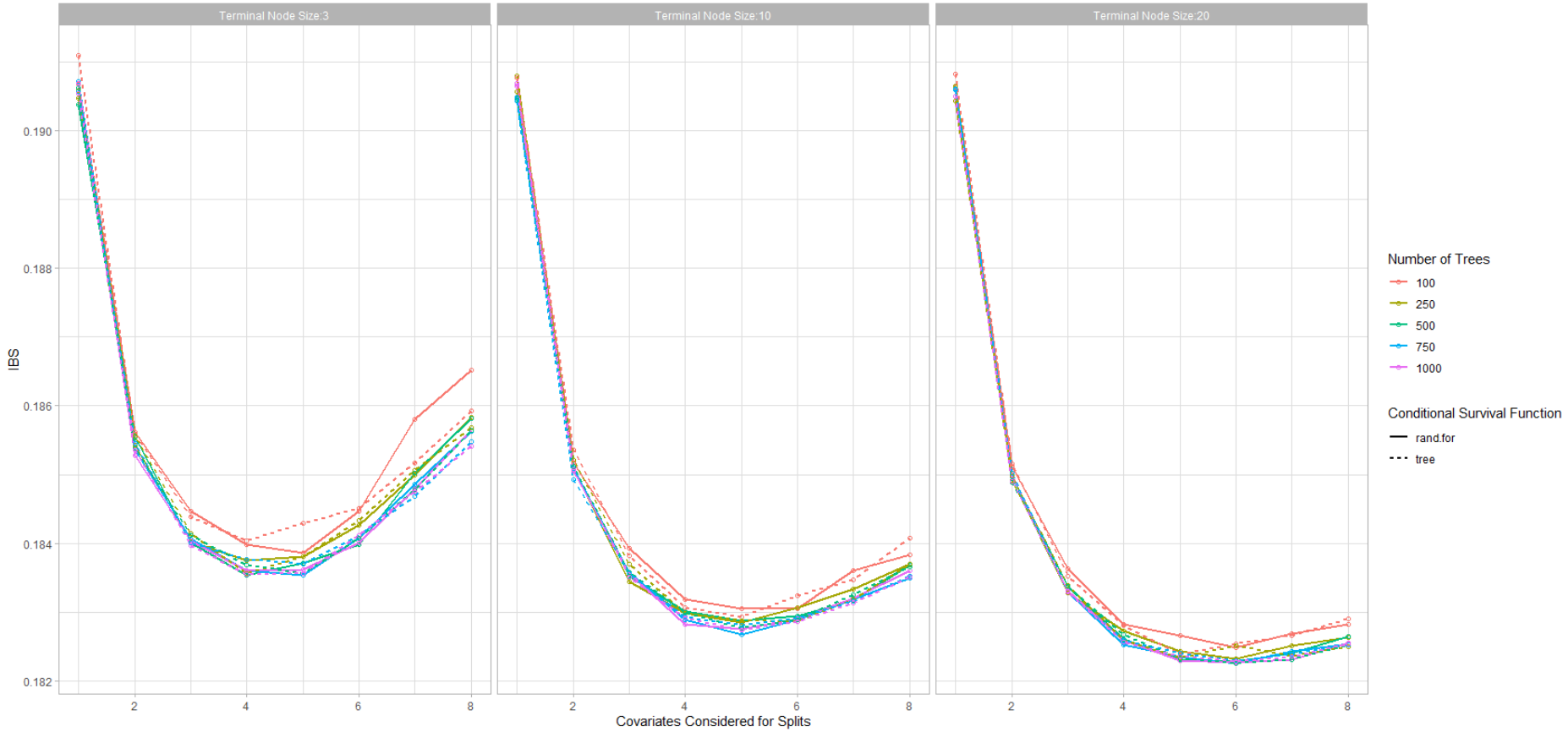

Figure 29. CURE - RQ2 RH Model Selection

Appendix C-7 RIST

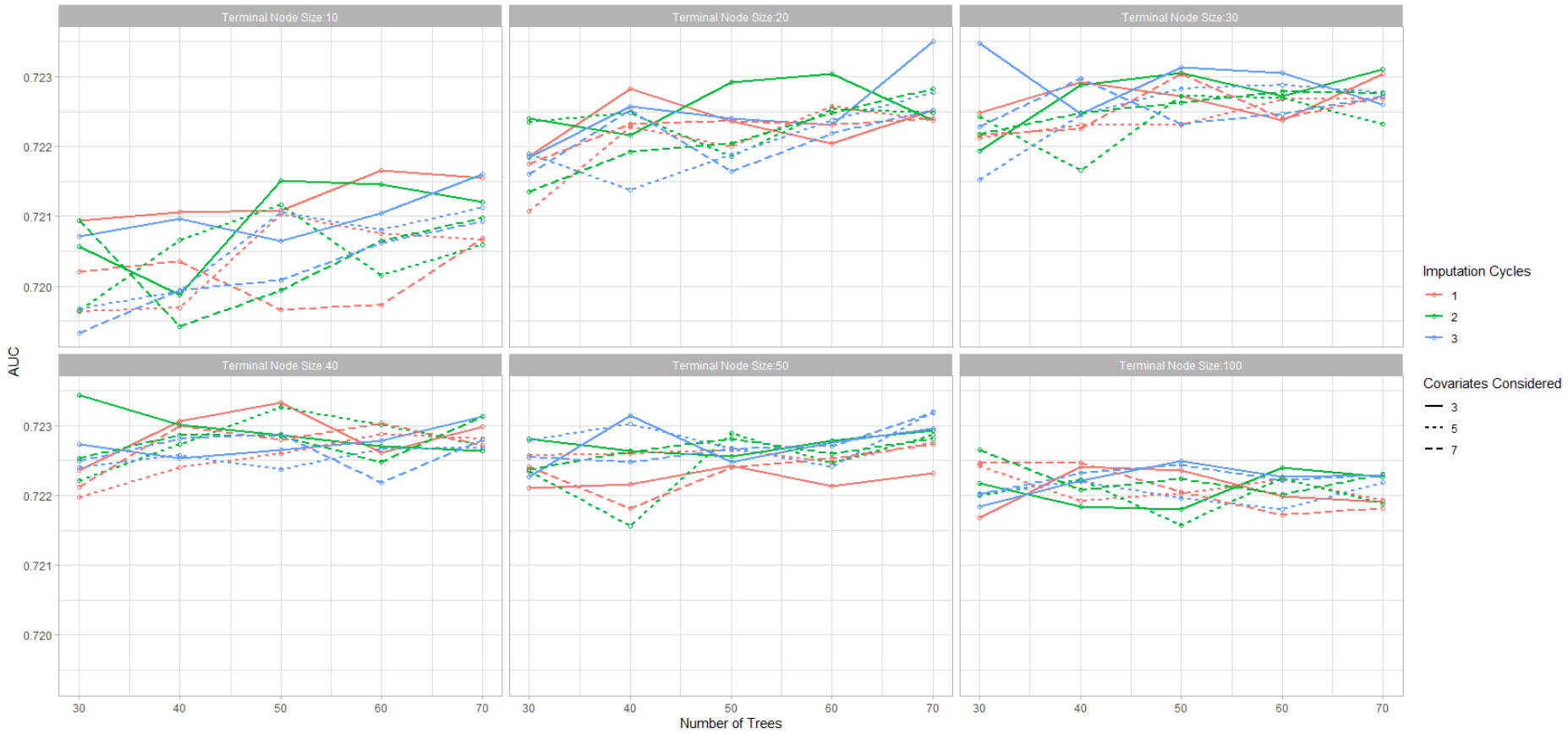

Figure 30. RIST - RQ1 GCUH Model Selection

