UNIVERSITY OF CAPE TOWN

MASTERS THESIS

Development of a risk score for constrictive pericarditis using the Investigation of the Management of Pericarditis randomised clinical trial dataset

Author:

Hayli GEFFEN

Supervisors:

Assoc. Professor Freedom GUMEDZE

Professor Mpiko NTSEKHE

A mini-dissertation submitted in fulfilment of the requirements for the degree of

MASTERS OF SCIENCE,

specialising in Biostatistics in the

DEPARTMENT OF STATISTICAL SCIENCES

September 4, 2022

Abstract

Despite the recent global decline of tuberculosis infections, constrictive pericarditis, one of the most serious consequences of tuberculous pericarditis, continues to be a major cause of morbidity and mortality in sub-Saharan Africa. Currently, while the risk of constrictive pericarditis in individuals with tuberculous pericarditis does not appear to be uniform, there is no defined risk score available to predict an individual's baseline risk of constrictive pericarditis. Therefore the main aim of this research was to employ supervised learning classification using the data from 1400 participants enrolled in the first Investigation of the Management of Pericarditis randomised clinical trial to derive a risk score for constrictive pericarditis. While various supervised learning classification methods, including tree-based algorithms, support vector machines and artificial neural networks, were compared to stratify individuals according to low, medium and high risk for constrictive pericarditis, the final risk score was developed using logistic regression. Significant associations were found between constrictive pericarditis and the following predictors: HIV, New York Heart Association functional class, cardiac tamponade and effusive-constrictive pericarditis. Although prednisolone treatment was associated with a reduced relative risk of constrictive pericarditis in high (risk ratio = 0.59; 95% CI = 0.378 – 0.925) and medium (risk ratio = 0.12; 95% CI = 0.016 – 0.971) risk individuals, prednisolone treatment did not seem to benefit the individuals predicted to be at low risk (risk ratio = 0.92; 95% CI = 0.084 - 10.047) for constrictive pericarditis. These results confirm that the baseline risk of developing constrictive pericarditis in individuals with suspected or confirmed tuberculous pericarditis is not uniform. Importantly, interventions such as adjunctive prednisolone should only be recommended for individuals suspected to be at either medium or high risk for constrictive pericarditis as they are the most likely to benefit while prednisolone treatment should potentially be avoided in treating individuals with tuberculous pericarditis that are suspected to be at low risk for constrictive pericarditis as they are the least likely

to derive any benefit.

Acknowledgements

I would like to acknowledge my supervisor Assoc. Prof. Freedom Gumedze for the expert statistical guidance and constructive feedback throughout this process. Thank you for providing me with insightful direction while still allowing me to tackle this project independently. I would like to acknowledge my supervisor Prof. Mpiko Ntsekhe for the assistance with the clinical interpretations of the results. Thank you for guiding me to ensure that the analyses were clinically meaningful and the opportunity to work with this dataset. I am extremely grateful and proud to be a part of the IMPI team.

I would like to thank Dr Juwa Nyirenda, Dr Sebnem Er and Stefan Britz from the Department of Statistical Sciences for the exceptional Supervised Learning course from which the concepts I learned inspired me to take on this research project.

I would like to thank the ETDP SETA for funding my degree and supporting this research.

Finally, thank you to my family for their unconditional love and support.

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike information criterion
AIDS	Acquired immunodeficiency syndrome
AUC	Area under the curve
AUCPR	Area under the precision-recall curve
CART	Classification and Regression Trees
CI	Confidence interval
CP	Constrictive pericarditis
FN	False negative
FP	False positive
HIV	Human immunodeficiency virus
HR	Hazard ratio
IMPI	Investigation of the Management of Pericarditis
IQR	Interquartile range
MAR	Missing at random
MCAR	Missing completely at random
MIP	*Mycobacterium indicus pranii*

MNAR	Missing not at random
NPV	Negative predictive value
NRMSE	Normalised root mean squared error
NYHA	New York Heart Association
OR	Odds ratio
PH	Proportional hazards
PPV	Positive predictive value
RCT	Randomised clinical trial
ROC	Receiver operating characteristic
Sens	Sensitivity
SMOTE	Synthetic minority oversampling technique
SMOTE-NC	SMOTE-nominal continuous
Spec	Specificity
SVM	Support vector machine
TB	Tuberculosis
TBP	Tuberculous pericarditis
TN(R)	True negative (rate)
TP(R)	True positive (rate)

Chapter 1

Introduction

1.1 Background

Despite the recent global decline of tuberculosis (TB) infections [1], TB continues to be the leading cause of death in TB-endemic regions such as South Africa [2]. Additionally in South Africa, tuberculous pericarditis (TBP), which is characterised by TB infection and inflammation of the pericardium, is one of the most serious morbidities associated with TB infection [3].

While TBP is treatable with anti-tuberculous drugs [4], the human immunodeficiency virus (HIV) epidemic in sub-Saharan Africa has resulted in the mortality rate for patients presenting with both clinical features of HIV infection and TBP being more than double compared to patients with TBP who are not immunocompromised [5]. This is especially relevant in South Africa, because there are over 7.5 million people presently living with HIV, and over 200,000 people are concurrently infected with both TB and HIV [6].

Broadly, TBP manifests in three forms: (1) pericardial effusion, (2) constrictive pericarditis (CP), or (3) a combination of both pericardial effusion and constriction [7, 8]. CP, which is characterised by pericardial inflammation or thickening [9], is a major cause of morbidity and mortality in sub-Saharan Africa and occurs in up to 60% of patients with TBP [7].

Compared to pericardiocentesis, a safe and effective procedure used for treating effusive

pericarditis [10, 11]; pericardiectomy, the only definitive treatment for CP, is associated with a perioperative mortality rate of 10% [8]. Interestingly, there is a notably reduced risk of progressing to CP in patients with TBP who underwent pericardiocentesis [12]. This observation is possibly due to pericardiocentesis reducing the proinflammatory fluid in the pericardium [13, 14] thus leading to a potential reduction in the inflammatory response associated with the pericardium, thereby limiting the progression to CP [15].

These results suggest that the patients with TBP who are at the highest risk for developing CP could benefit the most from the use of pericardiocentesis at baseline. The identification of such high-risk individuals, and intervening in the early stages of TBP before the onset of CP, could reduce both the incidence of CP and the need for pericardiectomy. This is especially salient in resource-deficient regions such as sub-Saharan Africa which usually lack the resources and professional expertise required to perform a high-risk procedure like pericardiectomy [13].

Importantly, the use of adjunctive corticosteroids, such as prednisolone, is beneficial in treating both effusive pericarditis and CP [16]. Prednisolone is associated with a reduction both in mortality and the need for pericardiocentesis in patients with pericardial effusion [17] and is associated with reduced adverse events and the need for pericardiectomy in patients with CP [8]. However, prednisolone may be associated with an increased risk of malignancies in HIV-infected patients [18] suggesting that the use of adjunctive corticosteroids may only be most beneficial in the patients who are at the highest risk of progressing to CP.

Currently, there are no standard biomarkers, non-invasive techniques or a defined risk score available to estimate the likelihood of the progression to CP for a patient with a diagnosis of effusive TBP. However, findings from both the Investigation of the Management of Pericarditis (IMPI) Registry and the first IMPI (IMPI-1) randomised clinical trial (RCT), which were multi-centre, prospective studies of patients with TBP in sub-Saharan Africa, suggested that patients with clinical signs of HIV infection were less likely to progress to CP [5, 18].

Specifically, in the IMPI registry, the patients with suspected TBP and clinical characteristics

of HIV infection (a surrogate for advanced HIV/AIDS) had a significantly reduced risk of CP compared to the patients who were not immunocompromised [19]. Additionally, in the IMPI-1 RCT, the proportion of participants with HIV at baseline who developed CP was approximately 50% of those who were HIV-uninfected [18]. Moreover, adjunctive corticosteroids significantly reduced the incidence of CP in both HIV-infected and uninfected individuals [18]. However, there was no clear association between the risk of progressing to CP in patients with TBP who underwent pericardiocentesis at study enrolment [19].

These results suggest that the baseline risk of progressing to CP among patients is not uniform. This is important because it is possible that patients, who are at the highest risk of progressing to CP, could derive the most benefit with the least harm from the use of adjunctive prednisolone and other interventions such as pericardiocentesis as opposed to employing the same interventions in all patients with TBP. The ability to identify such high-risk factors at baseline may allow for risk-tailored treatments for patients who are most likely to benefit and avoid the use of potentially harmful interventions in patients who are least likely to derive any benefit.

1.2 Aims and objectives

In a pre-specified analysis of the data from the IMPI-1 RCT the aims and objectives for this thesis are as follows:

1. Determine the predictors associated with CP in the cohort of participants who did not receive prednisolone treatment at randomisation using baseline demographic, clinical and echocardiographic variables.

2. The derived predictors from Objective 1 will be used to develop a risk score to stratify the whole cohort of participants into high, medium, and low risk for CP.

3. The impact of prednisolone therapy compared to the placebo will be assessed in the cohorts of participants determined by the risk score to have either a high, medium or low risk of CP.

4. The impact of pericardiocentesis at study enrolment on CP will be assessed in the cohorts of participants determined by the risk score to have either a high, medium or low risk of CP.

5. Determine the risk of CP, hospitalisation and death as competing events in the cohorts of participants determined by the risk score to have either a high, medium or low risk of CP. This will be done using competing risks survival analysis.

1.3 Thesis outline

This introductory chapter has provided the background and rationale to the research problem and outlined the aims and objectives for the thesis. Chapter 2 presents a brief overview of the IMPI-1 RCT and provides an exploratory analysis of the baseline participant data from the trial. In Chapter 3 several data preprocessing methods are discussed and applied to the IMPI-1 data to prepare the data for further analysis. Chapter 4 introduces the basic concepts of the use of risk scores in healthcare with specific applications to the field of cardiology and provides a rationale for the use of supervised learning techniques in clinical risk score development. Chapter 5 discusses the concepts, notations and applications of several supervised learning techniques for the classification of CP. Chapter 6 presents the use of a single supervised learning classification technique to develop a risk score for CP and investigates the impacts of adjunctive prednisolone and pericardiocentesis by risk of CP. Time-to-event data analysis is introduced in Chapter 7 with a specific focus on the use of competing risks analysis for when there is more than one event outcome of interest and its application to the events of CP, hospitalisation and death in the IMPI-1 RCT. Finally, the conclusions to the thesis are provided in Chapter 8 with a discussion on the limitations and possible further research opportunities.

Chapter 2

Data description and exploration

2.1 The IMPI-1 RCT

IMPI-1 was the first large-scale, multi-centre RCT in Africa used to assess the efficacy and safety of adjunctive prednisolone and *Mycobacterium indicus pranii* (MIP) immunotherapy in the treatment of TBP [18]. The trial was conducted over five years (2009 – 2014) across 19 hospitals from eight African countries.

The primary study aim was to determine if treatment of TBP with prednisolone and MIP immunotherapy reduced the incidence of the primary composite outcome defined as the first occurrence of CP, cardiac tamponade requiring pericardiocentesis, or death. The secondary aim was to assess the efficacy of these two treatments in reducing the incidence of the individual outcomes that comprised the primary composite outcome and additionally reducing the incidence of hospitalisation. The safety of the two treatments was assessed by quantifying the incidence of opportunistic infections and malignancies.

The trial was conducted using a two-by-two factorial study design. Of the 1400 trial participants, 706 were randomised to receive an initial oral dose of 120mg of prednisolone which was decreased over six weeks while 694 participants received placebo treatment for an equivalent duration. Additionally, of the 1400 trial participants, 625 were randomised to receive five subcutaneous injections of MIP over three months while 625 received a placebo equivalent. Moreover,

approximately 60% of trial participants required the performance of pericardiocentesis at baseline.

While prednisolone did not reduce the incidence of the primary composite outcome compared to the placebo, prednisolone was associated with a reduction in the individual incidences of CP and hospitalisation compared to the placebo in participants both with or without clinical symptoms of HIV. Additionally, MIP was neither associated with the reduced incidence of the primary composite outcome nor the secondary individual outcomes relative to the placebo. Importantly, both prednisolone and MIP treatments corresponded with an increase in the incidence of malignancies in HIV-positive participants.

These findings from the IMPI-1 RCT suggest that while the use of prednisolone is beneficial in reducing the incidence of CP, prednisolone use is associated with other complications particularly in patients with clinical signs of HIV infection. Crucially, because only a minority of trial participants progressed to CP (6.1%), there must be underlying risk factors associated with CP.

Since this trial was not aimed to identify either the risk factors associated with or the potential effects of pericardiocentesis on CP, there is a current gap in our ability to determine the individuals with TBP who are the most at risk for progressing to CP. The identification of these high-risk individuals is imperative in ensuring that treatments such as prednisolone and pericardiocentesis are administered to the patients who could derive the most benefit while patients who could derive the least benefit are not exposed to treatment-related risks unnecessarily.

2.2 Baseline characteristics of participants

To determine if the continuous baseline variables measured in the trial were comparable across the participants randomised to receive either the prednisolone or placebo treatment, the median and interquartile range (IQR) of each variable was determined and compared using the Mann-Whitney U test. Similarly, for the prednisolone and placebo treatment groups, the proportions of participants for each categorical variable were determined and compared using the chi-square or Fisher's exact test. The baseline characteristics of participants were comparable across the placebo and

prednisolone treatments (Table 2.1).

Table 2.1: Baseline characteristics of participants stratified according to treatment. The median of each numeric variable is presented with its interquartile range (IQR) in brackets. The number of participants is shown for the numeric variables for which there was missing patient data at baseline. Categorical variables are presented as the number of participants with percentages in brackets.

Characteristic	Overall (n = 1400)	Treatment		p-value
		Prednisolone (n = 706)	Placebo (n = 694)	
Age (years), median (IQR)	35.56 (28.72 - 46.37)	35.90 (28.72 - 46.34)	35.39 (28.74 - 46.42)	0.76
Weight (kg), median (IQR)	(n = 1377) 58 (51 - 67)	(n = 694) 57.35 (51 - 67)	(n = 683) 58 (50.35 - 66)	0.70
Duration of TB/TBP symptoms (days), median (IQR)	30 (14 - 42)	30 (14 - 60)	30 (14 - 35.75)	0.40
Sex, n (%)				
Male	784 (56)	389 (55.1)	395 (56.92)	
Female	616 (44)	317 (44.9)	299 (43.08)	0.49
Country, n (%)				
South Africa	1014 (72.43)	509 (72.1)	505 (72.77)	
Other	386 (27.57)	197 (27.9)	189 (27.23)	0.78
Definite TBP status, n (%)				
Yes	238 (17)	116 (16.43)	122 (17.58)	
No	1162 (83)	590 (83.57)	572 (82.42)	0.57

Table 2.1 Continued: Baseline characteristics of participants stratified according to treatment. The median of each numeric variable is presented with its interquartile range (IQR) in brackets. The number of participants is shown for the numeric variables for which there was missing patient data at baseline. Categorical variables are presented as the number of participants with percentages in brackets.

Characteristic	Overall (n = 1400)	Treatment		p-value
		Prednisolone (n = 706)	Placebo (n = 694)	
On TB medication, n (%)				
Yes	1025 (73.21)	522 (73.94)	503 (72.48)	
No	58 (4.14)	25 (3.54)	33 (4.76)	
Unknown	317 (22.64)	159 (22.52)	158 (22.77)	0.25
NYHA functional class, n (%)				
Class I	256 (18.29)	137 (19.41)	119 (17.15)	
Class II	694 (49.57)	342 (48.44)	352 (50.72)	
Class III	330 (23.57)	163 (23.09)	167 (24.06)	
Class IV	117 (8.36)	63 (8.92)	54 (7.78)	
Unknown	3 (0.21)	1 (0.14)	2 (0.29)	0.57
Tachycardia (heart rate $>$ 100 BPM), n (%)				
Yes	790 (56.43)	390 (55.24)	400 (57.64)	
No	608 (43.43)	314 (44.48)	294 (42.36)	
Unknown	2 (0.14)	2 (0.28)	0	0.40

Table 2.1 Continued: Baseline characteristics of participants stratified according to treatment. The median of each numeric variable is presented with its interquartile range (IQR) in brackets. The number of participants is shown for the numeric variables for which there was missing patient data at baseline. Categorical variables are presented as the number of participants with percentages in brackets.

Characteristic	Overall (n = 1400)	Treatment		p-value
		Prednisolone (n = 706)	Placebo (n = 694)	
Hypotension (systolic blood pressure $\leq$ 90 mmHg), n (%)				
Yes	106 (7.57)	59 (8.36)	47 (6.77)	
No	1293 (92.36)	646 (91.50)	647 (93.23)	
Unknown	1 (0.07)	1 (0.14)	0	0.26
Peripheral oedema, n (%)				
Yes	575 (41.07)	289 (40.93)	286 (41.21)	
No	825 (58.93)	417 (59.07)	408 (58.79)	0.92
Palpable pulsus paradoxus, n (%)				
Yes	270 (19.29)	141 (19.97)	129 (18.59)	
No	1130 (80.71)	565 (80.03)	565 (81.41)	0.51
Anaemia (haemoglobin $\leq$ 10 g/dL), n (%)				
Yes	748 (53.43)	385 (54.53)	363 (52.31)	
No	645 (46.07)	319 (45.18)	326 (46.97)	
Unknown	7 (0.5)	2 (0.28)	5 (0.72)	0.45

Table 2.1 Continued: Baseline characteristics of participants stratified according to treatment. The median of each numeric variable is presented with its interquartile range (IQR) in brackets. The number of participants is shown for the numeric variables for which there was missing patient data at baseline. Categorical variables are presented as the number of participants with percentages in brackets.

Characteristic	Overall (n = 1400)	Treatment		p-value
		Prednisolone (n = 706)	Placebo (n = 694)	
Renal impairment (creatinine $>$ 105 μmol/L), n (%)				
Yes	160 (11.43)	80 (11.33)	80 (11.53)	
No	1112 (79.43)	564 (79.89)	548 (78.96)	
Unknown	128 (9.14)	62 (8.78)	66 (9.51)	0.86
White cell count $>$ 10x10^9/L, n (%)				
Yes	99 (7.07)	42 (5.95)	57 (8.21)	
No	1297 (92.64)	662 (93.77)	635 (91.50)	
Unknown	4 (0.29)	2 (0.28)	2 (0.29)	0.10
Effusion size (cm), n (%)				
Small ($<$1cm)	106 (7.57)	50 (7.08)	56 (8.07)	
Medium (1-2cm)	331 (23.64)	172 (24.36)	159 (22.91)	
Large ($\geq$2cm)	922 (65.86)	462 (65.44)	460 (66.28)	
Unknown	41 (2.93)	22 (3.12)	19 (2.74)	0.67

Table 2.1 Continued: Baseline characteristics of participants stratified according to treatment. The median of each numeric variable is presented with its interquartile range (IQR) in brackets. The number of participants is shown for the numeric variables for which there was missing patient data at baseline. Categorical variables are presented as the number of participants with percentages in brackets.

Characteristic	Overall (n = 1400)	Treatment		p-value
		Prednisolone (n = 706)	Placebo (n = 694)	
Cardiac tamponade at presentation, n (%)				
Yes	556 (39.71)	278 (39.38)	278 (40.06)	
No	423 (30.21)	213 (30.17)	210 (30.26)	
Unknown	421 (30.07)	215 (30.45)	206 (29.68)	0.91
Effusive-constrictive pericarditis at presentation, n (%)				
Yes	441 (31.5)	230 (32.58)	211 (30.40)	
No	559 (39.93)	279 (39.52)	280 (40.35)	
Unknown	400 (28.57)	197 (27.90)	203 (29.25)	0.48
Pericardiocentesis performed at randomisation, n (%)				
Yes	846 (60.43)	427 (60.48)	419 (60.37)	
No	554 (39.57)	279 (39.52)	275 (39.63)	0.97

Table 2.1 Continued: Baseline characteristics of participants stratified according to treatment. The median of each numeric variable is presented with its interquartile range (IQR) in brackets. The number of participants is shown for the numeric variables for which there was missing patient data at baseline. Categorical variables are presented as the number of participants with percentages in brackets.