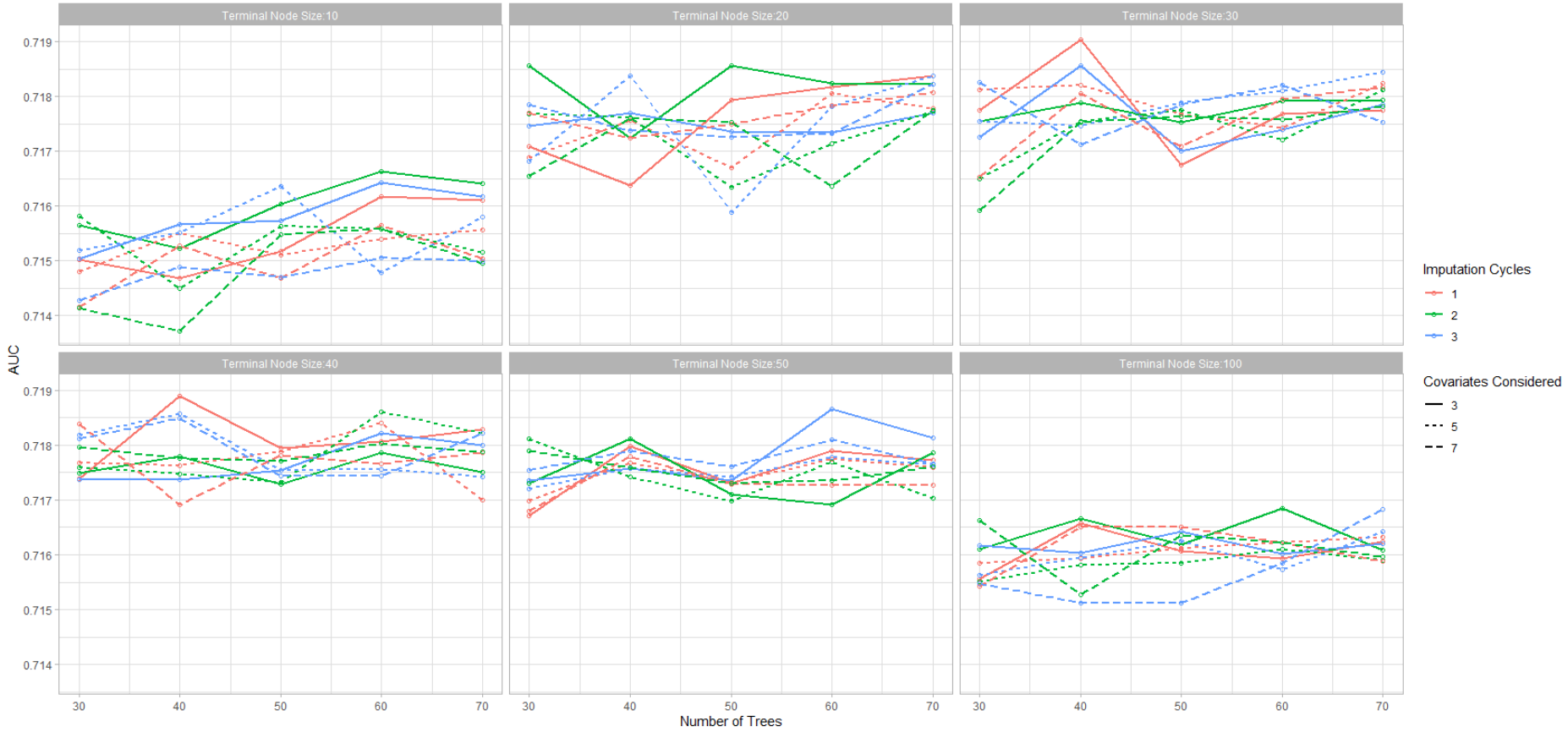

Figure 31. RIST - RQ1 RH Model Selection

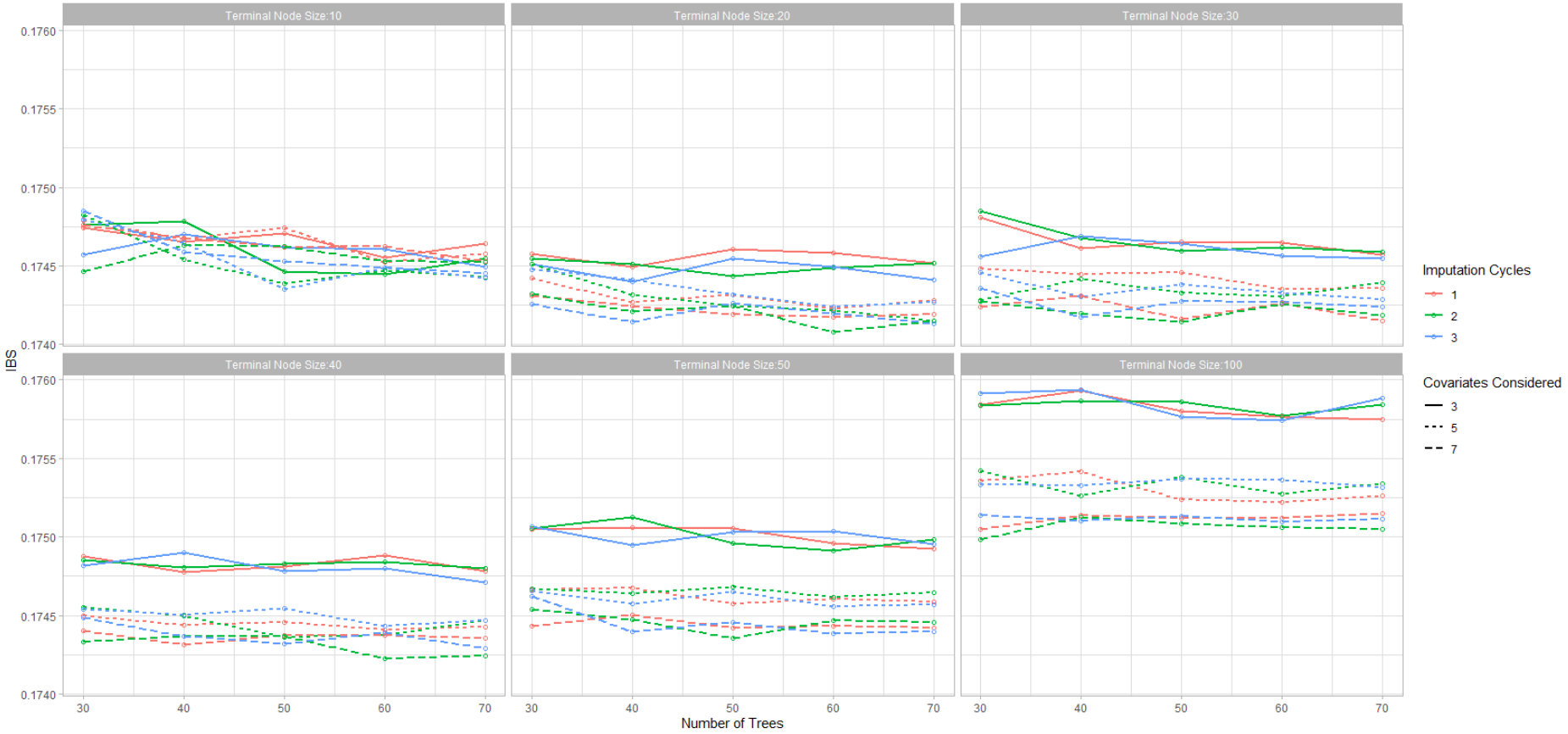

Figure 32. RIST - RQ2 GCUH Model Selection

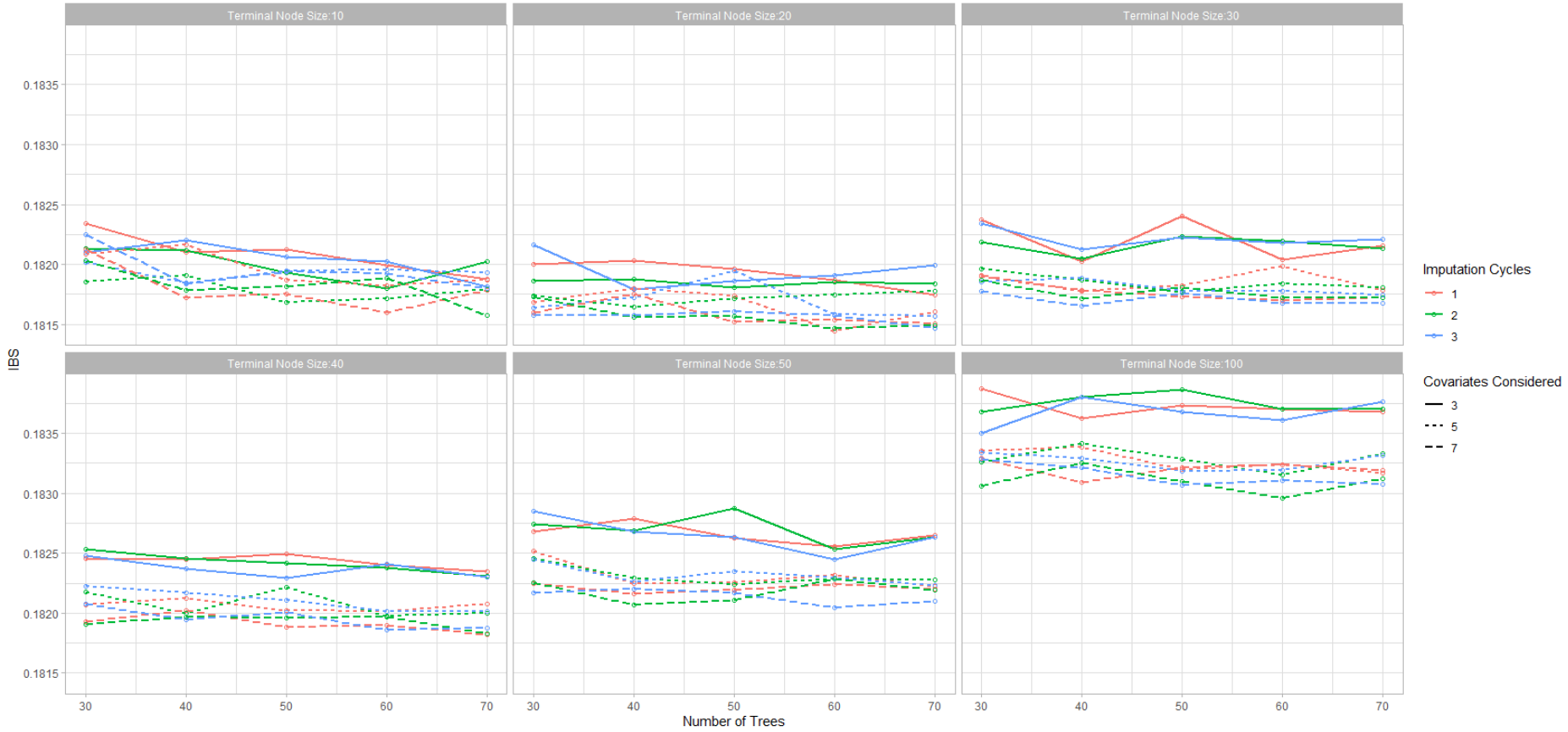