Characteristic	Overall (n = 1400)	Treatment		p-value
		Prednisolone (n = 706)	Placebo (n = 694)	
Pulmonary infiltrates on chest radiograph, n (%)				
Yes	420 (30)	217 (30.74)	203 (29.25)	
No	856 (61.14)	422 (59.77)	434 (62.54)	
Unknown	124 (8.86)	67 (9.49)	57 (8.21)	0.43
Atrial fibrillation on electrocardiogram, n (%)				
Yes	57 (4.07)	33 (4.67)	24 (3.46)	
No	1001 (71.5)	501 (70.96)	500 (72.05)	
Unknown	342 (24.43)	172 (24.36)	170 (24.50)	0.25
MIP treatment, n (%)				
MIP	625 (44.64)	316 (44.76)	309 (44.52)	
Placebo	625 (44.64)	315 (44.62)	310 (44.67)	
Treatment stopped	150 (10.71)	75 (10.62)	75 (10.81)	0.99

While all the participants were suspected of having TBP, only 17% had either a confirmed bacteriologic or histologic diagnosis of TBP (Table 2.1). Most trial participants were enrolled at South African centres (72.43%) and 73.21% of participants were using anti-tuberculous treatment

at enrolment (Table 2.1). Importantly, there were no strong correlations between the baseline continuous variables (Appendix A, Figure A.1).

Most participants were classified as not having hypotension, renal impairment, palpable pulsus paradoxus or atrial fibrillation on the electrocardiogram (Table 2.1). However, approximately half of the participants were categorised as having a New York Heart Association (NYHA) functional class II and 53.43% were anaemic (Table 2.1).

Almost two-thirds of the participants had a large pericardial effusion $\geq$2cm and over one-third were diagnosed as having cardiac tamponade at enrolment (Table 2.1). Consequently, 60.43% of study participants underwent pericardiocentesis at baseline (Table 2.1). Moreover, approximately one-third of participants were characterised as having effusive-constrictive pericarditis at enrolment (Table 2.1).

Crucially, this trial enrolled 939 (67%) participants who were either HIV-positive or had clinical signs of HIV infection, These participants had similar baseline characteristic profiles across the two treatment groups (Appendix A, Table A.1). Of these participants, only 21.19% were on anti-retroviral treatment at enrolment, 36.1% had a CD4 count of 50 – 200 cells/μL and 56.44% had HIV-related opportunistic infections (Appendix A, Table A.1)

2.3 Summary

This chapter has presented a summary of the rationale, methods and main findings of the original IMPI-1 RCT as a motivation behind the aims and objectives of the research for this thesis. The baseline characteristics of the trial participants were presented stratified according to the randomised baseline prednisolone and placebo treatments. Since all the presented baseline characteristics were similar according to the randomised treatment of prednisolone or placebo we can assume that these two groups of participants only differ by the treatment to which they were randomised at baseline.

Chapter 3

Data preprocessing

3.1 Introduction

Since the primary aim of this research was to develop a risk score for CP using the IMPI-1 RCT data, several preprocessing applications to the data need to be applied before the construction of the risk score for CP. Specifically, these preprocessing considerations are (1) training and test datasets, (2) multiple imputation of missing predictor observations, (3) standardisation of continuous predictors, and (4) addressing the low prevalence of the CP outcome. This chapter will discuss the rationale, methods and present the applications of the aforementioned data preprocessing steps to the IMPI-1 RCT data.

3.2 Training and test datasets

All statistical models commonly share the properties of bias and variance [20]. Model bias infers that the statistical model cannot truly capture the underlying biological process especially if the model is too simple. Therefore one would assume that to address model bias, a more complex model is required to better describe the real-world problem [20]. However, increasing the model's complexity naturally increases the model's variance which means that the model predictions are subject to large variation when the model is fit to new data [20].

Since a statistical model is built or trained on a particular dataset, referred to as the training dataset,

often using an iterative process, the model is capable of accurately learning the specific structures present in the data on which the model was trained thus allowing for incredibly accurate outcome predictions in that specific dataset [20]. However, to address the issues of bias and variance that are inherent in all statistical models, the generalisability of the model to new observations needs to be assessed using an independent dataset, referred to as the test dataset, to which the model was not previously exposed [20].

Commonly, we do not have two or more entirely separate datasets on which to train and test our statistical models. Therefore the observations of the initial dataset are randomly divided into two sets of observations that are used to comprise the training and test datasets for model training and testing respectively. The training dataset commonly comprising 70% of the observations from the original dataset and the remaining 30% as the test dataset.

3.2.1 Results

To construct a risk score to predict the baseline probability of developing CP, the model training only considered the 694 participants who did not receive the prednisolone treatment (Figure 3.1). This is because prednisolone use is known to be associated with a decreased incidence of CP [18]. Additionally, the participants who received the MIP treatment were not excluded from this selection since MIP was not determined to be associated with the CP outcome [18].

Of the 694 participants who were randomised to receive the placebo treatment as opposed to the prednisolone treatment, 70% (n = 486) were randomly sampled according to the binary outcome of CP and were used for training the statistical models while the remaining 30% (n = 208) were used to assess the model performance (Figure 3.1). The dataset was split according to the CP outcome to ensure that equal proportions of participants who did or did not experience CP were represented in both the training and test datasets.

Importantly, splitting the subset of participants who received the placebo treatment into the training and test datasets was performed before the other preprocessing steps such as multiple imputation

and standardisation. This was to avoid "data leakage" and to ensure that none of the information in the test dataset was included in the dataset used to train the statistical models [21].

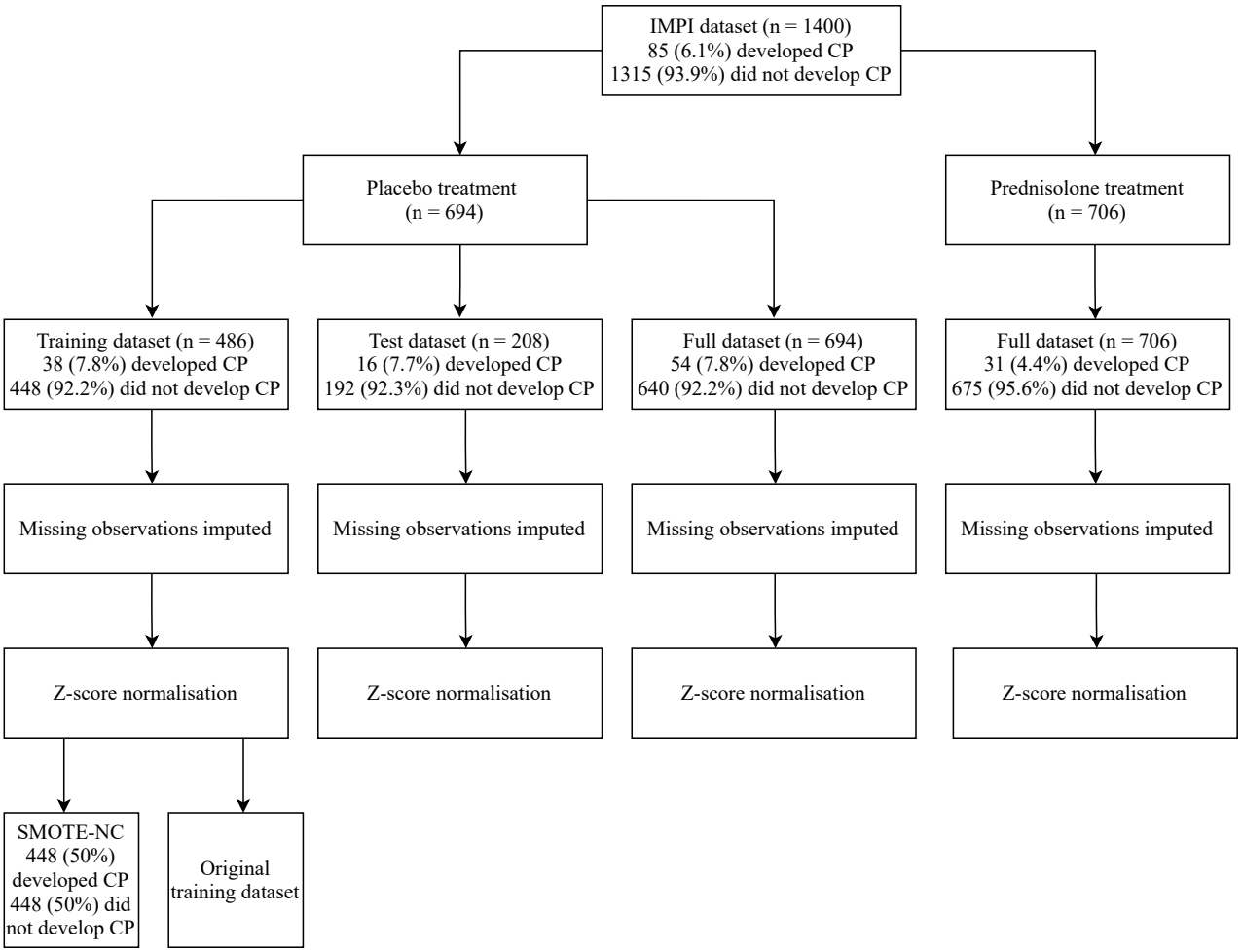

Figure 3.1: Overview of preprocessing techniques applied to IMPI-1 data subsets.

3.3 Multiple imputation

3.3.1 Characteristics of missing data

Commonly in medical data, especially from RCTs, most subject records do not contain a complete set of feature observations [22]. Since most statistical models require observations to have a complete set of feature values, if a classifier was applied to an incomplete dataset, all the observations with at least one missing predictor or outcome value would be excluded [22]. This

procedure is also known as complete case analysis [23].

Complete case analysis of a dataset that contains observations with missing values can be problematic as the complete cases potentially provide a biased representation of the full sample [23]. This results in biased estimates of the model parameters [23]. Furthermore, complete case analysis should only be used when the missing data are assumed to be missing completely at random (MCAR) [24].

The assumption of MCAR means that the missing feature values in the data are not reliant on either the observed or unobserved feature values in the dataset [24]. Typically, it is almost impossible to substantiate the assumption of MCAR observations [24] and hence complete case analysis is usually an inappropriate analysis method.

Alternatively, observations that are assumed to be missing at random (MAR) means that the probability of a missing feature value is reliant on the observed feature values in the dataset alone and independent of the unobserved feature values [24]. However, if the missing values depend both on the observed and unobserved values of other features, then the observations are assumed to be missing not at random (MNAR) [24]. While it can be hard to distinguish if the structure of missingness is due to MAR or MNAR, if the missing observations are assumed to be MAR, then the observed values of other features in the data can be used to predict the missing values of any predictor of interest [24].

Multiple imputation can be used to produce many datasets which contain predictions for missing variable values using the observed values of other features in the original dataset. The missing feature values are derived as an average of the predicted values across the many generated datasets. [24]. In the IMPI-1 dataset, since the baseline features for which the observations were missing were assumed to be MAR, multiple imputation was used to produce complete cases of observations.

3.3.2 Random forest multiple imputation

While several multiple imputation methods are available to predict missing feature values in a dataset using the observed values of other features, the majority are parametric meaning that certain distributional assumptions about the data are made [22]. Often there are complicated nonlinear structures and interactions in the data that make the specification of a parametric multiple imputation model complicated and potentially inaccurate [22].

The random forest supervised learning algorithm, which will be discussed in greater detail in Chapter 5, is an effective nonparametric classifier that makes no distributional assumptions about the data and is capable of learning nonlinear and interactive effects in the data [25]. The random forest algorithm can be extended to a multiple imputation model whereby a model is trained using the observed feature values in the data to predict missing values of other features [22].

For random forest imputation, we are interested in imputing the missing values $x_{mis}^{(s)}$ for a random feature $\boldsymbol{X_s}$, $s = 1, \cdots, p$. The data are divided according to:

- $\boldsymbol{y}_{obs}^{(s)}$ which are the observed values of the feature $\boldsymbol{X_s}$,
- $\boldsymbol{y}_{mis}^{(s)}$ which are the unobserved values of the feature $\boldsymbol{X_s}$,
- $\boldsymbol{x}_{obs}^{(s)}$ which are all other features excluding $\boldsymbol{X_s}$ that contain observed values,
- $\boldsymbol{x}_{mis}^{(s)}$ which are all other features excluding $\boldsymbol{X_s}$ that contain missing values.

The features $\boldsymbol{X_s}$ are sorted according to the number of missing observations in each feature. The feature containing the fewest missing values is imputed first [22]. The imputation for each variable is done by training a random forest model to first predict the observed values for each variable $\boldsymbol{y}_{obs}^{(s)}$ using the features $\boldsymbol{x}_{obs}^{(s)}$. Once the random forest is trained it is used to predict the missing values $\boldsymbol{y}_{mis}^{(s)}$ [22]. The procedure is then repeated for every feature with missing values. Importantly, two hyperparameters of random forest models namely (1) the number of trees built and (2) the number of variables considered in each tree split can be tuned when training the model for multiple imputation.

The overall performance of the random forest imputation model for continuous variables can be assessed using the normalised root mean squared error (NRMSE):

$$NRMSE = \sqrt{\frac{\text{mean}\left[(\boldsymbol{X}^{true} - \boldsymbol{X}^{imp})\right]^2}{\text{var}(\boldsymbol{X}^{true})}},$$

where $\boldsymbol{X}^{true}$ and $\boldsymbol{X}^{imp}$ denote the observed and imputed feature values respectively. The performance of the model for categorical variable imputation is assessed using the proportion of falsely classified observations [22]. Additionally, since the random forest algorithm typically randomly samples two-thirds of the training data to build the model, the remaining one-third can be used to assess the out-of-bag (OOB) error for each variable to determine the model-specific hyperparameters and as a measure of model performance [22].

Since the random forest algorithm can be applied to both continuous and categorical high-dimensional data and does not require any distributional assumptions about the data, random forest imputation tends to outperform other parametric imputation algorithms such as k-nearest neighbours imputation and multivariate imputation by chained equations [22]. Therefore, the random forest multiple imputation method was applied to the IMPI-1 data subsets to impute any missing values of the features considered in the models used to predict the CP outcome.

3.3.3 Results

The IMPI-1 data was partitioned into four subsets (placebo training, placebo test, placebo full and prednisolone full), all of which contained missing predictor observations. Therefore, the predictors that contained missing values in all four subsets of the IMPI-1 data were imputed separately using random forest multiple imputation (Figure 3.1).

Importantly, several variables were excluded from the imputation and consequently were excluded from all future analyses due to these variables potentially violating the assumptions of MCAR or MAR. Specifically, this included any predictor that was only measured in participants who were classified as HIV positive (Appendix A, Table A.1).

The OOB measures of NRSME for continuous variables and the proportion falsely classified for categorical variables were used to select the model-specific hyperparameters for each data subset (Table 3.1).

Table 3.1: Summary of hyperparameter tuning results in the random forest imputation model for each IMPI-1 data subset. For each data subset, the final OOB error is given separately for the imputed continuous variables (NRMSE) and the imputed categorical variables (proportion falsely classified)

	Data subset			
	Placebo training	Placebo test	Placebo full	Prednisolone full
Number of trees tested	50, 100, 200, 300, 400, 500, 600, 700, 800	50, 100, 200, 300, 400, 500	50, 100, 200, 300, 400, 500, 600	50, 100, 200, 300, 400, 500, 600, 700, 800
Number of variables tested in each tree split	2, 3, 4, 5, 6, 7, 8	2, 3, 4, 5, 6	3, 4, 5, 6, 7	3, 4, 5, 6, 7, 8
Final hyperparameter combination	600 trees with 6 variables considered at each tree split	200 trees with 4 variables considered at each tree split	400 trees with 5 variables considered at each tree split	600 trees with 5 variables considered at each tree split
OOB NRMSE for final hyperparameter combination	0.029	0.026	0.028	0.026
OOB proportion falsely classified for final hyperparameter combination	0.100	0.091	0.098	0.121

To assess the performance of the random forest multiple imputation model for each dataset, the density distributions of the continuous variables and the proportion of variable classes were

compared for the imputed and original data subsets.

The imputed categorical variables had similar proportions to the categorical variable proportions in the respective original datasets (Appendix A, Tables A.2 - A.5). The only continuous variable that required imputation was participant weight at baseline. The density distribution of the imputed weight variable overlapped with the density distribution of the weight variable in the original datasets (Appendix A, Figure A.2) suggesting that the random forest imputation algorithm converged and importantly, appropriate estimations of the missing values for each imputed predictor variable were approximated. Notably, multiple imputation was performed before standardisation and balancing of the placebo training dataset because both of these preprocessing steps require complete cases of observations.

3.4 Standardisation

Most clinical datasets consist of continuous variables that typically have different units and ranges of measurement. Standardisation of continuous variables is an important preprocessing step before statistical modelling and is performed to ensure that the ranges of continuous variables are consistent thus preventing a variable's scale and variance from heavily influencing its effect size in the model [20].

3.4.1 Z-score normalisation

To ensure consistent ranges of the numeric variables in the different IMPI-1 data subsets, the age, weight and duration of TB/TBP symptoms predictors were standardised after multiple imputation using z-score normalisation according to:

$$z = \frac{x - \bar{x}}{S},$$

where $\bar{x}$ and S are the sample mean and standard deviation of the variable x respectively. The use of z-score normalisation resulted in the continuous variables having a mean of zero and a standard deviation of one. Standardisation was performed before balancing the placebo training

dataset (Figure 3.1) as SMOTE-NC, which is discussed in the following section, utilises a statistical modelling algorithm that benefits from data standardisation.

3.5 Imbalanced data

Often in a clinical context, we are interested in predicting if certain individuals will or will not develop a particular health outcome. The prediction of a binary outcome is known as classification [20]. However, if the health outcome is rare in the population, the sample of data tends to consist largely of individuals who did not develop the outcome which makes the data sample imbalanced in terms of the outcome of interest [26]. For example, in the IMPI-1 dataset, only 6.1% of participants developed the CP outcome.

Most statistical models are susceptible to weak predictive performance on imbalanced datasets because the models are biased towards predicting the majority outcome class [26, 27]. This results in excessive misclassification of the minority outcome class [28] which, in healthcare, are usually the cases we are the most interested in identifying. Therefore, to build a predictive classification model that is capable of accurately classifying the subjects of the minority class, a dataset in which the outcome of interest is balanced is ideal.

Several data-based statistical strategies are available to create a balanced dataset including but not limited to random over-sampling of the minority class, random under-sampling of the majority class and synthetic minority over-sampling [29]. Neither random over-sampling nor under-sampling is typically the preferred method of addressing imbalance as they can result in model overfitting and biased model estimates respectively [28].

While there is no "gold-standard" technique recommended for addressing imbalanced data, synthetic minority over-sampling typically outperforms either random over- or under-sampling [30, 31]. The following section will discuss the synthetic minority over-sampling technique (SMOTE).

3.5.1 SMOTE

Contrastingly to random over-sampling of the minority class in which minority class cases are replicated, SMOTE uses the k-nearest neighbours algorithm to create fabricated minority class cases [31].

Specifically, SMOTE creates a sample of synthetic minority cases, s, according to:

$$s = \boldsymbol{x} + \phi(\boldsymbol{x} - v),$$

where $\boldsymbol{x}$ is the vector of predictors being considered, v is the selection of nearest neighbours to that feature vector and ϕ is known as the gap. The gap is a random number between 0 and 1 that is used to randomly select a position on the line segment that joins the two neighbours under consideration [30, 31]. The generation of artificial minority class cases allows for greater generalisability of the classification model.

Since SMOTE relies on the k-nearest neighbours algorithm which uses a measure of distance to generate synthetic minority class cases, SMOTE can only be applied to datasets in which all the features are continuous [31]. However, SMOTE-NC (synthetic minority over-sampling technique-nominal continuous) is an extension of the SMOTE algorithm that can be used for a mixed dataset, such as the IMPI-1 RCT dataset which contains both continuous and categorical features [31].

While SMOTE uses the Euclidean distance between neighbours as a measure of distance, SMOTE-NC generates artificial categorical predictor observations by using a majority vote of the category classes in the neighbouring observations [31].

3.5.2 Results

To limit the possibility of the classification models being biased towards predicting the majority class of participants who did not develop CP in the IMPI-1 dataset, SMOTE-NC was used to

create synthetic observations of participants who developed CP (Figure 3.1).

Importantly, unlike multiple imputation methods which are expected to preserve the original distributions and proportions of continuous and categorical variables respectively in a dataset, SMOTE-NC can result in altered variable distributions and proportions by synthesising more observations of the minority class. Therefore, there is no exploratory method to assess the performance of the SMOTE-NC algorithm.

Consequently, the placebo training dataset was duplicated and SMOTE-NC was only applied to one of the placebo training datasets thus creating SMOTE-NC balanced and original imbalanced versions of the placebo training dataset (Figure 3.1). This was done because each training dataset was used separately to train the classification models to predict CP. Therefore the placebo test dataset, which was not balanced, was used to compare the performance of the models built using either the SMOTE-NC balanced or original imbalanced placebo training datasets.

3.6 Summary

This chapter has presented several data preprocessing techniques that were applied to the IMPI-1 dataset. Specifically, the data subset of participants that received the placebo treatment in contrast to the prednisolone treatment was split into training and test data subsets. Multiple imputation, under the assumption of MAR, was used to address the missing predictor observations. Data standardisation was performed to ensure consistent scales and ranges of the categorical predictors. Finally, SMOTE-NC was applied to one of the two duplicated placebo training datasets to determine if addressing the imbalance in the data would improve the performance of the classification models.

The introduction to and application of the preprocessing techniques that are relevant to the IMPI-1 RCT data allow for the appropriate training of a classification model to predict the risk of CP. However, before the specific classification methods are introduced, the following chapter will discuss the concepts surrounding the use of risk scores in healthcare as an introduction to the

use of supervised learning techniques for the classification of a binary clinical outcome.

Chapter 4

Risk scores for clinical outcomes

4.1 Introduction

This chapter presents the application of predictive modelling in developing risk scores for clinical outcomes. The use of risk scores in healthcare are discussed and a summary of common attributes shared by risk scores are described. The application of risk scores in the field of cardiology, specifically heart failure, is introduced. Two popular examples of heart failure risk scores and their shortcomings are presented. Finally, this chapter introduces the recent use of machine learning classification methods and their applications to risk scores for clinical outcomes.

4.2 The use of risk scores in healthcare

Risk scores are a unique class of predictive models in statistics and are an integral component to precision and evidence-based medicine or healthcare [32]. Specifically, a risk score in healthcare combines multiple predictors or risk factors in a predictive model that is used to determine an individual's absolute risk for a medical outcome [33, 34].

In healthcare, risk scores serve to assist clinicians in making appropriate decisions regarding the diagnosis, prognosis and treatment of medical outcomes to improve the quality of life of an individual [35, 36]. Additionally, risk scores are used to provide crucial information to patients such that the patient can be involved in the decision process regarding treatment interventions or

lifestyle changes [32].

Generally, risk scores are derived from classification models which can range from being incredibly simplistic and easy to interpret such as logistic regression to vastly complex with little interpretability such as artificial neural networks [35]. While the statistical basis of the classification techniques used to define different risk scores may be highly variable, all predictive models used to create risk scores share similar characteristics [35].

Typically, a risk score is built using selected risk factors from a set of predetermined predictors which can include patient demographics, clinical history, symptoms and treatments [33]. Each selected risk factor is assigned a weight that is a proxy for the relative importance of that feature in determining the outcome [33]. The score is determined by combining the weights of the risk factors and is translated into a series of levels or risk categories based on a set of threshold values [33].

Importantly, all predictive models used to develop risk scores are assessed in terms of the model's ability to separate individuals into the different classes of the outcome, known as discrimination, and the model's ability to predict an individual's outcome class accurately, known as calibration [37]. Moreover, both internal and external validation techniques are crucial in determining the predictive performance of the classification model both within the current data sample and across different, ideally diverse, samples respectively [32, 33].

Since the recent developments of new statistical methods for classification and advancements in computational power over the last few decades, hundreds of published risk scores have been developed for use in a diverse range of clinical practices [32] including surgery [38], public health [32], and many specific health-related outcomes [39–41].

Although due to flawed methodology [42], lack of external validation [43] and clinical implementation [44], only a small fraction of the available published risk scores are utilised in widespread clinical practice [45]. Notably, some of the most well-known and successful

implementations of risk scores in healthcare are from the field of cardiology. Specifically, risk scores relating to heart failure and its causes have been at the forefront of predictive medicine [46].

4.3 Risk scores relating to heart failure

Unsurprisingly, risk scores and predictive models related to heart failure and its causes have dominated the literature surrounding risk scores for decades [46] as causes of heart failure are among some of the leading sources of morbidity and mortality worldwide [47]. Moreover, the morbidity associated with heart failure arising from diseases such as cardiovascular disease is often due to a lack of early disease detection and diagnosis [47] which motivates the need for being able to determine an individual's level of risk.

The Framingham Risk Score is used to predict the 10-year risk of cardiovascular disease [48] and resulted from research conducted in the Framingham Heart Study. The Framingham Risk Score is arguably one of the most successful examples of a risk score used in widespread clinical practice [49]. Another popular risk score that aims to predict patient survival after the occurrence of heart failure is the Seattle Heart Failure Model [50].

Despite the extensive use of both the Framingham Risk Score and the Seattle Heart Failure Model in cardiology [51, 52], they have both been criticised. Specifically, the Framingham Risk Score is known to both over and underestimate the risk of cardiovascular disease in external study cohorts [49, 53] and the robustness and generalisability of the Seattle Heart Failure Model are questionable [52].

The Framingham Risk Score and Seattle Heart Failure Model make use of multivariate logistic regression and multivariate standard survival analysis respectively [48, 50]. While these models are simple to implement and interpret, there are several downfalls such as the assumptions of linear relationships between the risk factors and the log odds or hazard of the outcome and additionally, that the risk factors are independent of one another. These assumptions are often not met when classifying a medical outcome based on several risk factors [54]. Moreover, the lack of model

flexibility of simpler classification methods such as the logistic regression model can lead to low predictive power and poor sensitivity and specificity when the model is used to predict a medical outcome [53].

4.4 Supervised learning for clinical outcome classification

The shortcomings of the previously described heart failure models have led to increased uptake of the use of other supervised learning classification techniques, such as support vector machines, classification trees, ensemble methods, and artificial neural networks, in developing predictive models for heart failure [55–58].

The main advantages of these machine learning methods for the classification of clinical outcomes are that these algorithms do not assume a linear relationship between the risk factors and the outcome nor make any assumptions about the independence of the risk factors. This allows for increased model flexibility and more importantly, potentially increased predictive power and model performance which is crucial in predicting clinical outcomes [56–58].

The development of these classification techniques and the ability to implement them in statistical software allow for the discovery of previously unidentified risk factors associated with clinical outcomes and more importantly, the establishment of novel risk scores in clinical practice.

4.5 Summary

This chapter has provided a brief introduction to the rationale and applications of risk scores in healthcare with a specific focus on the field of cardiology. Importantly, the commonalities and shortcomings of predictive modelling and the development of risk scores for clinical outcomes were highlighted as the motivation for assessing the use of various supervised learning classification techniques to build a risk score for CP. The following chapter aims to illustrate the important concepts and notations associated with six popular supervised learning classification techniques for predictive modelling and their application to the prediction of CP.

Chapter 5

Supervised learning for classification

5.1 Introduction

Supervised learning is a branch of statistical modelling that is used to predict and make inferences about an outcome or response using a set of input features or risk factors [20]. Classification is the field of supervised learning that is used when the outcome of interest is binary or categorical [20].

Classification is commonly used in medical research because often we are interested in being able to classify subjects or patients as either having a clinical outcome or not [55]. Specifically, medical researchers are typically interested in classifying or predicting either the presence or absence of disease given a set of patient-specific risk factors [55]. Therefore, clinicians require appropriate classification algorithms that are both accurate and comprehensible to make informed decisions regarding medical outcomes [55].

Classification algorithms range from being simple and easily interpretable at the cost of potentially lower accuracy such as logistic regression to having high predictive power at the cost of interpretability such as artificial neural networks [20]. Historically, medical research has primarily utilised the logistic regression model as the primary supervised learning algorithm to classify a binary health outcome [59]. However, in recent years, several other supervised learning techniques have become increasingly popular classification tools in the medical sciences [59].

Although numerous classification algorithms are available, the most successful and frequently used techniques for modelling a binary health outcome include logistic regression, tree-based methods, support vector machines and artificial neural networks [39, 55, 58–64]. However, the consensus of which algorithm is the best in terms of predictive accuracy is still unclear [39, 55, 58–64].

Nevertheless, complex machine learning techniques are becoming increasingly popular in the medical sciences to build predictive models using patient data arising from not only observational studies but also from randomised clinical trials [65–70].

This chapter introduces the concepts, features and notations of six widely utilised supervised learning classification methods that are used for the predictive modelling of a binary health outcome and their applications to the classification of CP using the baseline participant information from the IMPI-1 RCT.

The first section presents the logistic regression model as the standard classification technique for a binary outcome. Lasso regularisation for variable selection is introduced with a brief overview of k-fold cross-validation for internal validation of a supervised learning model. The second, third and fourth sections describe three tree-based classification algorithms namely (1) classification decision trees, (2) bootstrap aggregation and random forests, and (3) boosted trees. The fifth and sixth sections describe the concepts of support vector machines and artificial neural networks respectively. Common performance measures used to assess the calibration and discriminative ability of classification models will be discussed and applied to the IMPI-1 RCT data.

5.2 Logistic regression

This section discusses the concept and notation of logistic regression adapted from Dobson and Barnett [71], lasso regularisation for feature selection and k-fold cross-validation for internal validation adapted from James et al. [20] and Hosmer et al. [72].

Logistic regression falls into the class of generalised linear models and is one of the most commonly used and well-known classification methods to predict a medical outcome that is binary. Logistic regression predicts the probability that a binary outcome belongs to a particular class given a set of risk factors.

The binary outcome, Y, is defined as:

$$Y = \begin{cases} 1 & \text{if the outcome is present} \\ 0 & \text{if the outcome is absent.} \end{cases}$$

The probability that the outcome is present is denoted as $Pr(Y = 1) = \pi$ and the probability that the outcome is absent is denoted as $Pr(Y = 0) = 1 - \pi$.

5.2.1 The logistic regression model

Logistic regression aims to model the probability that $Y = 1$ for a subject i where $i = 1, \cdots, n$:

$$g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where g is the link function, $\boldsymbol{x}_i$ is the vector of predictors and $\boldsymbol{\beta}$ is the vector of regression parameters. The logit link function is used to restrict the probability of the outcome, π_i, to the interval $(0, 1)$:

$$g(\cdot) = \log\left(\frac{\pi_i}{1 - \pi_i}\right),$$

where the logit function is interpreted as the log of the odds of the outcome.

Therefore, the logistic model, which describes the relationship between the log odds of the outcome for a subject i and a given set of risk factors, $\boldsymbol{x_i}$, is denoted by:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x_i}^T \boldsymbol{\beta}.$$

The regression coefficients, $\boldsymbol{\beta}$, are estimated by maximising the log-likelihood function:

$$l(\pi; y) = \sum_{i=1}^{N} \left[y_i \log \pi_i + (n_i - y_i) \log(1 - \pi_i) + \log \begin{pmatrix} n_i \\ y_i \end{pmatrix} \right].$$

5.2.2 Feature selection

Feature selection is crucial in building a predictive model because we seek to find the simplest model with the best predictive performance to balance the trade-off between bias and variance [73]. While the inclusion of potentially irrelevant predictors can lead to a model with low bias, the variance will be high leading to a lack of model generalisability [73]. Several strategies are available to select the most appropriate features to include in a predictive model such as best subset selection, stepwise selection, ridge and lasso regression [20].

Best subset selection examines all possible explanatory variable combinations to determine a selection of features that will be included in the final model [20]. However, this method is commonly not used due to the computational intensity required to fit all possible models especially when the dataset contains many features [20].

A practical alternative to best subset selection is stepwise selection [20]. Feature selection using stepwise selection is performed by sequentially including and excluding predictors from the model using internal validation measures of variable importance or model performance such as the p-value and R^2 respectively [20]. While stepwise selection procedures are easily implementable and the model is usually simple to interpret, the discrete nature of adding and removing variables from the model tends to result in highly variable and biased model parameters [73].

Regularisation methods, such as ridge and least absolute shrinkage and selection operator (lasso) regression, are more modern feature selection techniques that often are a better alternative to stepwise selection procedures [73]. While ridge regression shrinks the model parameters that contain minimal explanatory power close to zero, lasso regression both shrinks and completely reduces some of the model parameters to zero [73]. Therefore lasso regression results in models

that are simpler and easier to interpret than ridge regression [73]. Additionally, lasso regression tends to outperform ridge regression in terms of the model's predictive performance when the explanatory variables are highly correlated [73].

5.2.3 Lasso regression

Lasso regression is a popular feature selection method that allows for a reduction in the number of risk factors in the final predictive logistic regression model [73]. While lasso regression is commonly used when the number of predictors in the dataset exceeds the number of observations, it can also be used for datasets where the number of features is large but not greater than the number of observations [73].

In lasso regression a hyperparameter, λ, is used to simultaneously estimate and select the model parameters that minimise the classification error of the model [73]. Lasso regression is performed on the logistic model by maximising the log-likelihood function of the logistic model parameters subject to the hyperparameter, λ [74]:

$$l_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i x_i^T \boldsymbol{\beta} - \log(1 + e^{x_i^T \boldsymbol{\beta}}) \right] - \lambda \sum_{j=1}^{p} \mid \boldsymbol{\beta}_j \mid .$$

By increasing the tuning parameter, λ, some of the regression coefficients that are estimated from the likelihood function will be equal to zero [73].

In contrast to stepwise selection which uses internal model validation techniques, lasso regression uses cross-validation to find an appropriate estimate for the hyperparameter λ [73]. Compared to internal model validation, cross-validation results in less biased regression coefficients [73]. Commonly, k-fold cross-validation is performed to select the model-specific tuning parameters and is discussed in the following section.

5.2.4 k-fold cross-validation

Several machine learning models can be built to predict the same outcome given a set of risk factors or model-specific tuning parameters [20]. However, we need to choose a final model or choose the tuning parameters that optimise the predictive performance of the model. Cross-validation is a way of internally checking the performance of a machine learning model on data that the model has never seen.

A model's generalisability cannot be assessed based on its ability to accurately predict those outcomes of the training data set on which it was constructed. The generalisability of a model can be assessed by the model's performance on data to which it was not previously exposed. Since we often have limited training data, we can use a small sample of the training data to validate the performance of the model and the model's ability to predict outcomes for observations it has not seen previously.

For k-fold cross-validation, the observations in the training dataset are randomly divided into k samples of roughly equal size. The first sample is used to validate the model's performance while the model is fit to the data in the $k-1$ folds. The classification error is determined for the i^{th}, where $i = 1,2,\ldots,k$, validation sample. The procedure is then repeated for all k folds such that in each iteration a different sample of the training dataset is used as the validation set. This results in k estimates of the classification error for the model. The k classification errors are used to compute the k-fold cross-validation error estimate by averaging the classification errors for the k samples:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i, \qquad \text{where } Err_i = I(y_i \neq \hat{y}_i).$$

Generally, k is set to ten as this is computationally feasible and tends to result in little bias and variance [75]. While k-fold cross-validation has been presented as the method to select the regularisation hyperparameter λ in lasso regression, cross-validation was used for every machine learning classification method to select the model-specific hyperparameters that were used in training the final model to predict CP.

5.2.5 Results

To determine the associations between the baseline characteristics of the participants in the IMPI-1 trial and the CP outcome, logistic regression with lasso regularisation were performed separately on both the placebo balanced and imbalanced training datasets. The hyperparameter, λ, that resulted in the minimum cross-validation error was determined using 10-fold cross-validation. However, to obtain a more parsimonious model, the one-standard-error rule was applied [20] to obtain λ values of 0.015 and 0.041 for the models built using the balanced and imbalanced training datasets respectively.

Notably, the use of a λ value of 0.041 for the imbalanced training dataset resulted in the elimination of all the predictors in the logistic model. Hence, for the imbalanced training dataset, the λ value of 0.019. which resulted in the lowest cross-validation error, was chosen.

Importantly, while the baseline predictor of atrial fibrillation on the electrocardiogram was not eliminated from the logistic regression model with lasso regularisation built using the balanced training dataset, this variable was manually removed from the final model as the odds ratio (OR) was very small (8.79×10^{-8}) and estimated with a large standard error (487.35). This was because none of the participants who developed CP had atrial fibrillation on the electrocardiogram which made this variable a linear separator.

Lasso regression was applied for feature selection. However since lasso regression is not used to obtain standard errors and confidence intervals [76], the estimates in Tables 5.1 and 5.2 are obtained from refitting the models using logistic regression with the features selected from lasso regression. The model estimates provided in Tables 5.1 and 5.2 have not undergone shrinkage, a feature of lasso regression [76]. Therefore, some of the odds ratios were very large and estimated with very wide confidence intervals. This was possibly a result of some degree of linear separation of the outcome by the predictors included in the models.

Table 5.1: Multiple logistic regression model with lasso regularisation from the SMOTE-NC balanced placebo training dataset used to predict the probability of CP.

Predictor	Odds ratio	Std. Error	p-value	95% CI
(Intercept)	5×10^{-4}	1.46	1.92×10^{-7}	2×10^{-5} - 0.006
Sex: female vs male	0.11	0.29	1.74×10^{-14}	0.058 - 0.184
Country: South Africa vs other	4.38	0.31	1.85×10^{-6}	2.410 - 8.131
NYHA functional class: II vs I	23.35	0.77	3.85×10^{-5}	6.444 - 150.832
NYHA functional class: III vs I	58.63	0.79	2.86×10^{-7}	15.089 - 392.590
NYHA functional class: IV vs I	103.37	0.82	1.67×10^{-8}	24.872 - 719.733
On TB medication: yes vs no	18.05	1.22	0.02	2.416 - 401.719
Peripheral oedema: yes vs no	3.49	0.26	1.93×10^{-6}	2.099 - 5.885
Hypotension: yes vs no	0.08	1.10	0.02	0.004 - 0.468
Tachycardia: yes vs no	3.23	0.25	2.42×10^{-6}	1.993 - 5.289
Anaemia: yes vs no	0.13	0.25	3.16×10^{-16}	0.075 - 0.205
White cell count $> 10\times10^{9}$/L: yes vs no	0.05	1.07	0.01	0.003 - 0.283
Renal impairment: yes vs no	0.10	0.58	4.79×10^{-5}	0.028 - 0.276
Cardiac tamponade at presentation: yes vs no	0.47	0.26	0.003	0.279 - 0.773
Effusive-constrictive pericarditis at presentation: yes vs no	3.66	0.28	3.76×10^{-6}	2.137 - 6.439
Pulmonary infiltrates: yes vs no	0.49	0.26	0.01	0.298 - 0.814
Definite TBP: yes vs no	0.78	0.30	0.42	0.437 - 1.415

Several statistical associations between the baseline predictors and odds of CP were determined from the multiple logistic regression model trained using the balanced training dataset (Table 5.1).

Notably, there was a reduced odds of CP in female participants relative to male participants (OR = 0.11; 95% CI = 0.058 – 0.184). Similar trends in the reduction of the odds of CP were observed for the predictors hypotension (OR = 0.08 yes vs no; 95% CI = 0.004 – 0.468), anaemia (OR = 0.13 yes vs no; 95% CI = 0.075 – 0.205), white cell count (OR = 0.05 yes vs no; 95% CI = 0.003 – 0.283), renal impairment (OR = 0.10 yes vs no; 95% CI = 0.028 – 0.276) and pulmonary infiltrates (OR = 0.49 yes vs no; 95% CI 0.298 – 0.814) (Table 5.1).

Participants who had an NYHA functional class categorisation of II, III or IV had respectively increasingly larger odds of CP relative to participants who were categorised as NYHA functional class I (Table 5.1). Furthermore, an increased odds of CP was observed for the features TB medication at baseline (OR = 18.05 yes vs no; 2.416 – 401.719), peripheral oedema (OR = 3.49 yes vs no; 95% CI = 2.099 – 5.885), tachycardia (OR = 3.23 yes vs no; OR = 1.993 – 5.289) and if the participants presented with effusive-constrictive pericarditis at baseline (OR = 3.66 yes vs no; 95% CI = 2.137 – 6.439) (Table 5.1).

Remarkably, far fewer predictors were selected for the logistic model with lasso regularisation trained using the imbalanced training dataset (Table 5.2). This was possibly due to the small proportion of participants in the imbalanced training dataset who developed CP (n = 38). The three predictors that were commonly selected in the logistic regression models built using the different training datasets, namely (1) NYHA functional class, (2) tachycardia and (3) anaemia, had similar trends in their effects on the odds of CP across the two logisitc models (Table 5.2).

While the logistic regression model trained using the imbalanced training dataset included the palpable pulsus paradoxus predictor, the effect of palpable pulsus paradoxus on CP was not determined to be significant in this model (95% CI = 0.634 – 3.498) (Table 5.2).

Table 5.2: Multiple logistic regression model with lasso regularisation from the imbalanced placebo training dataset used to predict the probability of CP.

Predictor	Odds ratio	Std. Error	p-value	95% CI
(Intercept)	0.03	0.75	1.28×10^{-6}	0.004 - 0.090
NYHA functional class: II vs I	2.65	0.77	0.20	0.721 - 17.130
NYHA functional class: III vs I	4.94	0.80	0.05	1.224 - 33.353
NYHA functional class: IV vs I	7.57	0.88	0.02	1.521 - 56.265
Palpable pulsus paradoxus: yes vs no	1.51	0.43	0.34	0.634 - 3.498
Tachycardia: yes vs no	1.95	0.39	0.09	0.932 - 4.337
Anaemia: yes vs no	0.31	0.38	0.002	0.141 - 0.637

Although logistic regression is easy to implement and the output provides an intuitive understanding of the relationship between the outcome and the risk factors, it is constrained by the assumptions of linearity between the odds of the outcome and the risk factors and that the risk factors are uncorrelated. Therefore non-parametric tree-based methods can be a useful alternative to the logistic model for the classification of a medical outcome. Several tree-based supervised learning methods are discussed and applied to the IMPI-1 data in the following sections.