Figure 33. RIST - RQ2 RH Model Selection

Appendix C-8 NNET Survival

Figure 34. NNET Survival - RQ1 GCUH Model Selection

Figure 35. NNET Survival - RQ1 RH Model Selection

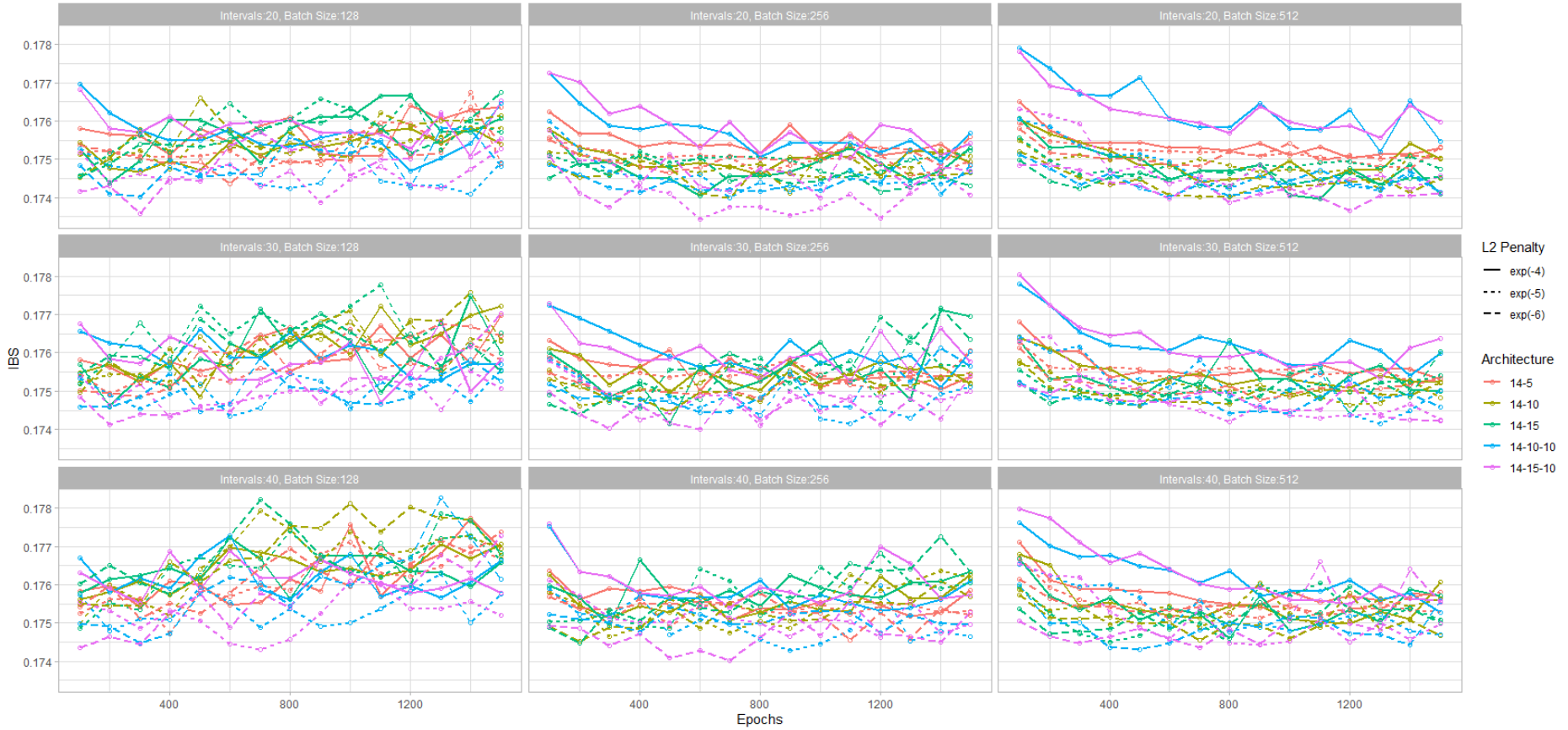

Figure 36. NNET Survival - RQ2 GCUH Model Selection

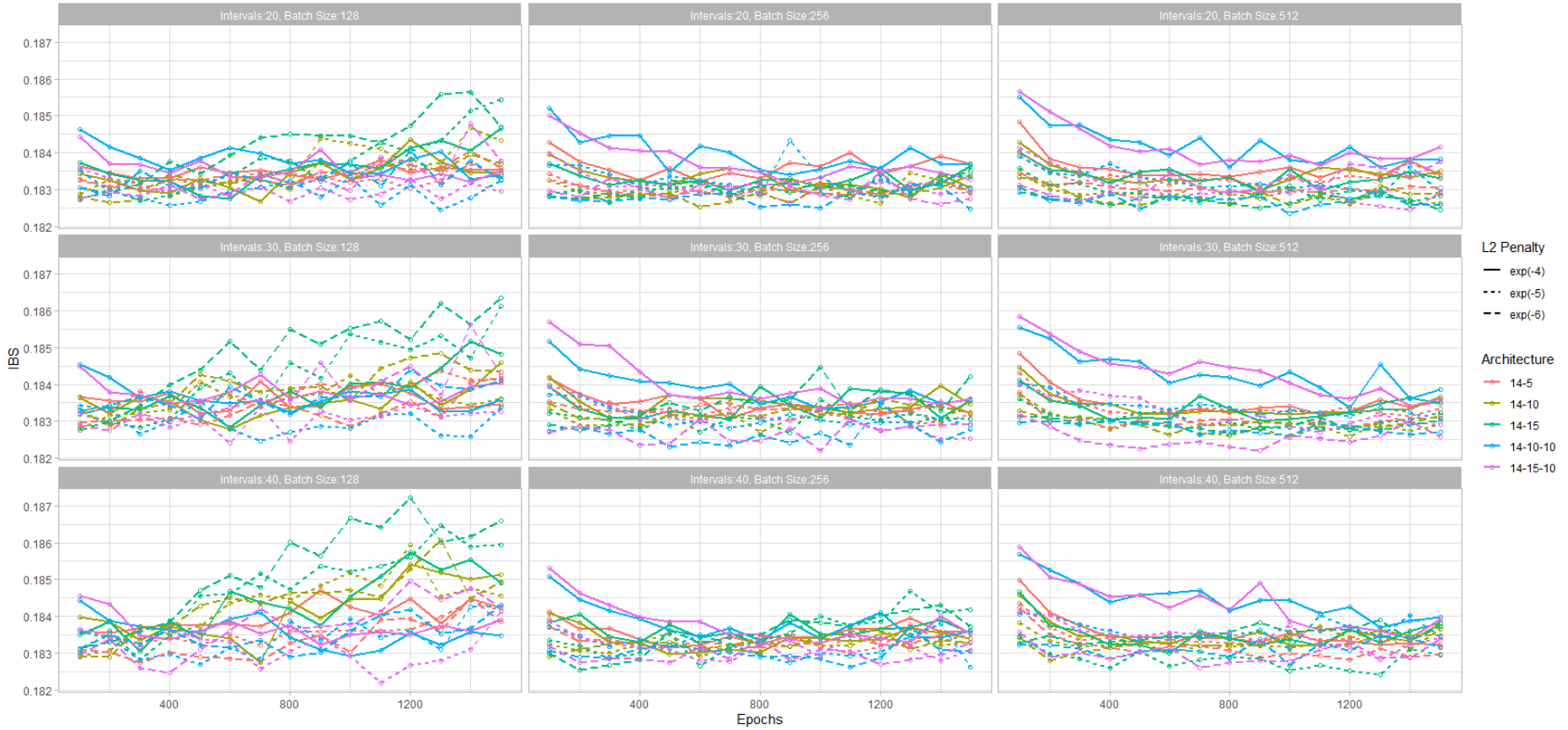

Figure 37. NNET Survival - RQ2 RH Model Selection

Appendix C-9 Time-Coded ANN