5.3 Classification trees

This section discusses the concepts and notation of decision tree models for classification adapted from James et al. [20]. Classification trees fall into the broader class of a non-parametric machine learning method called decision trees. Decision trees are graph-like, top-down tree-like structures (Figure 5.1) that provide a set of rules for determining an outcome based on a set of risk factors [77].

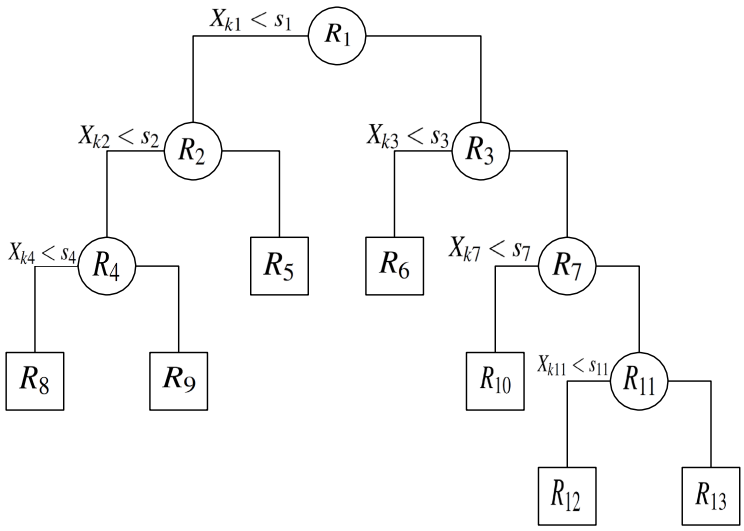

Figure 5.1: Example of a classification tree. Non-terminal nodes are represented by circles and terminal nodes or leaves are represented by squares. The lines connecting the non-terminal and terminal nodes represent the branches.

A decision tree is comprised of nodes, branches and leaves (Figure 5.1). Initially, all the observations are grouped into a single node called the root node at the top of the tree (Figure 5.1). A set of decisions are used to branch the root node into further non-terminal nodes which partition the feature space of the data into distinct subsets. The terminal nodes of the tree are called leaves whereby no further splits occur (Figure 5.1).

The tree classifies each new observation according to the majority outcome class for the observations in each leaf of the tree. The probability of an observation being classified in a specific class can be determined using the proportion of observations in the leaf that are of the same class as the observation of interest out of the total number of observations in the leaf.

The Classification and Regression Trees (CART) algorithm [77] is a popular choice used to build classification trees. CART works by recursive binary splitting at each non-terminal node to find the optimal split in the tree that increases the homogeneity or reduces the impurity of the samples of observations in the subsequent node in the tree. The algorithm ideally aims to produce trees with leaves that would contain only a single class of outcomes for all the observations in a leaf. The

two main parts of the CART algorithm namely (1) growing a classification tree and (2) pruning a classification tree are discussed below.

5.3.1 Growing a classification tree

If Y is a binary outcome, $Y \in \{0,1\}$, then the risk factors are divided using recursive binary splitting into J distinct non-overlapping regions, $R_1, \cdots, R_J$, until a certain stopping criterion is met (Figure 5.1). An initial risk factor X_k is selected with a division, s, in risk factor X_k that partitions the observations into two regions: $R_1 = X|X_k < s$ and $R_2 = X|X_k \geq s$ (Figure 5.1). This partitioning procedure is applied to every non-terminal node in the tree to produce leaves that are homogenous in terms of the outcome classes of the observations in each leaf. However, we need to decide on which predictors to grow the tree and how to partition the training sample based on these predictors.

A common method to construct the classification tree is using a measure of deviance whereby the classification tree is viewed as a probability model. For the construction of a classification tree, using deviance as the splitting criterion, let y_i denote the set of categorical outcomes in leaf node j. y_i is a random sample of size n_j from the multinomial distribution:

$$p(y_j) = \begin{pmatrix} n_j \\ n_{j1} \cdots n_{jK} \end{pmatrix} \prod_{k=1}^{K} p_{jk}^{n_{jk}},$$

where p_{jk} is the probability of leaf node j being classified as outcome k.

The likelihood over all J leaf nodes is:

$$l = \prod_{j=1}^{J} p(y_j) \propto \prod_{j=1}^{J} \prod_{k=1}^{K} p_{jk}^{n_{jk}}.$$

We define deviance as:

$$\begin{aligned} D &= -2 \log l \\ &= -2 \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} \log p_{jk}. \end{aligned}$$

For each split in the classification tree, the split that results in the smallest deviance is selected.

The Gini index, G, is another method of constructing the classification tree. For all the leaf nodes, $j = 1, \cdots, J$, in the tree, the Gini index measures the impurity of node G_j:

$$G = \sum_{j=1}^{J} G_j, \qquad \text{where } G_j = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk}),$$

where $\hat{p}_{jk}$ is the proportion of observations in the outcome category k within each leaf node j. The split is chosen during the tree growth that results in the greatest decrease in the Gini index.

5.3.2 Pruning a classification tree

Ideally, we would grow a classification tree that results in leaves containing only a single class of outcomes however, a very large tree would need to be built to perfectly classify each training observation based on the set of risk factors. Growing such a large tree would ultimately lead to the model overfitting the training data and a lack of generalisability of the model to new observations. Therefore it is useful to grow a very big tree initially on the training data, but then remove certain nodes from the tree in a process called pruning.

To prune the tree we specify a criterion on which to remove the nodes in the tree known as the cost complexity criterion. To prune a classification tree, T, the cost complexity criterion is:

$$C_\alpha(T) = C(T) + \alpha|T|,$$

where $C(T)$ is the classification error rate of tree T and α is the complexity hyperparameter which is a penalty enforced for every leaf added to the tree. The value of α which is the best compromise between the bias and variance of the classification tree can be determined using k-fold cross-validation.

5.3.3 Results

A classification tree, using the reduction in deviance as the splitting criterion, was applied separately to both the balanced and imbalanced training datasets and cost complexity pruning with 10-fold cross-validation was used to determine the number of terminal nodes that resulted in the smallest cross-validation error. For the balanced training dataset, a large tree consisting of 66 terminal nodes was grown and an α value of 3, determined by 10-fold cross-validation, was used to prune the classification tree to 15 terminal nodes (Figure 5.2).

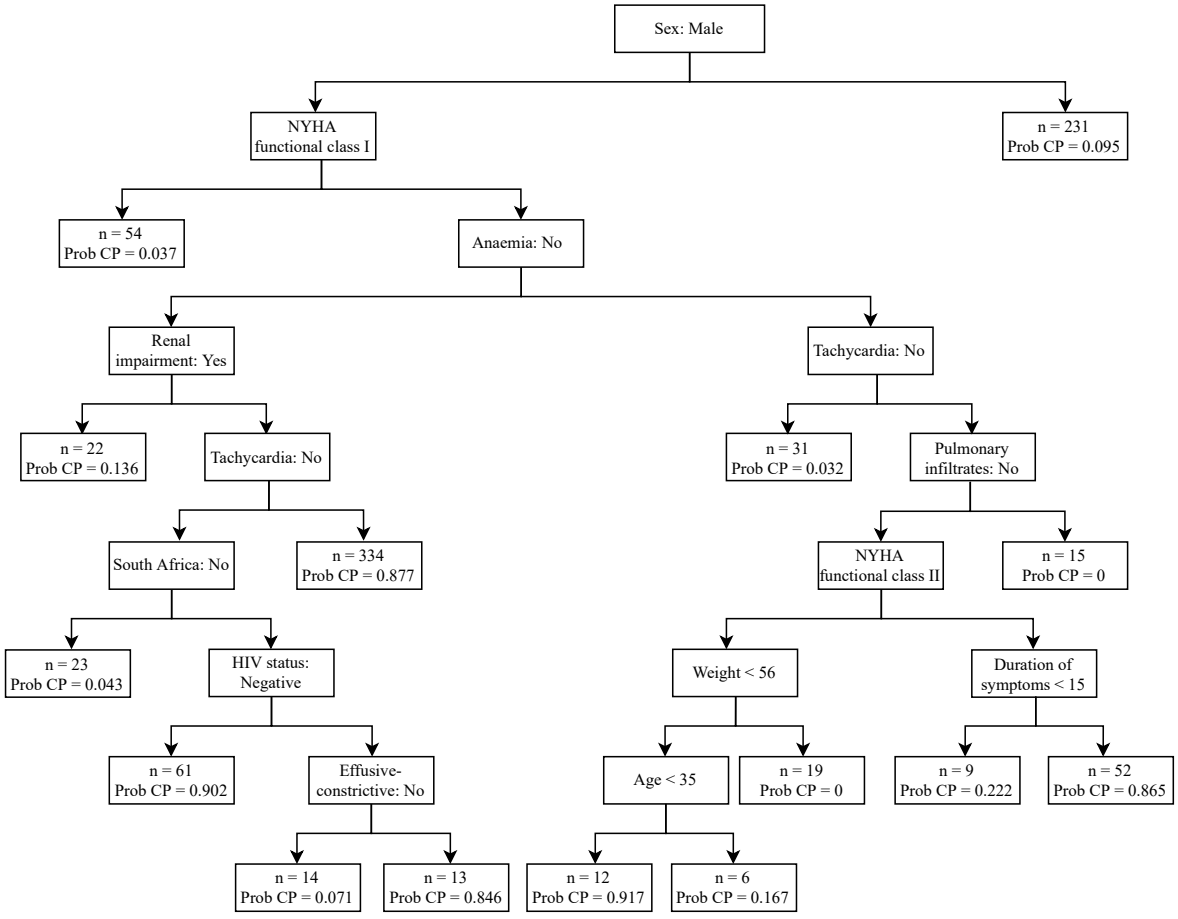

Figure 5.2: Classification tree trained using the SMOTE-NC balanced placebo training dataset determined by cost complexity pruning. Each internal node is labelled with the condition of the predictor on which that feature was split. Branches to the left of each condition reflect the child or terminal node when that condition is true while branches to the right of each condition reflect the child or terminal node when that condition is false. Terminal nodes are labelled with the number of terminal-node observations and the predicted probability of CP.

Despite all the predictors being considered for the tree construction, only 12 features were included in the final classification tree built using the balanced training dataset (Figure 5.2). Based on their order in tree, the sex and NYHA functional class were considered the first and second most important predictors for the CP outcome respectively. The advantage of classification trees is that they allow for an intuitive understanding of how the predictors are associated with the outcome. For example from the tree built using the balanced training dataset, if a participant is male and has an NYHA functional class I, their expected probability of CP is 0.037 (Figure 5.2).

For the imbalanced training dataset, a large tree consisting of 31 terminal nodes was grown. The α value of 0.71, determined by 10-fold cross-validation, resulted in the tree with the lowest deviance. However, this tree had only one terminal node which would have resulted in all the observations being classified as the majority class of no CP. Therefore the α value of 0.5, which resulted in the second greatest reduction in deviance, was selected to prune the classification tree to consist of eight terminal nodes (Figure 5.3).

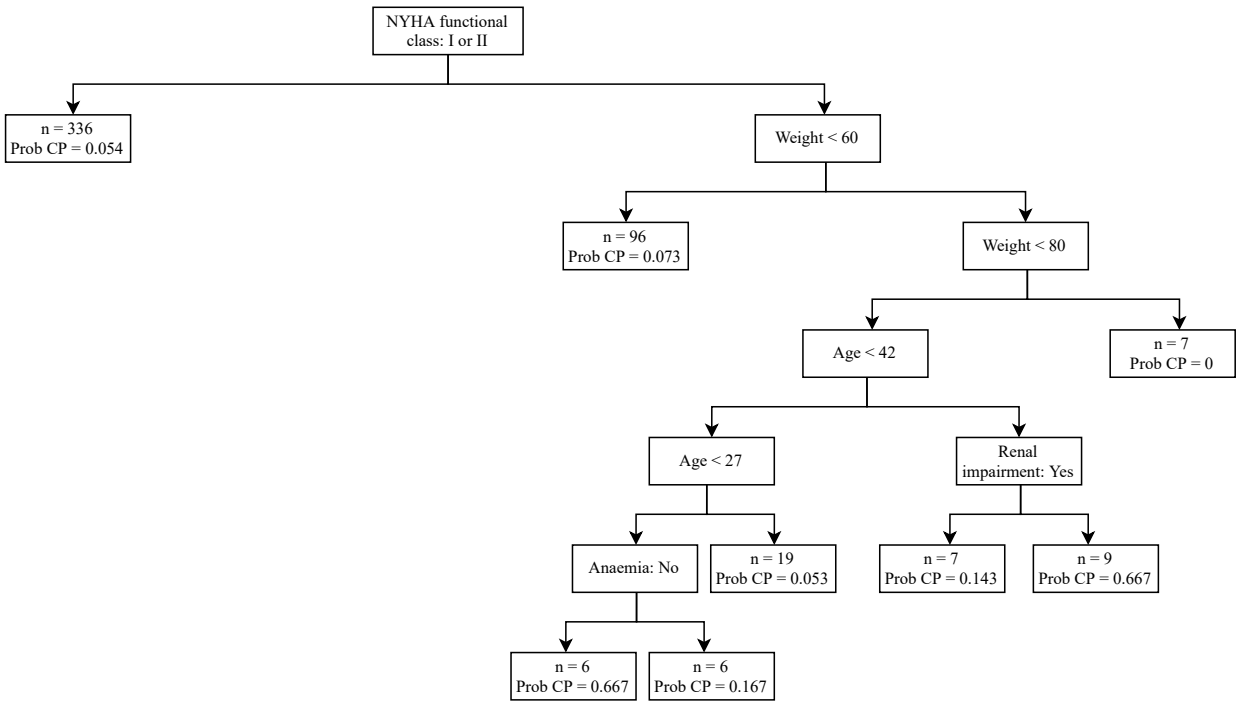

Figure 5.3: Classification tree trained using imbalanced placebo training dataset determined by cost complexity pruning. Each internal node is labelled with the condition of the predictor on which that feature was split. Branches to the left of each condition reflect the child or terminal node when that condition is true while branches to the right of each condition reflect the child or terminal node when that condition is false. Terminal nodes are labelled with the number of terminal-node observations and the predicted probability of CP.

Of all the features considered in building the tree, only five were useful in predicting CP (Figure 5.3). Contrastingly to the tree trained using the balanced training dataset, based on its order in the tree, the NYHA functional class was the most important feature in the tree trained using the imbalanced dataset. The age and weight predictors were also included in the final tree which was interesting as these predictors had not been selected in any of the previously discussed models.

Classification trees are incredibly useful in predictive modelling for clinical outcomes as they are easy to understand and visualised in a single graph of decision rules that can be used to classify new observations. Additionally, no specification of the relationship between the risk factors and the outcome is required which allows the tree to not be constrained by the assumptions of linearity

as in logistic regression. This allows the tree to learn and identify complex interactions that exist in the data. However, classification trees have a high sampling variability which often leads to poorer predictive performance than more complex tree-based methods such as random forests and boosted trees [20] which will be discussed in the following two sections.

5.4 Random forests

While classification trees are easily visualised and have a simple interpretation, a single classification tree can be highly variable from one training sample to another which can result in a low predictive performance when the tree is applied to new data [20]. Therefore, instead of training a single classification tree to predict an outcome, multiple trees, known as a random forest, can be trained to collectively predict an outcome with greater accuracy than a single tree. This section will discuss the concepts of the bootstrap aggregation algorithm and its specific application to random forests with concepts and notation mainly adapted from James et al. [20] unless otherwise specified.

5.4.1 Bootstrap aggregation

Bootstrap aggregation or bagging is a method to overcome the high sampling variability of a single classification tree by training several classification trees that are built from training datasets containing observations that are sampled independently from the original training dataset [78]. A common method to try reduce the sampling variability of a single classification tree is to combine the results of multiple classification trees which can lead to better predictive performance when the model is applied to new data.

The bagging procedure can be applied to classification trees to build what is known as a random forest [25]. Suppose there are B independently bootstrapped training datasets, a classification tree is built for each dataset. The outcome, based on a given set of features, is predicted for each classification tree. The average of the predictions for the outcomes across the B trees is calculated as:

$$\hat{y}_{ave} = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_b,$$

where $\hat{y}_b$ is the predicted outcome, y, for the b^{th} tree.

The B trees, collectively known as the random forest, are built using bootstrap sampling of observations with replacement from the original training dataset to create smaller B datasets of equal size. Therefore for each new observation, B predictions are made. For a binary outcome, the classes of new observations are determined by the most frequently occurring class among the B predictions made for the classification trees.

Typically, two-thirds of the training data are randomly sampled with replacement to build the random forest. Therefore approximately one-third of the data on average is not selected to build the classification trees in the random forest. These unused observations are known as out-of-bag (OOB) observations. Because the OOB observations were not involved in the training of the random forest, they can be used to calculate the OOB classification error. The OOB classification error assesses the performance of the random forest on unseen data which is comparable to the cross-validation error described previously.

While bagging is a useful method of reducing the sampling variability of the model, if a specific predictor describes a large amount of the variation in the outcome, it is likely that the majority of classification trees grown in the bagging procedure will select this predictor by which to partition the data. The use of the same predictors over and over again to build many classification trees results in highly correlated trees.

5.4.2 Decorrelating bagged trees

The random forest algorithm reduces the correlation between the trees by restricting the number of predictors from which the splits in the tree can consider. For every tree in the random forest, a random sample of m predictors is considered such that $m < p$ where p is the number of risk factors in the full training dataset. This results in the trees in the random forest being less correlated as on

average $\frac{p-m}{p}$ of the splits in the trees will not consider a dominant predictor.

Several tuning parameters can be considered when building the random forest including the number of features on which the node splitting can consider, the number of classification trees to build and the sizes of the trees. The OOB classification error can be used to select the best combination of tuning parameters to use for training the final model.

Although random forest models do not provide an easy visual interpretation of the model output like a single classification tree, we can derive a measure of the importance of each risk factor when training the random forest model. The importance of the risk factors in the model is assessed by how much the splitting criterion, either deviance or Gini index, is improved for each tree when considering a certain predictor on which to split the tree. The average improvement in the splitting criterion for a specific risk factor is then averaged for the B trees in the forest. Large variable importance for a specific risk factor indicates the importance of that risk factor in training the model.

5.4.3 Results

To reduce the correlation between the trees, a grid search of hyperparameters was conducted separately for each training dataset to determine the optimal number of predictors considered for each split and the minimum terminal node size of the random forest that resulted in the smallest OOB misclassification error. For each training dataset, the grid search trained 10,000 trees and considered a combination of 1-13 predictors for every tree split and 1-5 terminal nodes.

The combination of hyperparameters that resulted in the smallest OOB error of 0.0603 in the balanced training dataset was seven predictors considered at each split in each tree and three terminal nodes (Appendix B, Figure B.1). However, for the imbalanced training dataset, the lowest OOB error from the grid search of hyperparameters was 0.0782 and was achieved with one variable considered at each tree split and one terminal node in each tree (Appendix B, Figure B.1).

Similar to the CART model, the two most important features, determined by the average reduction in the Gini index, in the random forest trained using the balanced training dataset were sex and NYHA functional class (Figure 5.4). However, contrastingly to the CART model trained using the imbalanced training dataset, the random forest trained using this dataset determined that the age, weight and duration of TB/TBP symptoms were the most important features in the model and not the NYHA functional class predictor (Figure 5.4).