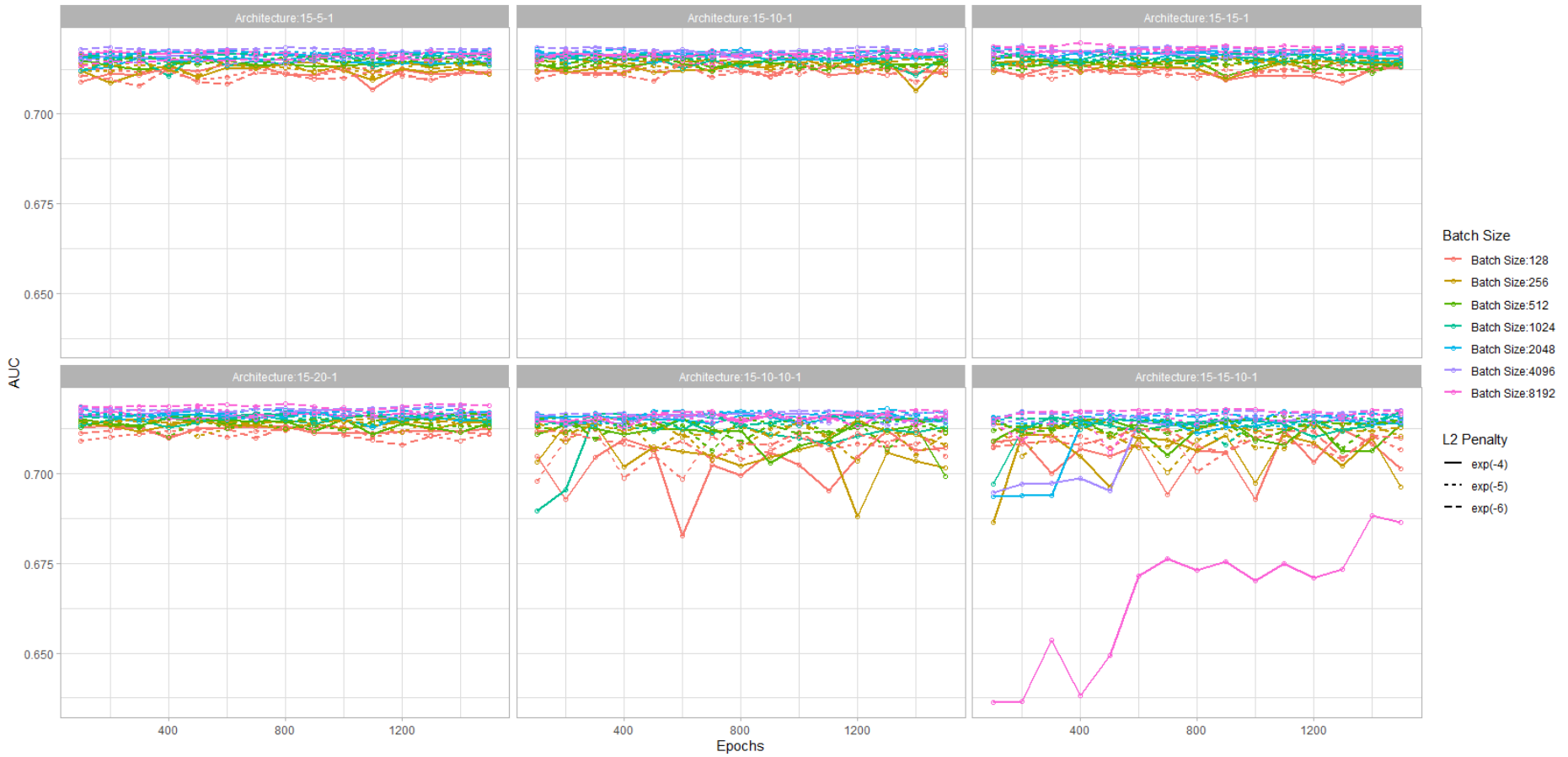

Figure 38. Time-Coded ANN - RQ1 GCUH Model Selection

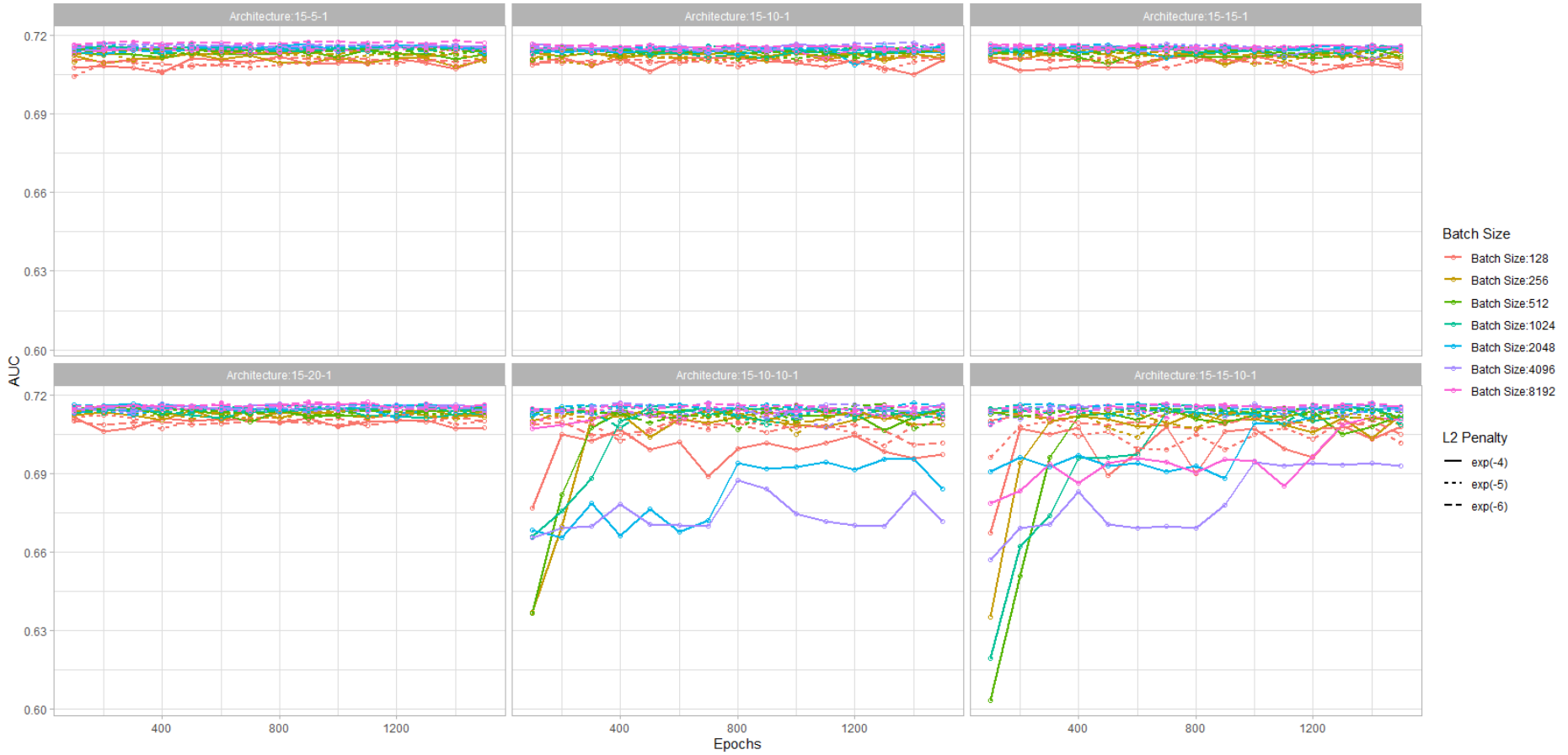

Figure 39. Time-Coded ANN - RQ1 RH Model Selection