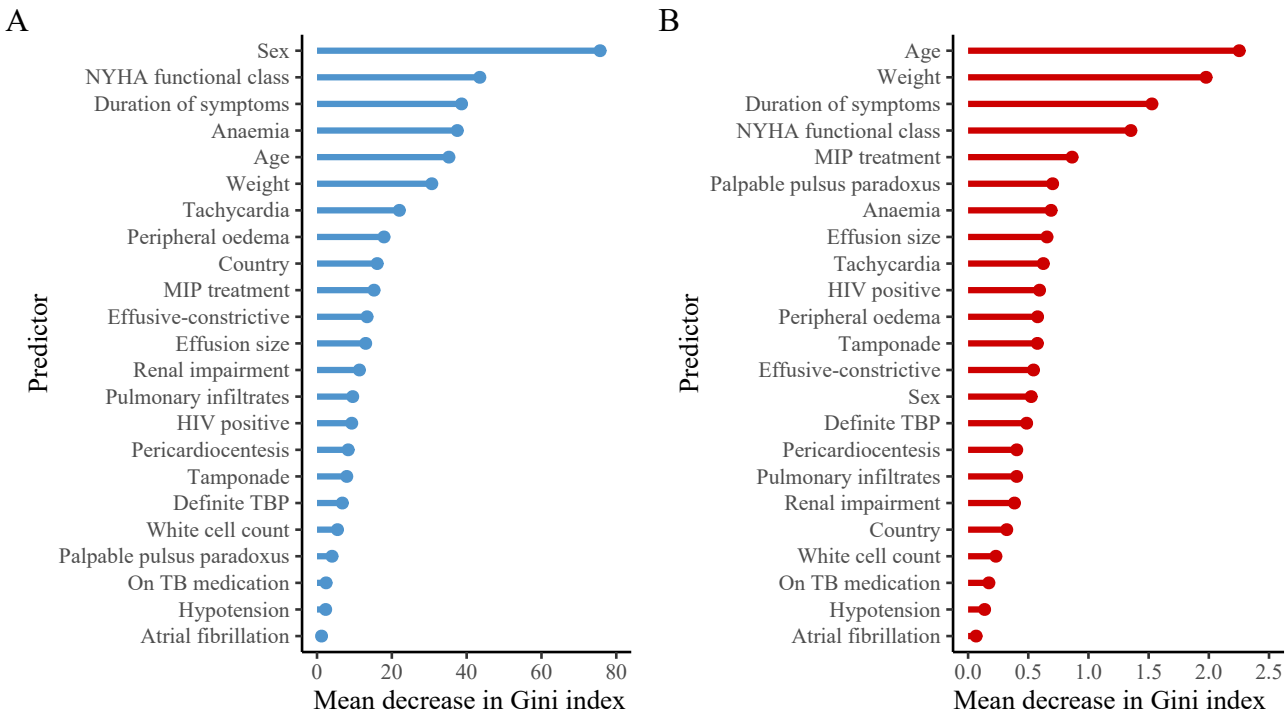

Figure 5.4: Comparison of variable importance between the random forest models built using either the SMOTE-NC balanced placebo training dataset or imbalanced placebo training dataset. A: SMOTE-NC balanced placebo training dataset, B: imbalanced placebo training dataset. Predictors are ranked according to their mean decrease in Gini index when used for splitting.

Interestingly, the random forests trained using either dataset both determined that the features TB medication, hypotension and atrial fibrillation were the least important in training the random forest (Figure 5.4). However, unlike the previously discussed logistic regression and CART models, the random forest models do not give a clear indication of how these features were associated with the CP outcome.

Notably, there was an obvious scale difference between the mean decrease in Gini index in the random forest models trained on the different datasets (Figure 5.4). The predictors in the model built using the balanced training dataset contributed to much larger decreases in the Gini index compared to the model built using the imbalanced training dataset (Figure 5.4).

Random forests are an example of ensemble learning which aims to combine several simple, weaker models into a single, stronger model used for prediction at the cost of model interpretability. Boosted trees, which are discussed in the following section, are another example of ensemble tree-based methods that can be used instead of a single classification tree to enhance a classifier's predictive performance.

5.5 Boosted trees

Boosting, similar to bagging combines many weak models to produce a single more powerful model in terms of predictive ability. A boosted classification model can produce very accurate predictions by combining the decisions made from many weak classifiers [20, 79]. This section discusses the concepts and notation of the boosting algorithm and its application to classification trees adapted from James et al. [20] and Elith et al. [79] unless otherwise specified.

5.5.1 Introduction to boosting

In general, boosting reapplies a supposedly weak classification model to data that has been reweighted [80]. For every iteration of the model's algorithm, observations for which the outcome was incorrectly predicted are given greater weight than the observations for which the outcome was correctly predicted. A final prediction is then made by combining the sequence of weighted weak classifiers. This method of ensemble learning allows the model to become a more accurate classifier by learning from the mistakes made in a previous iteration of the model.

While random forests and boosting are both examples of ensemble learning, the boosting ensemble method is different to random forests because the trees that are built using the boosting algorithm are all dependent compared to a random forest model whereby the trees are grown independently.

5.5.2 Boosted trees

The boosting algorithm is applied to classification trees by growing small trees with d splits sequentially. Let r^b be the residuals after fitting b trees and $\hat{r}^b$ be the predicted values of the outcomes in tree b. The boosted, $b = 1, \cdots, B$, trees with d splits are built:

$$r^b = r^{b-1} - \lambda \hat{r}^b,$$

where λ is the learning rate of the model. The trees that are first grown capture the greatest variation in the data and contribute the most to the reduction in the size of the residuals. The more trees that are built, the smaller their impact will be on the reduction in residual size. We can control the pace of the learning process by imposing a tuning parameter, λ, that reduces the contribution of each subsequent tree to the final model. By controlling this learning rate, we can allow the model to learn slower which contributes to a better predictive performance of the final model.

The final predictions of the boosted classification tree model for B trees are:

$$\begin{aligned} \hat{y} &= y - r^b \\ &= \sum_{b=1}^{B} \lambda \hat{r}^{(b)}, \end{aligned}$$

where $\hat{y}$ are the model predictions for y observations.

Several tuning parameters including the number of trees, B, the number of splits in the trees, d, and the learning rate, λ, can be determined using k-fold cross-validation.

While boosted classification trees use ensemble learning to improve the classification accuracy of the model, they are not easily interpreted, nor do they provide a simple visual output of the model like a single classification tree. However, similar to random forests, the importance of predictors in the boosted model can be assessed by a measure of relative influence which is determined by

how much that variable improves the splitting criterion on average when that predictor is used to split the tree.

5.5.3 Results

To determine the combination of hyperparameters used to train the final boosted model, 10-fold cross-validation was used in a grid search of the following hyperparameter combinations for both the balanced and imbalanced training datasets:

1. Number of trees = 1000, 2000, 3000, 4000, 7000, 10,000.
2. Number of tree splits = 1, 2, 3, 4, 5.
3. Learning rate = 0.01, 0.005, 0.001.

For the balanced training dataset, it was determined that 7000 trees, with five splits in each tree and grown at a learning rate of 0.01 resulted in the largest cross-validation reduction in deviance (Appendix B, Figure B.2). However, for the imbalanced training dataset, similar to the random forest model, the smallest cross-validation deviance was achieved with 1000 trees each with only one split in each tree and grown at a learning rate of 0.001 (Appendix B, Figure B.2). Additionally, while substantially increasing the number of trees in the model trained using the balanced dataset corresponded to a large reduction in the cross-validation deviance, only a small reduction in the cross-validation deviance was observed with an increase in the number of trees in the model trained using the imbalanced dataset (Appendix B, Figure B.2).

In agreement with the previous tree-based methods, sex, duration of TB/TBP symptoms and NYHA functional class were some of the most important features used to train the boosted model using the balanced training dataset, while TB medication, atrial fibrillation on the electrocardiogram and hypotension were some of the least important predictors (Figure 5.5). Importantly, every feature included in the boosted tree model trained using the balanced training dataset had some influence on the prediction of the CP outcome (Figure 5.5).

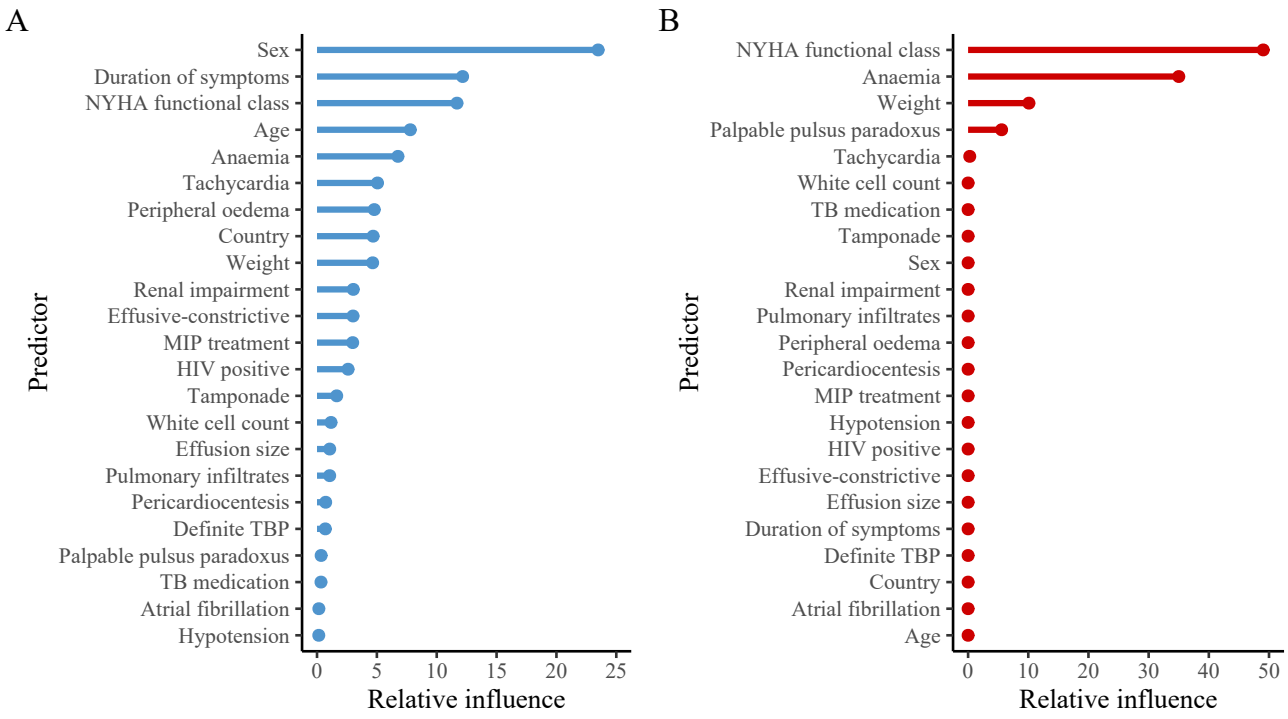

Figure 5.5: Comparison of variable importance between the boosted tree models trained using either the SMOTE-NC balanced placebo training dataset or imbalanced placebo training dataset. A: SMOTE-NC balanced placebo training dataset, B: imbalanced placebo training dataset. Predictors are ranked according to their mean decrease in deviance when used for splitting.

Contrastingly, the only predictors that were influential in the boosted model trained using the imbalanced training dataset were NYHA functional class, anaemia, weight, palpable pulsus paradoxus and tachycardia (Figure 5.5). This was surprising as both the random forest and CART models trained using this dataset determined that participant age and duration of TB/TBP symptoms were important features in predicting the CP outcome (Figures 5.3, 5.4).

While ensemble tree-based classification models are neither easily interpreted nor visualised, they are potentially capable of achieving greater predictive performance than more simple classification models such as logistic regression or CART. Support vector machines, which are another powerful classification technique used to predict clinical outcomes, will be discussed in the next section.

5.6 Support vector machines

Support vector machines (SVMs) are extensions of the support vector classifier which aims to find a hyperplane in a high dimensional feature space that separates classes of an outcome such that the observations from the different classes lie on opposite sides of the hyperplane [81]. This section discusses the concepts and notation of SVMs adapted from James et al. [20] and Friedman et al. [74].

In a SVM model, the distances between observations and the proposed hyperplane are calculated and are used to determine the margin. The margin is defined as the minimum distance between the observations of the different classes. The SVM model attempts to find a hyperplane that maximises the margin such that the distance between the hyperplane and the observations from the different classes is maximised.

Observations that lie directly on the constructed margin that separates the two classes or on the wrong side of the margin for a particular class are known as support vectors. The support vectors allow the model to misclassify some observations to reduce the potential overfitting of the model to the training data and allows for increased model generalisability.

5.6.1 The support vector machine model

The SVM model aims to determine a hyperplane that best separates an outcome, y, where an observation, i, belongs to a binary outcome class $y_i \in \{-1,+1\}$. The hyperplane is described by a vector $\boldsymbol{w}$ and the space between the two outcome classes is defined as the margin $\frac{1}{\|\boldsymbol{w}\|}$ [82]. The SVM function is defined as:

$$f(\boldsymbol{w},\boldsymbol{x},b) = sgn(<\boldsymbol{w},\boldsymbol{x}> +b),$$

where sgn is a signum function that determines the classification of observations as either $+1$ or -1 and $<\boldsymbol{w},\boldsymbol{x}>$ is the dot product of the hyperplane, $\boldsymbol{w}$, with an intercept b, and the vector of locations of the observations, $\boldsymbol{x}$.

The SVM model aims to determine the hyperplane $\boldsymbol{w}$ which maximises the distance between the observations of the different classes and the hyperplane to classify as many training observations correctly as possible:

$$\max \frac{1}{\| \boldsymbol{w} \|}.$$

This is equivalent to minimising the norm of the hyperplane vector, $\boldsymbol{w}$, for all points in the feature space $\boldsymbol{x_i}, y_i$:

$$\min \frac{1}{2} \| \boldsymbol{w} \|^2 + \sum_i (C)(\xi_i),$$

subject to $y_i(< \boldsymbol{w}, \boldsymbol{x} > +b) \geq 1 - \xi_i$ where ξ_i is the margin of error for the observations that are incorrectly classified and C is the cost-complexity parameter that determines the margin width. Since this is a constrained optimisation problem, the solution requires constructing a dual problem where a Lagrange multiplier α_i is introduced and is associated with every constraint in the primary optimisation problem:

$$L(w, b, \alpha) = \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_i \alpha_i (y_i(< \boldsymbol{w}, \boldsymbol{x_i} > +b) - 1 + \xi_i).$$

The optimisation is then transformed such that $0 \leq \alpha_i \leq C$:

$$\max \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j < x_i, x_j > .$$

A kernel function $K(x_i, x_j)$ can be substituted for the dot product $< x_i, x_j >$ which allows for a solution to a classification problem where the boundary between the classes is nonlinear.

5.6.2 Kernel functions

Kernel functions provide a convenient solution to determine a nonlinear boundary to separate two classes. Kernel functions allow for the data to be projected into a higher dimensional feature space

where the data become linearly separable.

While many kernel functions exist, four of the basic kernel functions include the linear kernel function:

$$K(x_i, x_j) = x_i^T x_j + c,$$

the polynomial kernel function:

$$K(x_i, x_j) = (\sigma x_i^T x_j + c)^d,$$

the radial basis kernel function:

$$K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right),$$

and the sigmoid kernel function:

$$K(x_i, x_j) = \tanh(\sigma x_i^T x_j + c),$$

where c is a constant, σ is a kernel parameter and d is the degree of the polynomial in the polynomial kernel function. These tuning parameters can be determined using k-fold cross-validation.

5.6.3 Results

To determine the Cost (C), kernel function, and kernel-specific tuning parameters that resulted in the minimum model cross-validation error, a grid search using 10-fold cross-validation was performed separately on both the balanced and imbalanced training datasets using the following hyperparameter combinations:

1. Linear kernel with C = 0.0001, 0.001, 0.01, 0.05, 1, 5, 10, 15,

2. Polynomial kernel with C = 0.0001, 0.001, 0.01, 0.05, 1, 5, 10.
   σ = 0.005, 0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10,
   d = 2, 3, 4, 5.

3. Radial basis kernel with C = 0.0001, 0.001, 0.01, 0.05, 1, 5, 10,
   σ = 0.005, 0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10.

4. Sigmoid kernel with C = 0.0001, 0.001, 0.01, 0.05, 1, 5, 10,
   σ = 0.005, 0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10.

For the balanced placebo training dataset, the SVM containing a radial basis kernel with $C = 1$ and $\sigma = 0.5$ resulted in the lowest cross-validation classification error of 0.0535 (Appendix B, Table B.1). Interestingly, the grid search of hyperparameters performed on the imbalanced training dataset resulted in the cross-validation classification error for all three different kernel functions equalling 0.0783 (Appendix B, Table B.1). Therefore an SVM model containing a linear kernel function with $C = 0.0001$ was fit to the imbalanced placebo training dataset.

As a measure of feature importance for the SVM models trained using either the balanced or imbalanced placebo training dataset, the predictor model weights were determined and ranked according to their absolute value (Figure 5.6).

The variables sex and anaemia were of the greatest influence in the SVM model trained using the balanced training dataset (Figure 5.6). However, the NYHA functional class predictor did not have a large absolute weight value in the SVM model trained using the balanced training dataset which was surprising considering that the NYHA functional class feature was relatively important in most of the other models (Figure 5.6).

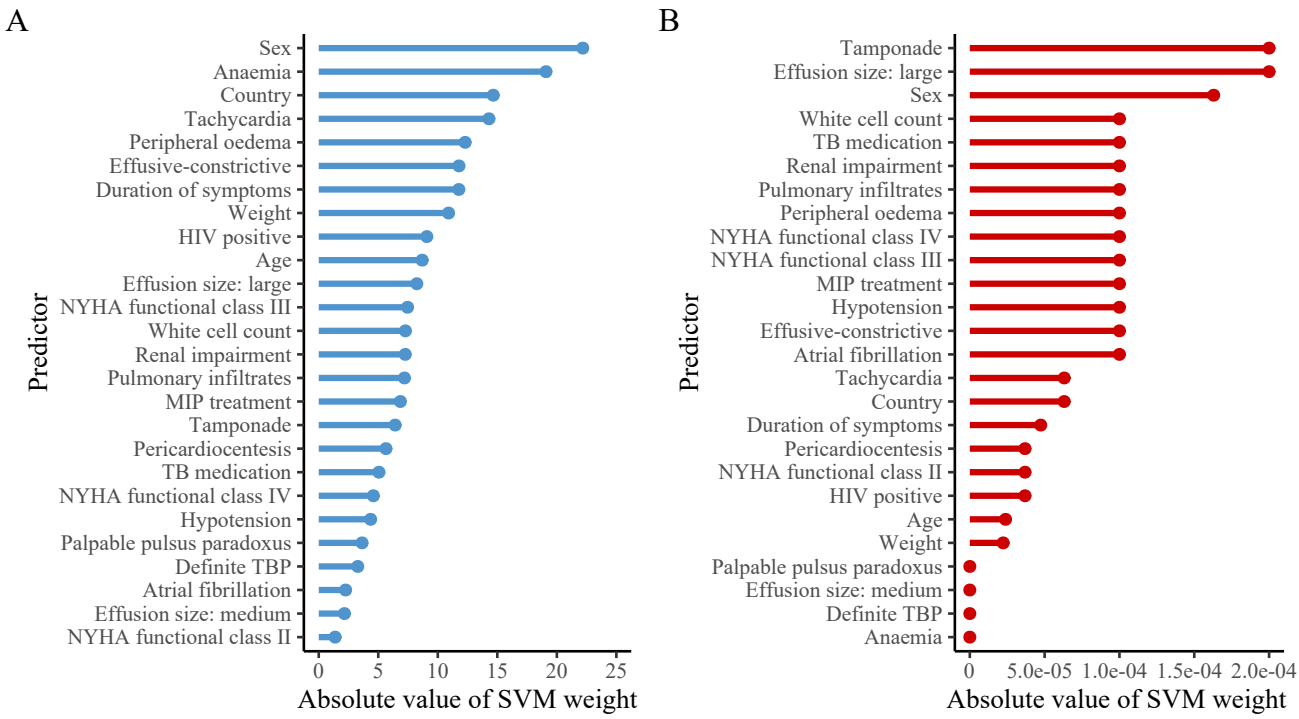

Figure 5.6: Comparison of absolute values of the weights obtained for each predictor of CP in the SVM models built using either the SMOTE-NC balanced placebo training dataset or imbalanced placebo training dataset. The features are ranked from largest to smallest absolute weight value as an indication of feature importance. A: SMOTE-NC balanced placebo training dataset, B: imbalanced placebo training dataset.

While the sex predictor was also a highly weighted feature in the SVM model built using the imbalanced placebo training dataset, the tamponade and large effusion size features were the highest weighted predictors in the model (Figure 5.6). Notably, these variables were not considered to have a large relative influence in any of the other models built using this dataset. Surprisingly, the age and weight features that were considered to be relatively important in the tree-based methods trained using the imbalanced training dataset were ranked among some of the least important predictors in the SVM model (Figure 5.6).

While SVM models, similar to random forests and boosted trees, allow for increased model flexibility and potentially increased predictive power compared to simpler classification algorithms like logistic regression or classification trees [54], they provide little interpretability or visual

understanding of the specific relationships between the risk factors and the outcome. Similarly, artificial neural network models, which are discussed in the following section, provide another modelling approach to overcome the constraints and limitations posed by simpler supervised learning classification methods.

5.7 Artificial neural networks

Similar to how a biological brain detects sensory input from the surroundings and responds, the artificial neural network model is designed to process a set of input features and respond with an output [83]. Typically, an artificial neural network is comprised of three layers commonly called the input, hidden and output layers [84]. Within each layer, the network is comprised of neurons that process the training data and weighted connections which connect the neurons from the previous layer to the neurons in the next layer.