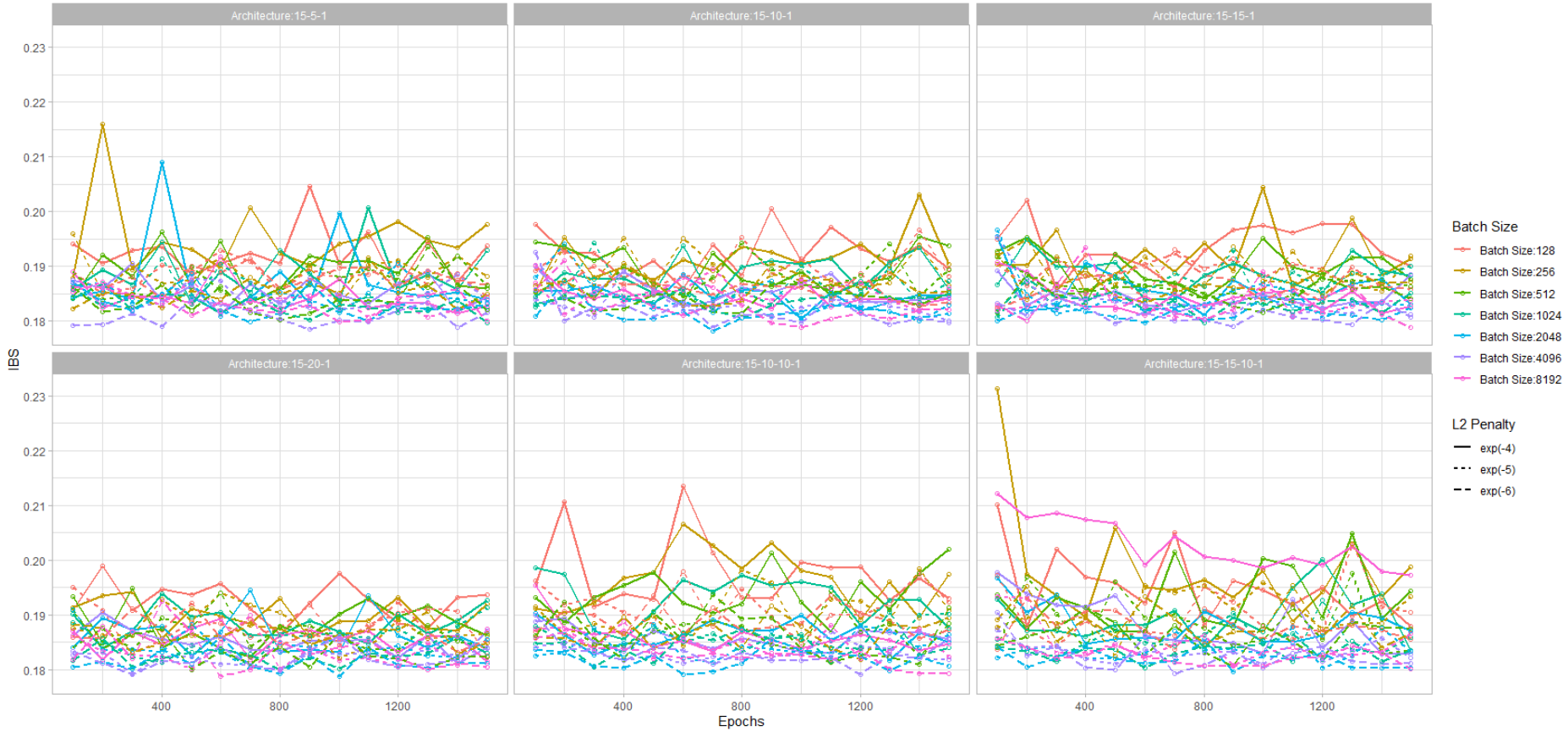

Figure 40. Time-Coded ANN - RQ2 GCUH Model Selection

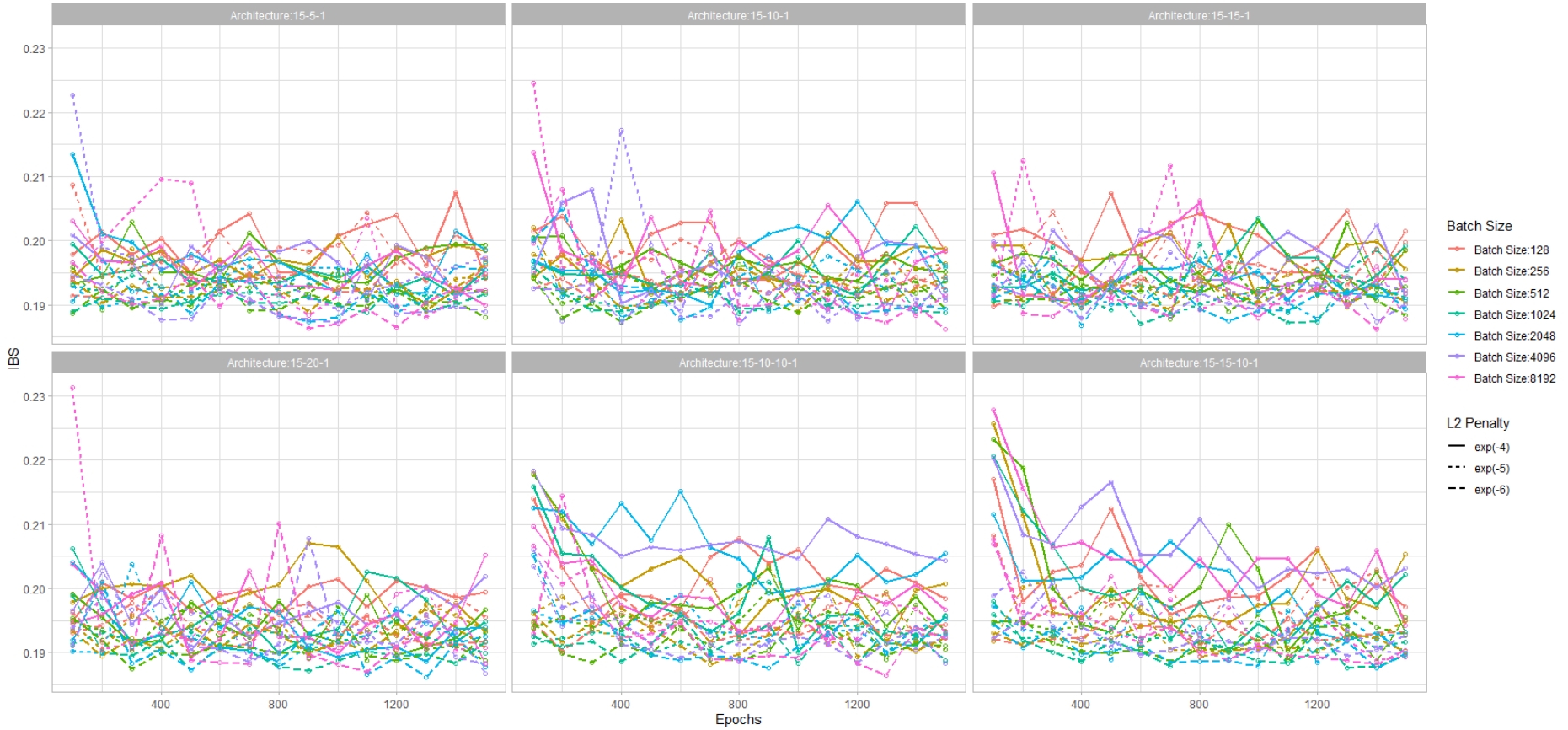

Figure 41. Time-Coded ANN - RQ2 RH Model Selection