Artificial neural networks are commonly called feedforward networks because the information flows sequentially through the model from the input to the hidden to the output layers [84]. This section will outline the concepts and notation of artificial neural networks adapted from Friedman et al. [74] and Nielson [85].

5.7.1 The artificial neural network model

A feedforward neural network is made up of a series of linearly weighted connections which combines the input data and an activation function that determines the output of the neurons to the next layer in the network:

$$z_j^{(l)} = g\left(\sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)} a_i^{(l-1)}\right),$$

where:

$l =$ the layer of the network with $l = 0, \cdots, L$,

$d^l =$ the number of neurons in layer l,

$w_{ij}^l =$ weight from the i^{th} neuron in the $(l-1)^{th}$ layer to the j^{th} neuron in the l^{th} layer,

z_j^l = the sum of the weighted input of the of the j^{th} neuron in the l^{th} layer,
a_j^l = the output of the j^{th} neuron in the l^{th} layer,
$g(\cdot)$ = activation function applied.

A simple example of an artificial neural network is shown in Figure 5.7.

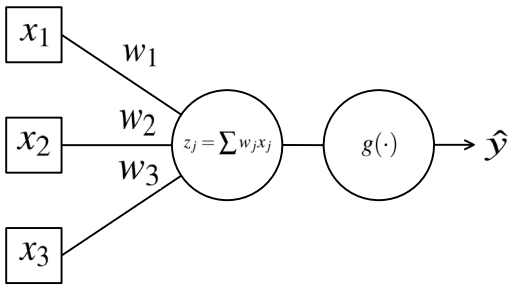

Figure 5.7: Example of an artificial neural network. The neural network is used to predict a single output, y, from a set of three input features, x_1, x_2, x_3.

Initially, random values are assigned to the weights in the network (Figure 5.7). The training input data are presented to the network where the data are linearly combined with the weights (Figure 5.7). A specified activation function is applied to the linear combination of input data in a procedure known as forward propagation (Figure 5.7). The model classifies the observations based on the input data and the values of the weights and the error of the model is computed. The weights are then updated according to the model error in a procedure known as backpropagation.

5.7.2 Backpropagation: updating the weights

In a classification setting, for observations that are classified correctly by the model, the weights are not changed. However, if an observation is misclassified then the error, $J(w)$, is determined:

$$J(w) = -\frac{1}{2}\sum_{i=1}^{n}\left[y^{(i)}\log\left(\hat{y}^{(i)}\right) + (1 - y^{(i)}\log\left(1-\hat{y}^{(i)}\right)\right].$$

The model aims to minimise the error, $J(w)$, with respect to the weights, w. The weights are updated in a process known as backpropagation:

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(w),$$

where α is the learning rate, $0 < \alpha \leq 1$, that determines the amount of change that is allowed to occur. The weights are updated until all the training observations are correctly classified or a specified number of iterations has passed. By specifying a certain number of iterations, we can control the complexity of the model to reduce the model overfitting the training data. The learning rate and the number of iterations can be determined using k-fold cross-validation.

5.7.3 Activation functions

Even though the artificial neural network is comprised of a set of neurons connected by a linear combination of weights, artificial neural networks have great flexibility to model nonlinear relationships between the input features and the outcome of interest using activation functions.

Activation functions transform the linear combination of input features and weights from the output of each neuron in the network to model nonlinear relationships between the input features and the outcome. Several different activation functions exist with popular ones including the hyperbolic tangent function:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

the rectified linear unit function:

$$r(x) = \max(0, x) = \begin{cases} x & x > 0 \\ 0 & \text{elsewhere} \end{cases},$$

and the maxout activation function:

$$h(x) = \max(w_1^T x + b_1, w_2^T x + b_2, \cdots, w_n^T x + b_n).$$

5.7.4 Results

To determine the neural network hyperparameters and activation function that resulted in the minimum model cross-validation error, 10-fold cross-validation was performed separately for both the balanced and imbalanced training datasets using a grid search of the following tuning parameter combinations:

1. Activation functions = hyperbolic tangent, rectified linear unit, maxout,
2. Hidden layers = 1 hidden layer with 2 - 10 neurons, 2 hidden layers each with 1 neuron or 2 hidden layers each with 2 neurons,
3. L1 regularisation = 0.01, 0.001,
4. Iterations = 1000, 5000.

The performance of the 10-fold cross-validation was evaluated using the log-loss function defined as:

$$\text{Log-loss} = -\frac{1}{N}\sum_{i=1}^{N} w_i \left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i) \right],$$

where p is the predicted probability of the outcome.

For the balanced placebo training dataset, the artificial neural network containing a single hidden layer with eight neurons, hyperbolic tangent activation function, L1 regularisation of 0.001 and run using 1000 iterations resulted in the smallest cross-validation log-loss of 0.2794. For the imbalanced placebo training dataset, the artificial neural network containing a single hidden layer with four neurons, hyperbolic tangent activation function, L1 regularisation of 0.01 and run using 1000 iterations resulted in the smallest cross-validation log-loss of 0.2603.

The relative importance of each variable is a measure of the standardised magnitude of the weights used in the model with a larger weight indicating a larger predictor effect in the model relative to other variables in the model [86]. As a measure of relative feature importance, the weights of each

predictor from the artificial neural network models built using either the balanced or imbalanced placebo training datasets were compared and ranked (Figure 5.8).

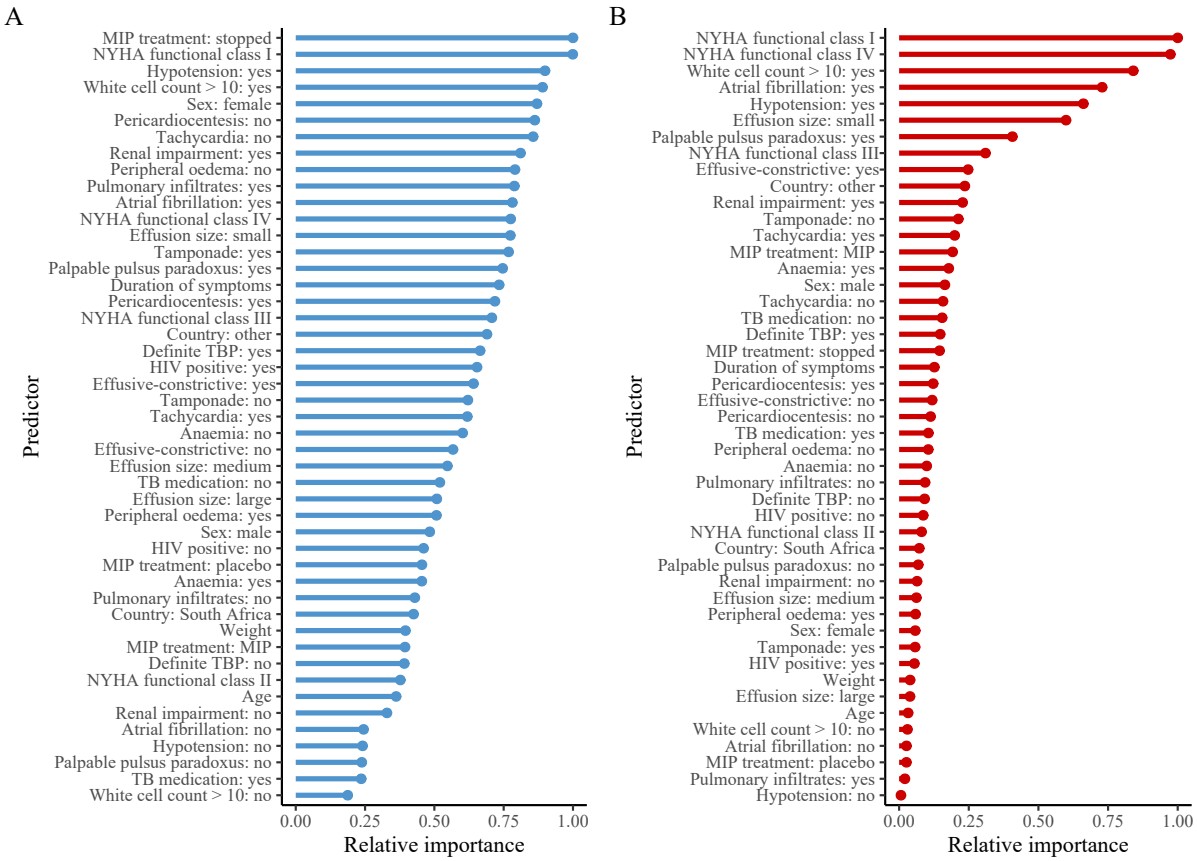

Figure 5.8: Comparison of the relative importance of predictors in the artificial neural network model built using either the SMOTE-NC balanced placebo training dataset or imbalanced placebo training dataset. The features are ranked according to their importance in predicting CP relative to the other features included in the model. A: SMOTE-NC balanced placebo training dataset, B: imbalanced placebo training dataset.

Surprisingly, the predictor categorising the participants who stopped MIP treatment was of the greatest relative importance in the artificial neural network model trained using the balanced training dataset (Figure 5.8). Additionally, the hypotension and white cell count predictors had a relatively large influence on the CP outcome in this model which was surprising considering that these features were not deemed to be influential in the tree-based models trained using the balanced placebo training dataset (Figure 5.8).

Notably, in the artificial neural network built using the imbalanced placebo training dataset, the majority of features were ranked with relatively low importance except for NYHA functional class, white cell count, atrial fibrillation on the electrocardiogram and effusion size (Figure 5.8). This was contrasted by the model trained using the balanced placebo training dataset in which most of the features had relatively substantial importance in predicting the CP outcome (Figure 5.8).

While artificial neural networks, similar to SVMs, can be very powerful in terms of predictive performance and are capable of learning complex data structures with many features that are nonlinearly related to the outcome, they are incredibly complex and do not provide a simple interpretation or visual understanding of how the features relate to the outcome. This can be especially limiting in a clinical context when attempting to understand the factors that impact disease risk. Therefore both the predictive performance and clinical relevance of a supervised learning model need to be considered when developing a risk score to predict a clinical outcome.

The previous sections of this chapter have demonstrated the use of different supervised learning classification algorithms to predict a binary outcome. While these models differ in their statistical approach and interpretation, they are all subject to the same set of performance measures that are used to determine the discriminative capacity and calibration of the model when applied to a test dataset. Common classification performance measures are discussed and applied to the models trained to predict CP in the following section.

5.8 Measures of classification model performance

This section discusses common error and accuracy measurements that are used in classification and the use of receiver operating characteristic (ROC) and precision-recall curves to assess the discriminative capacity of a classification model. This section briefly outlines the concepts and notation of measures used to assess the performance of a classification model adapted from James et al. [20] unless otherwise specified.

5.8.1 Classification error measurements

To assess the performance of a classification model in classifying the outcomes of new observations, the classification error can be measured which is determined by the proportion of incorrectly predicted outcomes:

$$\text{Error} = \frac{1}{N}\sum_{i=1}^{N}(Y_i \neq \hat{Y}_i).$$

However, in a classification setting, we are not only concerned with the overall classification error of the model but with the specific breakdown of the errors. To determine the specific breakdown of model errors we define four possible classification scenarios for a binary outcome:

- True Positive (TP) where the model correctly classifies $\hat{Y} = 1$ when the actual observed outcome $Y = 1$.
- True negative (TN) where the model correctly classifies $\hat{Y} = 0$ when the actual observed outcome $Y = 0$.
- False Positive (FP) where the model incorrectly classifies $\hat{Y} = 1$ when the actual observed outcome $Y = 0$.
- False negative (FN) where the model incorrectly classifies $\hat{Y} = 0$ when the actual observed outcome $Y = 1$.

5.8.2 Accuracy measures for a classification model

Using the number of true positives, true negatives, false positives and false negatives determined by the model, we can define the accuracy or calibration of a classification model as the number of observations for which the outcomes are classified correctly out of the total number of observations:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}.$$

The true positive rate (TPR) also known as sensitivity or recall is the probability of predicting that an observation has the outcome when the observation has the outcome:

$$\text{TPR} = \frac{TP}{TP+FN}.$$

The true negative rate (TNR) also known as specificity is the probability of predicting that an observation does not have the outcome when the observation does not have the outcome:

$$\text{TNR} = \frac{TN}{TN+FP}.$$

Positive predictive value (PPV) also known as precision is the probability that an observation has the outcome when they are predicted by the model to have the outcome:

$$\text{PPV} = \frac{TP}{TP+FP}.$$

Negative predictive value is the probability that an observation does not have the outcome when they are predicted by the model to not have the outcome:

$$\text{NPV} = \frac{TN}{TN+FN}.$$

5.8.3 Receiver operating characteristic and precision-recall curves

For all classification predictions, we define some threshold that determines if the predicted outcome of the observation is classified as 0 or 1 (no outcome or outcome) depending on the predicted probability of the observation to be equal to 1. If we set the threshold to 0.5 then we strive to approximate the Bayes classifier which attempts to reduce the overall model error. We classify the predicted outcome of any observation as $\hat{Y} = 1$ if the predicted probability of $\hat{Y} = 1 > 0.5$. However, using a threshold of 0.5 implicitly assumes that the false positive and false negative errors are equally unacceptable which is not always the case for a clinical outcome [87].

The ROC curve is a useful visualisation tool to demonstrate the discriminative ability of a classification model as the threshold for classifying the outcome is varied. The ROC curve is plotted by determining the sensitivity and specificity over various classification thresholds. The area under the curve (AUC) also known as the concordance (c) statistic is used as a measure of the discriminative ability of the classification model. More specifically, the AUC determines the probability that a random observation that has the outcome is predicted by the model to be classified as having the outcome compared to a random observation that does not have the outcome.

However in the case of imbalanced data, specifically where there are far fewer observations for which the outcome of interest was observed compared to the number of observations for which the outcome was not observed, it can be more useful to determine the F1 score and the area under the precision-recall curve (AUCPR) [88]. The F1 score is a measure of the precision and robustness of the classification model and is defined as:

$$\text{F1 Score} = 2 \times \frac{PPV \times TPR}{PPV + TPR},$$

where the score ranges from 0 – 1. A model that has an F1 score close to 1 indicates that the model is both precise and robust in predicting the observations that have the outcome of interest.

Often we are more interested in correctly predicting the observations who have the outcome rather than those who do not. Therefore we are interested in both the number of correctly classified observations and ensuring that the model does not misclassify too many observations for whom the outcome is present.

The precision-recall curve is a plot of the precision versus the recall of the model at different classification thresholds. Similar to the AUC for the ROC curve the AUCPR is a measure of the model's performance. However, the AUCPR is particularly useful for imbalanced datasets when there are far fewer individuals who experienced the outcome than those who did not experience the outcome. The AUCPR is used to determine if the model is capable of both identifying (precision) and accurately predicting (recall) the majority of individuals who experienced the outcome.

The AUC ROC curve does not account for imbalanced outcome classes, the baseline AUC ROC curve threshold is always 0.5. However, since the precision-recall curve does consider the baseline prevalences of the outcome classes, the baseline AUCPR threshold is equivalent to the proportion of observations that have the outcome of interest which makes the precision-recall curve more suitable to imbalanced data, but also less intuitive than the ROC curve [89].

5.8.4 Results

To determine the performances of the final classification models trained using either the balanced or imbalanced placebo training dataset, the final models for each classification technique were used to predict the CP outcome for the placebo test dataset observations. While all the previously mentioned measures of classification model performance were determined (Table 5.3 and Appendix B, Figure B.3), the AUCPR from the application of the models to the placebo test dataset was the most important metric used to compare the performance and discriminative abilities of the trained models (Figure 5.9).

Since 7.7% of the participants developed CP in the placebo test dataset, the baseline threshold or reference AUC of the precision-recall curve was 0.077. Importantly, all the classification models performed better than the baseline AUCPR of 0.077 (Figure 5.9). While the artificial neural network model resulted in the highest AUCPR (0.320) of all the classification models trained using the imbalanced placebo training dataset, the logistic regression model resulted in the highest AUCPR of 0.236 compared to the models trained using the balanced placebo training dataset (Figure 5.9).

Interestingly, although the artificial neural network model trained using the balanced placebo training dataset performed well on the placebo test dataset, the logistic regression model trained using the imbalanced placebo training dataset had the worst performance of all the classification models trained using the imbalanced training dataset (Figure 5.9).

Table 5.3: Comparison of classification accuracy measures between models used to predict CP in the placebo test dataset.

Model	Classification accuracy	Sens	Spec	PPV	NPV	F1 score
Logistic regression - balanced dataset	0.726	0.438	0.750	0.127	0.941	0.197
Logistic regression - imbalanced dataset	0.923	0	1	NA	0.923	NA
CART - balanced dataset	0.736	0.375	0.766	0.118	0.936	0.179
CART - imbalanced dataset	0.894	0.063	0.964	0.125	0.925	0.083
Random forest - balanced dataset	0.875	0.188	0.932	0.188	0.932	0.188
Random forest - imbalanced dataset	0.923	0	1	NA	0.932	NA
Boosted tree - balanced dataset	0.779	0.250	0.823	0.105	0.929	0.148
Boosted tree - imbalanced dataset	0.923	0	1	NA	0.923	NA
Support vector machine - balanced dataset	0.889	0	0.963	0	0.920	NA
Support vector machine - imbalanced dataset	0.923	0	1	NA	0.923	NA
Neural network - balanced dataset	0.841	0.188	0.896	0.130	0.930	0.154
Neural network - imbalanced dataset	0.543	0.688	0.531	0.109	0.953	0.188

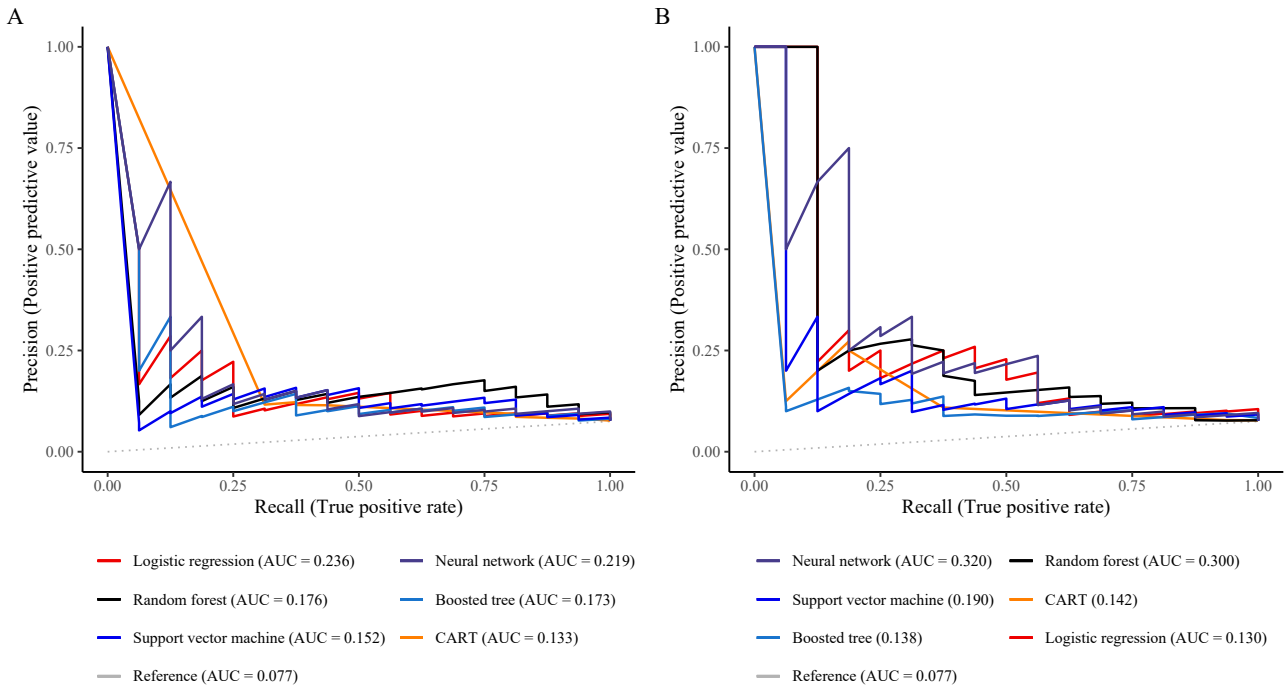

Figure 5.9: Precision-recall curves of the different classification models used to predict CP in the placebo test dataset. A: Final models trained using the SMOTE-NC balanced placebo training dataset, B: final models trained using the imbalanced placebo training dataset.