Appendix C-10 Hybrid Cox-ANN

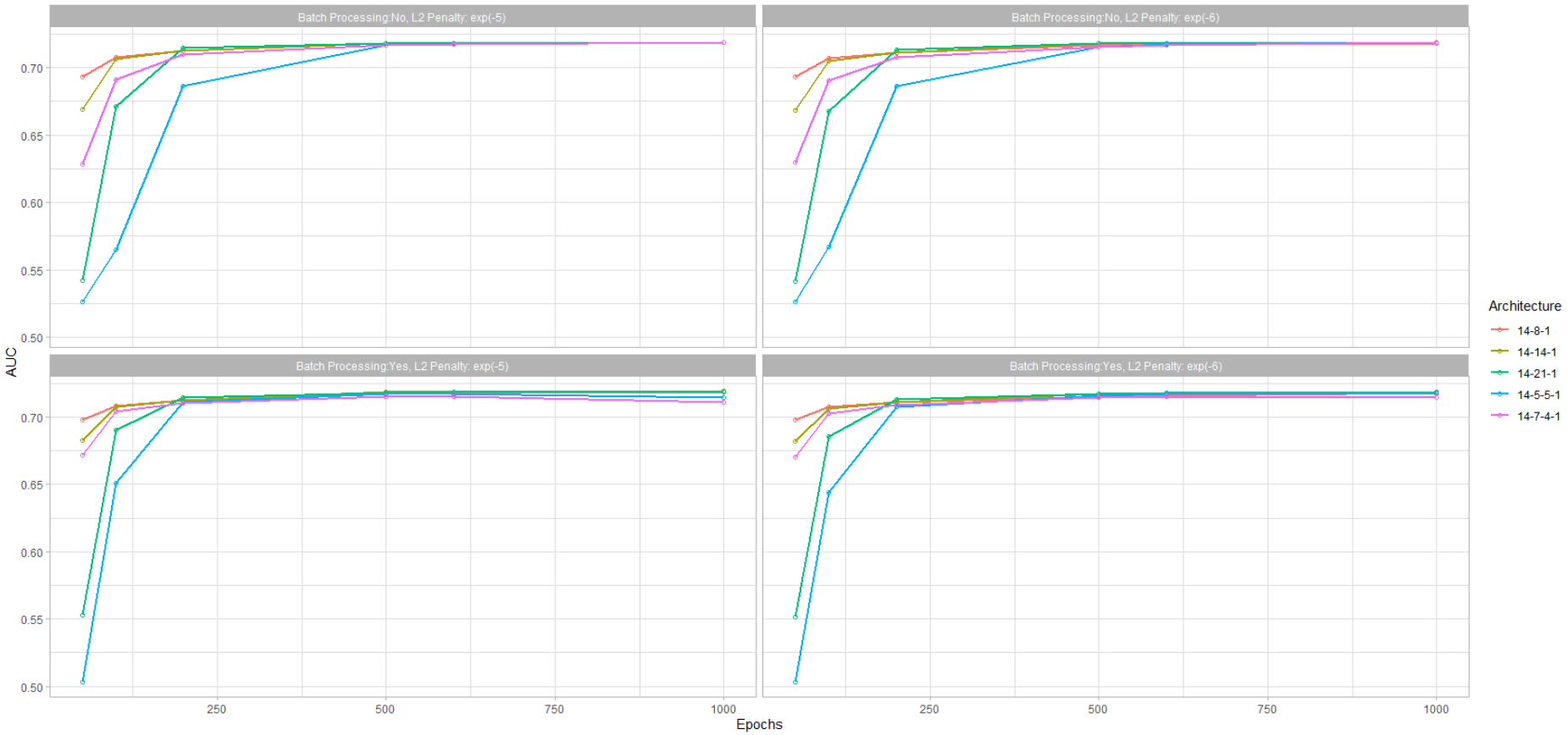

Figure 42. Hybrid Cox-ANN - RQ1 GCUH Model Selection

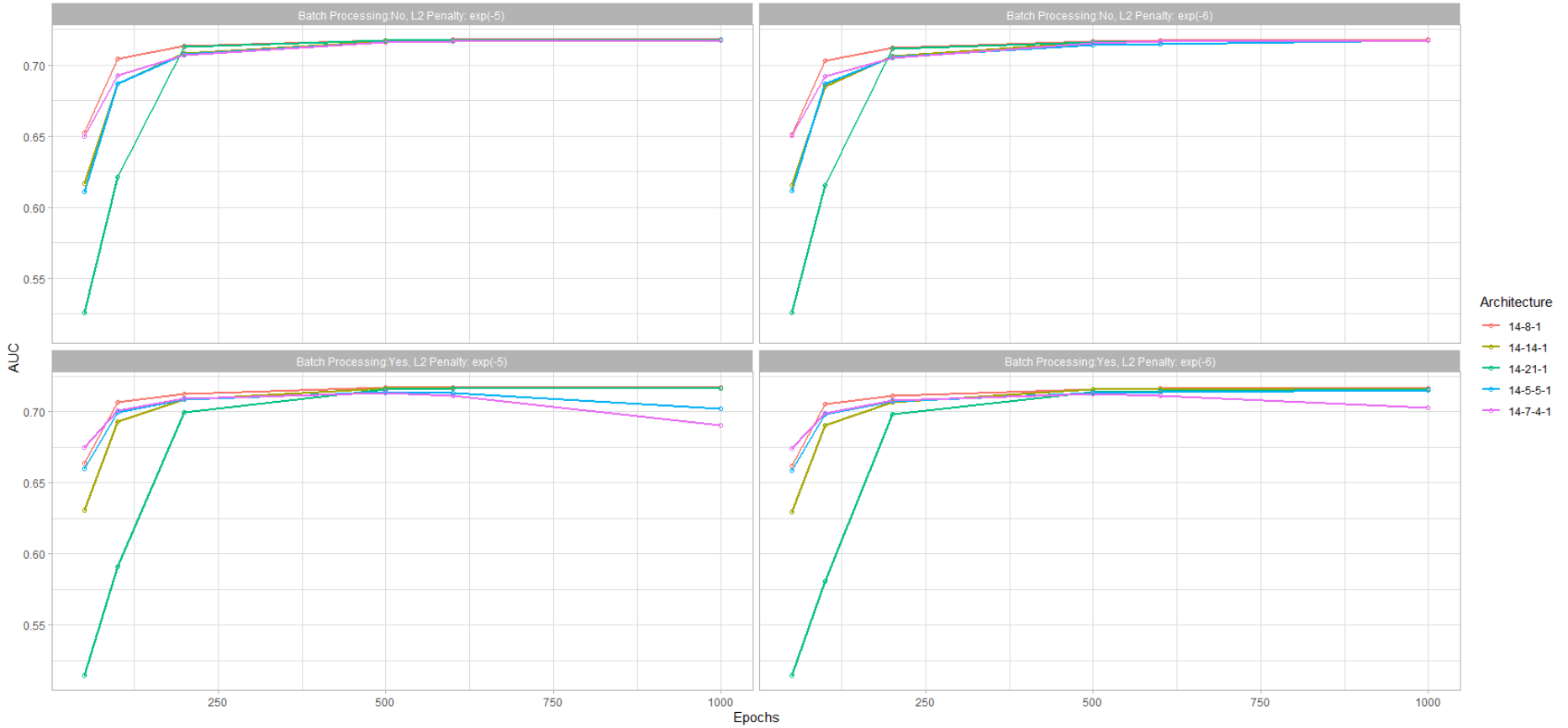

Figure 43. Hybrid Cox-ANN - RQ1 RH Model Selection