5.9 Discussion

This chapter aimed to address the primary objective of this thesis which was to determine the baseline demographic, clinical and echocardiographic predictors associated with CP in the cohort of IMPI-1 participants who did not receive prednisolone treatment. Given that there are several different methods available for supervised learning classification, and that the performance of these methods differs greatly within the literature, six of the most common classification techniques were included in the analysis. Importantly, these classification techniques were used to predict the probability of the binary CP outcome as opposed to time to the occurrence of CP. The motivation for this decision is discussed in Chapter 7.

Furthermore, since the IMPI-1 dataset was imbalanced in terms of the CP outcome, the SMOTE-NC preprocessing technique was used to assess if having a balanced dataset improved

the classification performance of the different models.

Although several baseline predictors were found to have strong associations with or influences on the CP outcome in the different models, the subsets of predictors that were deemed to be important differed both between the models trained using either the balanced or imbalanced training datasets and between the different classification techniques applied to the same training dataset. Some of the most influential predictors in the models trained using the balanced dataset were sex, NYHA functional class, tachycardia, anaemia and age. However for the models trained using the imbalanced dataset, the most commonly occurring influential predictors were NYHA functional class, anaemia, age and weight.

Importantly, the feature of pericardiocentesis performed at trial randomisation was not selected by lasso regularisation in either of the multiple logistic regression models constructed using the different training datasets nor was it deemed by any of the other classification methods to be a highly influential predictor of CP.

Crucially, the HIV baseline predictor was not considered to be highly influential in any of the models except for the CART model trained using the balanced dataset. This was surprising considering that clinical signs of HIV infection is one of the only baseline features previously reported to have a significant association with CP [19]. However, the anaemia predictor was featured as a relatively important predictor in most of the models trained using the balanced dataset and some of the models trained using the imbalanced dataset. Anaemia is known to be associated with HIV infection [90] and was found by both logistic regression models to be associated with a decreased odds of CP. In these analyses, the anaemia predictor may be a proxy for clinical signs of HIV infection.

Interestingly, the models trained using the balanced training dataset consistently featured more predictors of CP in the final trained model than the models trained using the imbalanced training dataset. One possibility for this observation is that SMOTE-NC was used to create synthetic minority class observations to increase the prevalence of CP in the balanced training dataset.

Therefore, there might have been an increased prevalence of the predictors associated with CP in the balanced dataset that were not as prevalent in the original imbalanced training dataset thus allowing for the models trained using the balanced dataset to detect stronger associations between CP and the baseline risk factors. However, the creation of the synthetic observations in the balanced dataset may have resulted in a noisier dataset leading to spurious associations between the baseline predictors and CP.

While various measures of classification performance can be used to assess a model's predictive capacity, the AUCPR was determined to be the most important criterion on which to stratify the models in terms of predictive power. This metric was favoured the original IMPI-1 dataset was imbalanced in terms of the CP outcome and the majority of trial participants did not develop CP. Crucially in this research, it is arguably more important to identify and accurately predict the participants at baseline who are the most likely to progress to CP which is highlighted by the AUCPR.

While the AUC ROC curve is more frequently reported and more intuitive than the AUCPR [89], it is important to note that the AUC ROC curve and AUCPR metrics resulted in different rankings of the models in terms of predictive performance especially for the models trained using the balanced dataset (Figure 5.9, Appendix B, Figure B.3). Specifically, while the AUCPR indicated that the logistic regression model trained using the balanced dataset had the most power to predict CP in the test dataset, the random forest had the highest AUC ROC value when applied to the test dataset. Contrastingly, the logistic regression model built using the balanced dataset had one of the lowest AUC ROC values (Appendix B, Figure B.3).

Furthermore, the AUCPR was used rather than a single accuracy metric because all accuracy measures require the specification of an arbitrary classification threshold above which participants will be classified by the model to have the CP outcome and vice versa. The classification accuracy measurements reported are all based on a classification threshold of 0.5 which is meant to approximate the Bayes classifier [20]. This choice of threshold is completely discretionary and implies that equal weighting is given to both outcome classes. However, the implications of failing

to detect a future case of CP are potentially greater than incorrectly predicting a case of CP.

It is considered best practice to consider the AUCPR rather than AUC ROC or any single classification accuracy measure when dealing with heavily imbalanced datasets because imbalances in the outcome result in biased models that are trained to predict the majority class very well and the minority class poorly. For example, the models that resulted in the highest classification accuracy when applied to the placebo test dataset were the logistic regression, random forest, boosted tree and SVM models all built using the imbalanced training dataset each achieving a classification accuracy in the placebo test dataset of 0.923 (Table 5.3).

However, all of these models resulted in a specificity of 1, sensitivity of 0 and PPV that was not estimated (Table 5.3) implying that these models were incapable of predicting a single new case of CP correctly nor were they capable of detecting any of the true cases of CP in the dataset and instead were classifying every observation as not having CP.

Additionally, the application of the neural network built using the imbalanced placebo training dataset to the placebo test dataset resulted in the lowest classification accuracy of 0.543 but the highest sensitivity of 0.688 compared to all the other models (Table 5.3). Similarly, the logistic regression model built using the balanced dataset had the highest F1 score of 0.197 compared to all the models when applied to the placebo test dataset (Table 5.3). These results highlight the importance of determining an appropriate classification metric based on the specific goals of the statistical model.

Since only one model could be selected as the final model used to stratify the IMPI-1 participants according to their risk of CP, the logistic regression model with lasso regularisation trained using the balanced training dataset was chosen. This was because while the neural network and random forest models built using the imbalanced placebo training datasets both outperformed the logistic regression model trained using the balanced placebo training dataset in terms of the AUCPR, these models provide little interpretability and are potentially not as clinically useful for inference as the intuitive logistic regression model.

From a statistical viewpoint, the goal of attempting to develop a classifier for risk estimation requires that the classifier be accurate and have good discriminative capacity. However, from a clinical viewpoint, the classifier needs to be comprehensible for the clinician such that they can understand the associations between the predictor variables and the outcome to better inform disease diagnosis, prognosis, and treatment approaches. Collectively, these reasons provide the motivation for the use of the multiple logistic regression model with lasso regularisation trained using the balanced placebo training dataset as the final model to construct the risk score for CP.

5.10 Summary

This chapter has provided a background to the concepts and notations of six popular classification algorithms namely logistic regression, classification trees, random forests, boosted trees, support vector machines and artificial neural networks. The concept of k-fold cross-validation was introduced with its application to model selection and hyperparameter tuning for each classification technique. Classification error measurements and model accuracy are defined as well as ROC and precision-recall curves as a means to assess the performance of a classification model.

The application of these classification techniques was demonstrated using the IMPI-1 RCT data with the intent to build a classification model for CP risk stratification. The logistic regression model with lasso regularisation trained using the SMOTE-NC balanced training dataset was selected as the final classifier to be used to stratify the IMPI-1 participants according to their risk of CP as it provided reasonable discriminative ability relative to the other models and importantly, is a clinically useful model for interpretation.

The next chapter will present the findings of the final selected classification model in its application to the stratification of the entire cohort of IMPI-1 RCT participants by the predicted risk of CP and compare the impacts of adjunctive prednisolone and pericardiocentesis treatments in the trial participants according to the predicted risk of CP.

Chapter 6

A risk score for constrictive pericarditis

6.1 Introduction

Previously, supervised learning was used to train an appropriate statistical model to predict the CP outcome. This chapter presents the development of a risk score for CP using the predictors derived from the final logistic regression model discussed in Chapter 5. Specifically, to develop the risk score, the logistic regression model, with features determined by lasso regression, will be used to stratify the IMPI-1 trial participants into three risk categories namely high, medium and low risk for CP. The impacts of prednisolone therapy compared to the placebo and pericardiocentesis at study enrollment will be assessed in the different cohorts of participants determined to be at either high, medium or low risk of CP.

6.2 Constrictive pericarditis risk stratification

To develop the final model used to stratify the IMPI-1 trial participants by risk of CP, the logistic regression model trained using the balanced placebo training dataset was reapplied to the entire cohort of participants who did not receive prednisolone treatment (Table 6.1).

Table 6.1: Multiple logistic regression model used to predict the probability of CP.

Predictor	Odds ratio	Std. Error	p-value	95% CI
(Intercept)	0.003	1.52	0.0002	9×10^{-5} - 0.045
Sex: female vs male	0.75	0.32	0.364	0.388 - 1.392
Country: South Africa vs other	1.75	0.44	0.2	0.771 - 4.333
NYHA functional class: II vs I	2.29	0.65	0.206	0.721 - 10.198
NYHA functional class: III vs I	5.48	0.67	0.011	1.658 - 25.137
NYHA functional class: IV vs I	6.23	0.76	0.016	1.498 - 32.297
On TB medication: yes vs no	3.84	1.09	0.216	0.670 - 73.429
Peripheral oedema: yes vs no	1.83	0.34	0.072	0.950 - 3.582
Hypotension: yes vs no	0.65	0.77	0.572	0.100 - 2.374
Tachycardia: yes vs no	1.48	0.35	0.255	0.764 - 2.996
Anaemia: yes vs no	0.55	0.35	0.08	0.274 - 1.067
White blood cell count $> 10 \times 10^9$/L: yes vs no	0.52	0.66	0.322	0.114 - 1.657
Renal impairment: yes vs no	1.02	0.45	0.967	0.396 - 2.358
Cardiac tamponade at presentation: yes vs no	0.47	0.34	0.026	0.243 - 0.916
Effusive-constrictive pericarditis at presentation: yes vs no	2.82	0.36	0.004	1.406 - 5.759
Pulmonary infiltrates: yes vs no	1.50	0.32	0.203	0.797 - 2.779
Definite TBP: yes vs no	1.87	0.36	0.080	0.912 - 3.726
Clinical signs of HIV infection: yes vs no	0.47	0.34	0.025	0.241 - 0.909

Although the HIV status of the participants was not one of the features selected from the lasso regression cross-validation procedure during model training, this predictor was retained in the

final CP prediction model to determine if the results from this analysis supported the previous findings that clinical signs of HIV infection was associated with a reduced incidence of CP [19]. Importantly, from the final logistic regression model, participants with clinical signs of HIV infection were on average 53% less likely than participants without clinical signs of HIV to develop CP (95% CI = 0.241 – 0.909; p-value = 0.025) (Table 6.1) which confirmed the findings from previous research [19].

In contrast to previous analyses, the final logistic regression model also found significant associations between NYHA functional class and CP with NYHA functional class III or IV being associated with an average increased odds of CP of 5.48 (95% CI = 1.658 – 25.137; p-value = 0.011) and 6.23 (95% CI = 1.498 – 32.297; p-value = 0.016) respectively relative to participants with NYHA functional class I (Table 6.1). Unsurprisingly, the participants who had effusive-constrictive pericarditis at presentation were on average 2.82 times as likely as participants who did not have effusive-constrictive pericarditis to develop CP (95% CI = 1.406 – 5.759; p-value = 0.004). Interestingly from this model, it appeared that participants who presented with cardiac tamponade at presentation were on average 53% less likely than participants who did not present with cardiac tamponade to develop CP (95% CI = 0.243 – 0.916; p-value = 0.026) (Table 6.1).

To determine if there was any evidence of multicollinearity in the final logistic regression model, the variance inflation factor was calculated for each of the predictors and determined that there were no strong correlations between any of the features included in the final model (Appendix B, Table B.2). Additionally, Cook's distance and standardised residual plots were used to determine if any of the observations were influential or outliers. From the residual analysis, there was no evidence of outlying or influential observations (Appendix B, Figure B.4).

There is no currently defined risk score or probability threshold by which to determine a person's baseline risk of CP. Moreover, the baseline risk of progressing to an outcome is relative to the population from which the data were sampled [91]. To stratify the IMPI-1 trial participants as either low, medium or high risk for CP, the probability of CP was predicted for each participant

who was randomised to receive the placebo treatment, as opposed to prednisolone, from the final logistic regression model (Appendix B, Figure B.5). From the predicted probability distribution of CP in the group of participants that did not receive the prednisolone treatment, the entire IMPI-1 cohort was stratified by risk of CP according to the following conditions:

- Low risk if $Pr(CP = 1) < 0.025$.
- Medium risk if $0.025 \leq Pr(CP = 1) < 0.05$.
- High risk if $Pr(CP = 1) \geq 0.05$.

These probability thresholds were determined due to both the prevalence of the CP outcome in the dataset and the distribution of the predicted probabilities of CP by the final logisitc regression model in the IMPI-1 cohort who did not receive the prednisolone treatment (Appendix B, Figure B.5).

6.3 The impact of adjunctive prednisolone by risk of CP

To determine the association between prednisolone treatment and CP risk, the risk ratio between the participants who received either the prednisolone or placebo randomised treatment was determined in the cohorts of participants predicted to be at either low, medium or high risk for CP (Table 6.2). To account for the small number of participants who developed CP in some of the risk stratifications, the risk ratios and confidence intervals were determined by unconditional maximum likelihood and normal approximation respectively with a small sample adjustment.

A total of 377 (26.93%) participants were classified as being at low risk for CP of which 180 received the placebo treatment and 197 received the prednisolone treatment (Table 6.2). Participants who were at low risk for CP and were treated with prednisolone were 0.92 (95% CI = 0.084 - 10.047; p-value = 0.676) times as likely as the participants who were at low risk for CP and were treated with the placebo to progress to CP over five years (Table 6.2).

There were 338 (24.14%) participants who were categorised as being at medium risk for CP of which 177 received the placebo treatment and 161 received the prednisolone treatment (Table 6.2).

The participants who were stratified to be at medium risk for CP and treated with prednisolone were 0.12 (95% CI = 0.016 - 0.971; p-value = 0.028) times as likely as the participants stratified to be at medium risk for CP and treated with the placebo to progress to CP over five years (Table 6.2).

Table 6.2: Summary of the effect of prednisolone on the CP outcome in the participants categorised as either low, medium or high risk for CP.

	Placebo			Prednisolone			
	CP outcome			CP outcome			
Risk	No CP	CP	Total	No CP	CP	Total	Risk ratio (95% CI)
Low	179	1	180	195	2	197	0.92 (0.084 - 10.047)
Medium	169	8	177	160	1	161	0.12 (0.016 - 0.971)
High	292	45	337	320	28	348	0.59 (0.378 - 0.925)
Total	640	54	694	675	31	706	0.55 (0.361 - 0.852)

Most of the participants, 685 (48.93%), were categorised to be at high risk for CP of which 337 received the placebo treatment and 348 received the prednisolone treatment (Table 6.2). Unsurprisingly, the majority of cases of CP were observed in the high-risk stratification with 45 and 28 cases of CP being observed in the placebo and prednisolone treatment groups respectively. Therefore the participants who were at high risk of CP and treated with prednisolone were 0.59 (95% CI = 0.378 - 0.925; p-value = 0.025) times as like as the participants who were at high risk of CP and treated with the placebo to develop CP over five years (Table 6.2).

To establish the absolute risk reduction and consequently, the number needed to treat to benefit

from prednisolone in the predicted low, medium and high risk for CP cohorts, the score method was used to determine the risk measures and confidence intervals while accounting for the small sample of participants who developed the CP outcome in each risk stratification (Table 6.3).

Prednisolone was associated with a 5.307 (95% CI = 0.696 - 10.067) absolute reduction of the risk of CP in participants predicted to be at high risk for CP (Table 6.3). Therefore, approximately 19 individuals (95% CI = 10 – 144) predicted to be at high risk for CP would need to be treated with prednisolone for one individual to benefit. Similarly, in the cohort of participants who were predicted to be at medium risk for CP, the absolute risk reduction of CP associated with prednisolone treatment was 3.899 (95% CI = 0.610 – 8.125) implying that one individual at medium risk for CP would benefit from prednisolone treatment for every 26 individuals at medium risk for CP treated with prednisolone (Table 6.3).

Table 6.3: Summary of the effect of prednisolone on absolute risk reduction and number needed to treat in participants categorised as either low, medium or high risk for CP.

	Risk score		
Risk measures, (95% CI)	Low	Medium	High
Absolute risk with placebo	0.556	4.520	13.35
Absolute risk with prednisolone	1.015	0.621	8.05
Absolute risk reduction	0.460 (-2.147 - 3.131) with no benefit from prednisolone	3.899 (0.610 - 8.125)	5.307 (0.696 - 10.067)
The number needed to treat to benefit	No prednisolone benefit	26 (12 - 164)	19 (10 - 144)

Unsurprisingly, since there were so few participants who developed CP in the cohort predicted to be at low risk for CP (Table 6.2), there was no clear absolute risk reduction of CP associated with prednisolone treatment in the low-risk cohort (Table 6.3). Specifically, in the low-risk cohort,

the confidence interval estimated for the absolute risk reduction indicated that the absolute risk reduction with prednisolone ranged from 3.131 times as likely as the placebo treatment to not have any benefit to 2.147 times as likely as the placebo treatment to have some benefit on CP (Table 6.3).

6.4 The impact of pericardiocentesis by risk of CP

To determine the impact of pericardiocentesis performed at randomisation on the risk of CP in the participants predicted to be at either low, medium or high risk for CP, the risk ratios and confidence intervals with a small sample adjustment were estimated (Table 6.4). However, since prednisolone treatment is known to be associated with a reduced risk of CP, especially in the cohorts of participants predicted to be at medium and high risk for CP (Table 6.2), the pericardiocentesis effect on CP was determined separately in the participants who received either the prednisolone or placebo treatments at baseline (Table 6.4).

The risk of CP in each of the low, medium and high-risk categories appeared to be reduced in the participants who had pericardiocentesis performed relative to the participants who did not have pericardiocentesis performed regardless of the randomised treatment at baseline (Table 6.4). However, the relative risk reduction associated with the pericardiocentesis procedure was not statistically significant in any of the CP risk stratifications for either of the randomised treatment groups (Table 6.4).

Table 6.4: Summary of the pericardiocentesis effect on the CP outcome in the participants categorised as either low, medium or high risk for CP.

Placebo							
	Pericardiocentesis not performed			Pericardiocentesis performed			
	CP outcome			CP outcome			
Risk	No CP	CP	Total	No CP	CP	Total	Risk ratio (95% CI)
Low	86	0	86	93	1	94	0.93 (-)
Medium	80	4	84	89	4	93	0.73 (0.189 - 2.832)
High	90	15	105	202	30	232	0.86 (0.482 - 1.523)
Total	256	19	275	384	35	419	1.15 (0.674 - 1.973)
Prednisolone							
	Pericardiocentesis not performed			Pericardiocentesis performed			
	CP outcome			CP outcome			
Risk	No CP	CP	Total	No CP	CP	Total	Risk ratio (95% CI)
Low	92	1	93	103	1	104	0.45 (0.029 - 7.124)
Medium	60	0	60	100	1	101	0.60 (-)
High	115	11	126	205	17	222	0.81 (0.392 - 1.675)
Total	267	12	279	408	19	427	0.96 (0.473 - 1.943)

6.5 Discussion

This chapter aimed to address several secondary objectives of the thesis including developing a risk score for the stratification of the IMPI-1 trial participants into high, medium and low risk for CP and determining the impacts of prednisolone and pericardiocentesis on CP in the different risk classes for CP.

Multiple logistic regression combined with lasso regularisation for feature selection was used to develop a predictive model to predict the probability of CP given several risk factors. These probabilities were used to devise risk categories to assign individuals according to low, medium and high risk for CP. Importantly, the parameter estimates in the logistic regression model were determined using the cohort of participants in the IMPI-1 trial who did not receive the prednisolone treatment. This was because prednisolone treatment was previously demonstrated to reduce the incidence of CP [18].

The final logistic regression model confirmed previous findings of a significant statistical association between HIV infection and reduced odds of CP and additionally, described several other significant statistical associations between CP and the baseline predictors including NYHA functional class, effusive-constrictive pericarditis at presentation and cardiac tamponade at presentation. While there was no clear evidence for precise protective or harmful associations between most of the predictors in the model and the odds of CP, the selection of these risk factors by lasso regression with 10-fold cross-validation was indicative of the potential importance of these baseline features in predicting the odds of CP. Therefore, this analysis has provided some guidance for further investigations to determine the precise associations between these risk factors and the odds of CP.

Furthermore, this analysis confirmed the previous conclusions from the IMPI-1 RCT that prednisolone treatment was associated with a significant reduction in the risk of CP relative to the placebo treatment [18]. However, this research further extended the findings from the IMPI-1 RCT by demonstrating that prednisolone was effective in reducing the relative and absolute risk

of CP in participants predicted to be at medium and high risk for CP. However, this effect was not observed for participants predicted to be at low risk for CP. This was possibly due to the small number of participants who progressed to CP in the low-risk stratification which resulted in an imprecise estimate of the risk ratio.

Crucially, these results confirmed the hypothesis that in individuals who are suspected to have TBP, the baseline risk of progressing to CP is not uniform. Moreover, prednisolone treatment was demonstrated to be associated with a benefit in participants predicted to be at medium and high risk for CP but not in participants predicted to be at low risk for CP. Therefore, prednisolone, which is potentially correlated with treatment-related severities such as malignancy in participants with clinical signs of HIV infection [18], should only be recommended for TBP patients who are suspected to be at medium or high risk for CP.