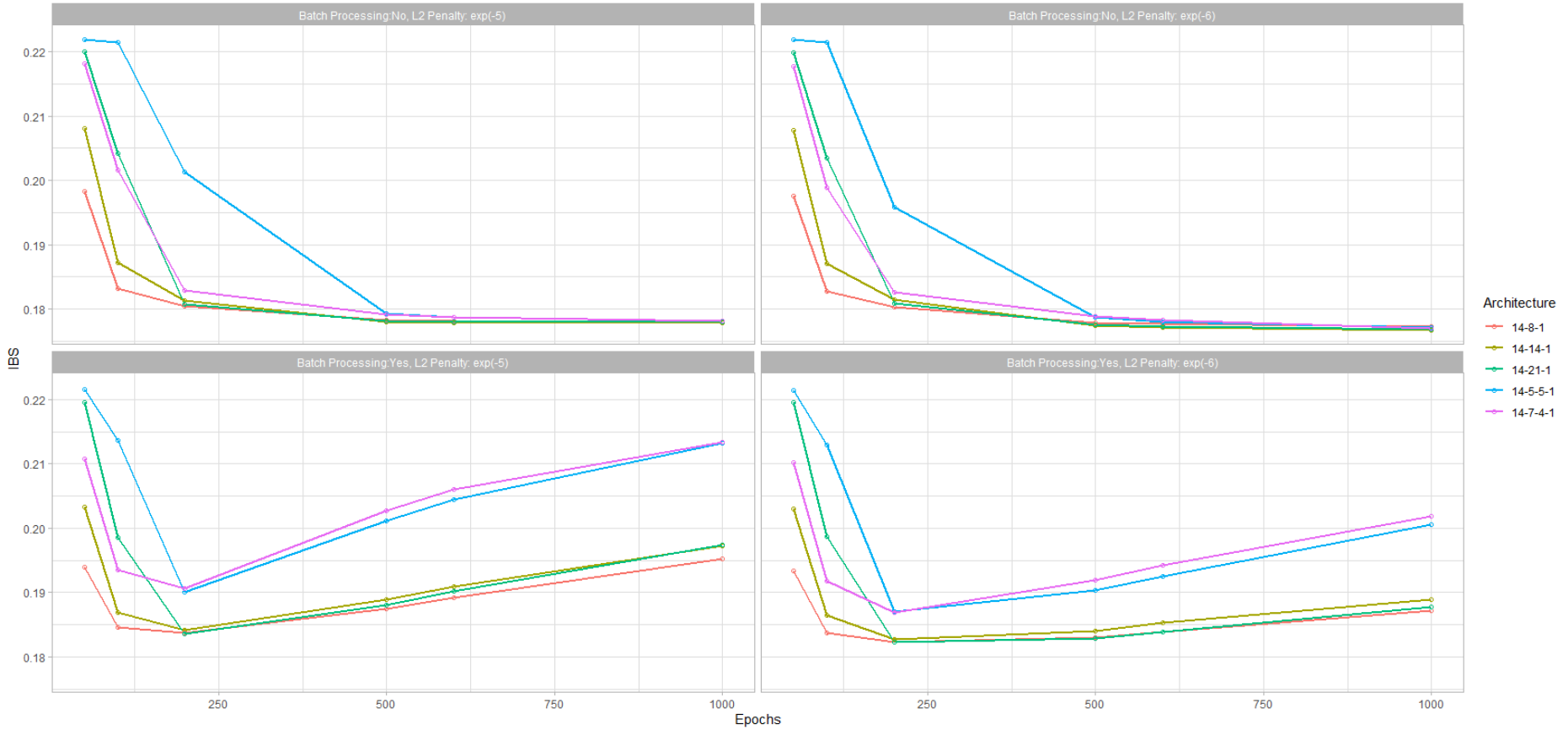

Figure 44. Hybrid Cox-ANN - RQ2 GCUH Model Selection

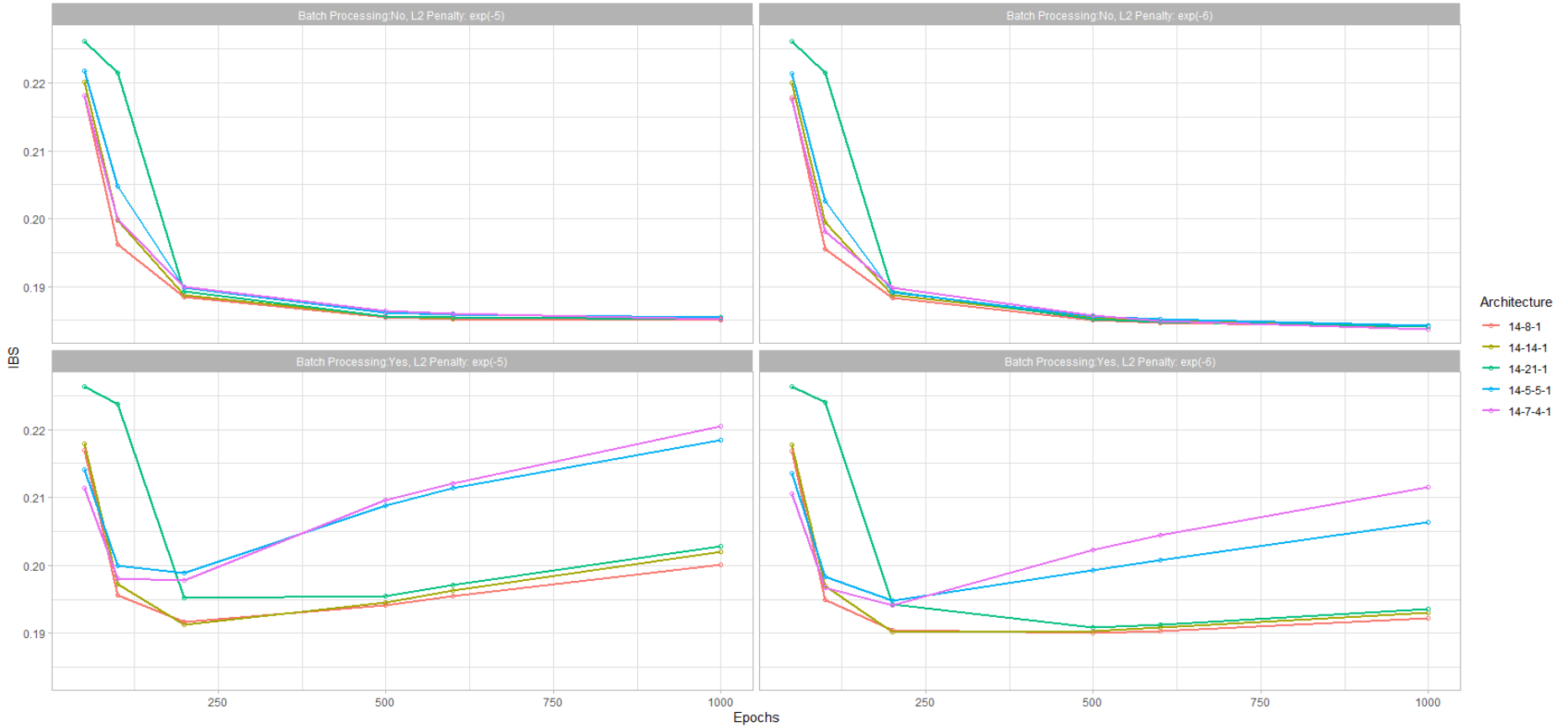

Figure 45. Hybrid Cox-ANN - RQ2 RH Model Selection