Notably, while pericardiocentesis was not directly associated with a statistically significant reduction in the relative risk of CP for participants predicted to be at either low, medium or high risk for CP, pericardiocentesis potentially was associated with the CP outcome indirectly. Specifically, the final logistic regression model suggested that the participants who had cardiac tamponade at presentation were on average associated with a significantly decreased odds of CP by 53%. Of the 556 IMPI-1 trial participants who presented with cardiac tamponade at baseline, 448 (80.58%) underwent pericardiocentesis. Therefore it is possible that in the participants with baseline features of cardiac tamponade, pericardiocentesis may have been associated with a reduced risk of effusive pericarditis thus indirectly potentially decreasing the odds of CP in these individuals. However, further research is required to determine both the clinical and statistical interactions between cardiac tamponade, pericardiocentesis and CP.

Lastly, it is important to note that the logistic regression model trained using the balanced training dataset was selected for the construction of the final risk score for CP as it performed decently when applied to the test dataset and provided an intuitive understanding of the associations between the baseline risk factors and the CP outcome. However, the artificial neural network model trained using the imbalanced training dataset resulted in a higher AUCPR than the logistic

regression model when applied to the test dataset.

Since this artificial neural network model would have been selected as the final model used to construct the risk score if not for its lack of clinical meaningfulness, the same analyses conducted in this chapter and the following chapter were performed using the final artificial neural network trained using the imbalanced training dataset. The results of the application of the artificial neural network to the IMPI-1 RCT data are presented in Appendix D.

6.6 Summary

This chapter has presented the results and findings from the application of a logistic regression model, trained and tested in participants who were randomised to receive placebo treatment in the IMPI-1 RCT, in using baseline clinical, demographic and echocardiographic predictors to develop a risk score to categorise individuals as being at either low, medium or high risk for CP.

The impacts of adjunctive prednisolone and pericardiocentesis were assessed in the cohorts of participants stratified according to the predicted risk categories of CP and confirmed that adjunctive prednisolone was associated with both a reduced relative and absolute risk of CP relative to the placebo but additionally, that the benefits of adjunctive prednisolone were only associated with the participants deemed to be at medium and high risk for CP and not at low risk for CP.

Furthermore, while these analyses confirmed that there is limited statistical evidence for a direct benefit of pericardiocentesis on reducing the risk of CP in patients at either low, medium or high risk for CP, there are potentially interacting effects between cardiac tamponade, pericardiocentesis and CP which requires further investigation.

The next chapter will present the findings of the application of the risk score for CP in survival analyses used to determine the effects of the risk of CP on the competing hazards of CP, hospitalisation and death.

Chapter 7

Analysis of time to constrictive pericarditis, hospitalisation and death

7.1 Introduction

Time-to-event or survival analysis is crucial in the field of medical sciences for the analysis of observational data and more importantly, data from randomised clinical trials, in determining both the time to an event or outcome of interest and the risk factors associated with that event [92]. Survival data are characterised by censored data in which we do not know the exact survival times for certain individuals due to administrative termination of the study, loss to follow-up or dropout due to death or another competing event [92].

Standard survival analysis is concerned with the analysis of the time from a subject's initial state to a single event of interest [92]. Consequently, for standard survival analysis, we are interested in the risk set at a particular time which are the subjects who remain in the study and have yet to experience the event of interest [92].

Although standard survival analysis, in which there is a single event of interest, is crucial in analysing the disease or recovery process, in medical research, there is likely more than one event of importance [92]. Additionally, there could be numerous other secondary events, so-called competing risks, which can change the likelihood or even prevent the primary event of interest

from occurring [92].

Consequently, to analyse the disease or recovery process with the interest of the time to an event and the risk factors associated with a particular event time in the presence of secondary events, competing risks survival models have been developed to extend the standard analysis of survival data to more than one event of interest [92].

This chapter provides a brief overview and intuitive understanding of the main methods of survival models which are used for the analysis of time-to-event data. The notation and estimation of the survival function will be introduced and applied to the IMPI-1 RCT data. The Cox proportional hazards model, which is used to model the effects of risk factors on the hazard of the event of interest, is presented. Extensions of the survival and Cox proportional hazards model will be addressed including a brief discussion of addressing the assumption of proportional hazards and additionally the use of survival models for competing risks where there is more than one event of interest. The application of competing risks survival analysis will be applied to the IMPI-1 RCT data to investigate the relationship between the predicted risk stratifications of CP and the competing risks of CP, hospitalisation and death.

7.2 Standard survival analysis

Survival analysis is used for data when the outcome of interest is the time until some event occurs [93]. Typically, when the outcome of interest is the time until an event occurs, we are following observations over some time interval. However, not all observations will experience the event of interest. Individuals that do not experience the event of interest are known as censored observations. For this discussion only right censoring will be considered which occurs when an individual is censored due to the study ending before the event was experienced or the individual was withdrawn or lost to follow-up from the study. In the case of right censoring the individual's complete survival time is not observed. This section discusses the concept and notation of standard survival analysis adapted from Kleinbaum and Klein [93] unless specified otherwise.

7.2.1 The survival function

The main outcome of interest is the time until one specific event occurs. We define the random variable T, known as the event time. The probability of the event of interest occurring at time t is described by the probability density function $f(t)$:

$$\begin{aligned} f(t) &= Pr(T = t) \\ &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t}. \end{aligned}$$

The probability density function describes the probability of the event of interest occurring at time t. The cumulative density function, $F(t)$, describes the probability of the event of interest being experienced before or at time t:

$$F(t) = Pr(T \leq t).$$

The survival function is therefore defined as the probability of not experiencing the event, also known as surviving, up to and beyond time t:

$$\begin{aligned} S(t) &= Pr(T > t) \\ &= 1 - F(t). \end{aligned}$$

The Kaplan-Meier method can be used to estimate the survival function for right-censored data [94]. It considers the number of participants in the study, $d(t_i)$, who experience the event at a specific time, t_i, out of the total number of participants present in the study that "survived", $n(t_i)$, at that specific time. Therefore observations that are censored are included in $n(t_i)$ at the time of censoring but are excluded from the estimation thereafter. The product of the survival times of the observations is determined for each successive time point, $t_1 < t_2 < \cdots < t_i$, where the event of interest is observed. The survival function, $\hat{S}(t)$, is estimated as:

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{d(t_i)}{n(t_i)}\right).$$

The log-rank test is used to determine if the survival functions of two or more groups are different based on the observed and expected number of events in each time interval [93]. The null hypothesis assumes that the survival functions of two or more groups are the same. The log-rank test statistic approximates the chi-square distribution with $k-1$ degrees of freedom where k is the number of groups:

$$\chi^2_{k-1} \sim \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i are the observed number of events and E_i are the expected number of events.

7.2.2 Results

Since several outcomes were recorded in the IMPI-1 RCT including CP, hospitalisation and death and the times from baseline until these respective outcomes occurred, non-parametric Kaplan-Meier curves were used to assess if there were associations between the predicted categorical risk of CP and the survival distributions of the times to CP, hospitalisation and death. The survival distributions of the participants were stratified according to the predicted risk of CP and the randomised prednisolone or placebo treatment at baseline. The log-rank test was used to determine if there were differences between the survival distributions of these different outcomes across the different risk cohorts of CP within each treatment group. Only the first three years of the study follow-up were presented as very few outcome events occurred after three years.

Unsurprisingly, for both treatment groups, the cumulative incidence of CP was the highest in the participants predicted to be at high risk for CP while participants predicted to be at low risk for CP had the lowest cumulative incidence of CP (Figure 7.1). Importantly, the cumulative incidence of CP was higher in the high and medium CP risk participants randomised to receive the placebo treatment relative to the high and medium CP risk participants who were randomised to receive the prednisolone treatment (Figure 7.1). However, a similar trend was not observed for the participants predicted to be at low risk for CP.

The median survival time for the CP outcome could not be estimated since the cumulative incidence of CP in the IMPI-1 cohort did not reach 0.5 (Figure 7.1). However, it was estimated that in the high-risk participants who received the placebo treatment, the probability of experiencing CP in the first year of follow-up was 13.64% (95% CI = 9.75% - 17.35%) while in the low-risk CP cohort the probability of experiencing CP in the first year was 0.57% (95% CI = 0% - 1.68%) (Figure 7.1). However in the prednisolone group, while the probability of experiencing CP in the low-risk cohort was similar to the placebo group, in the high-risk group, the probability of experiencing CP in the first year of follow-up was 8.51% (95% CI = 5.38% - 11.53%) (Figure 7.1).

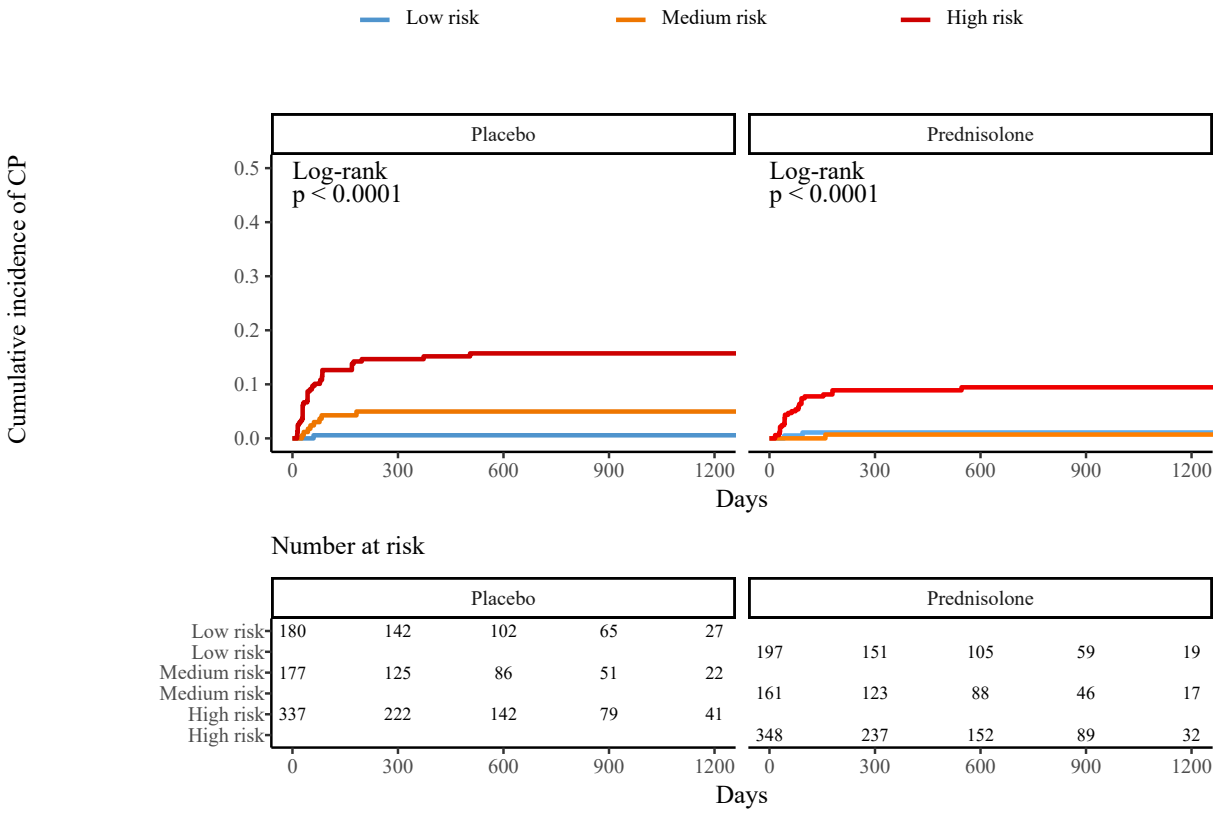

Figure 7.1: Kaplan-Meier estimates of the cumulative incidence of CP as a first event stratified according to the risk of CP and treatment.

For the hospitalisation outcome, while a significant difference was observed between the survival curves of the different CP risk groups in the placebo cohort (log-rank p-value = 0.0008), the distributions of the hospitalisation outcome in the various CP risk stratifications were similar in the cohort of participants who received prednisolone (log-rank p-value = 0.6) (Figure 7.2).

Unsurprisingly, the probability of hospitalisation in the placebo group was greatest in the cohort predicted to be at high risk for CP with a probability of hospitalisation within the first year of follow-up being 29.86% (95% CI = 24.62% - 34.74%) (Figure 7.2). The probability of hospitalisation in the first year in the placebo participants predicted to be at low risk for CP was the lowest at 14.31% (95% CI = 8.95% - 19.35%) (Figure 7.2). However, in the prednisolone treatment group, the probability of hospitalisation within the first year of follow-up was approximately 19% for all the predicted CP risk cohorts (Figure 7.2).

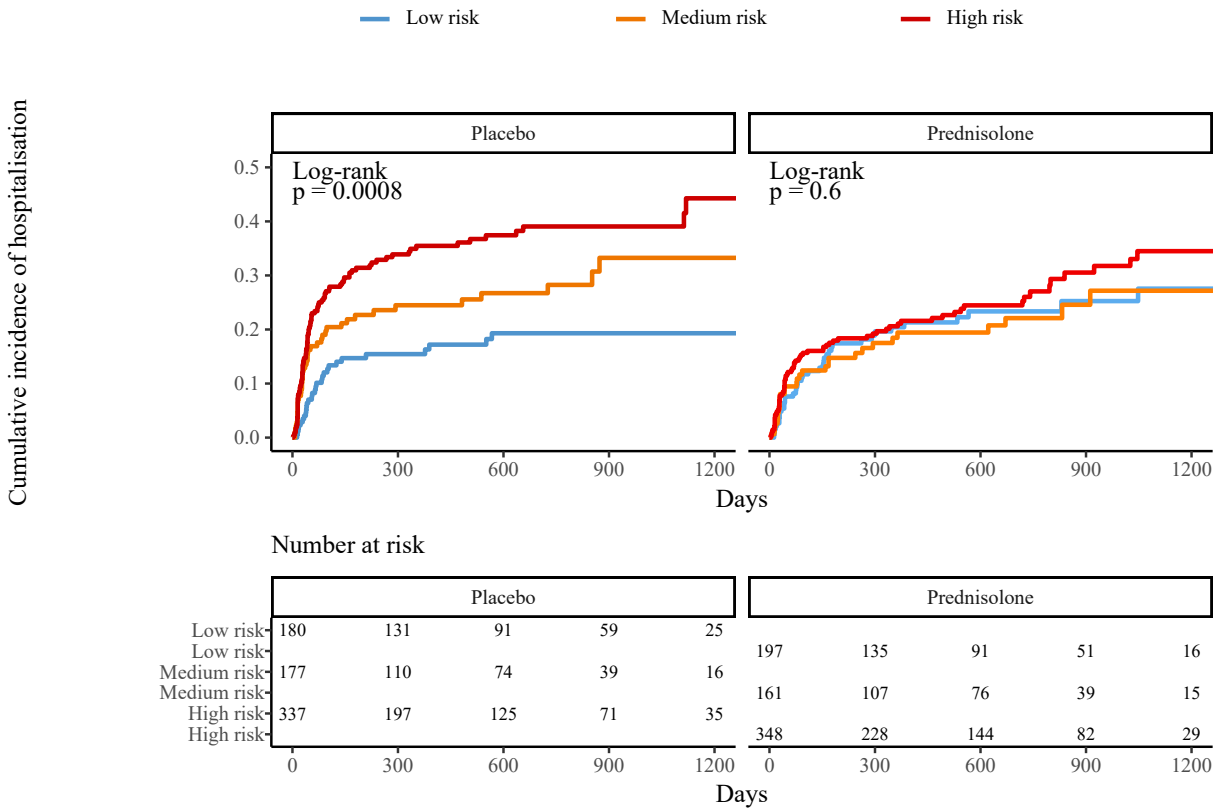

Figure 7.2: Kaplan-Meier estimates of the cumulative incidence of hospitalisation as a first event stratified according to the risk of CP and treatment.

For the death outcome, some of the survival distributions appeared to be significantly different between the CP risk cohorts in the placebo participants (log-rank p-value = 0.026), with participants at high risk for CP appearing to have the highest cumulative incidence of death while participants at medium and low risk for CP appeared to have a lower cumulative incidence of death compared to the high-risk cohort (Figure 7.3). However, the survival distributions of death were

not significantly different between the three CP risk cohorts of participants randomised to receive the prednisolone treatment (log-rank p-value = 0.2) (Figure 7.3).

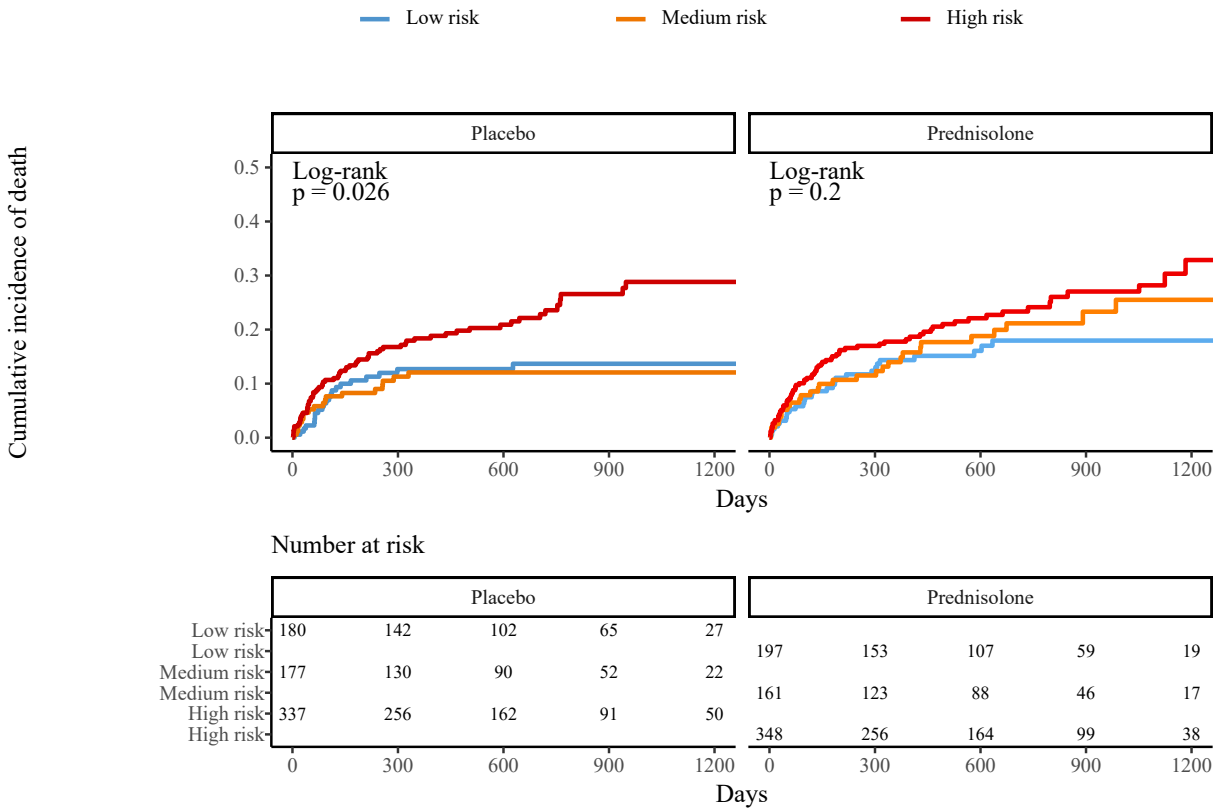

Figure 7.3: Kaplan-Meier estimates of the cumulative incidence of death as a first event stratified according to the risk of CP and treatment.

Within the cohort of participants who received the placebo treatment, the predicted probability of death occurring within the first year in participants predicted to be at high risk for CP was 16.78% (95% CI = 12.62% - 20.74%) (Figure 7.3). This was similar to the predicted probability of death within the first year in high-risk participants who received the prednisolone treatment (*Pr* = 16.29%; 95% CI =12.14% - 20.13%) (Figure 7.3). The predicted probability of death within the first year was approximately 11% for both the low and medium CP risk cohorts randomised to receive the placebo treatment (Figure 7.3).

While non-parametric Kaplan-Meier analysis is useful to visualise the differences between the survival distributions of the three CP risk stratifications in the different treatment groups, the Cox

proportional hazards model, which will be introduced in the following section, can be used to estimate the effects, and their respected uncertainty, of these predictors on the survival outcomes.

7.2.3 The Cox proportional hazards model

The Cox proportional hazards (PH) model describes the relationship between survival and a set of risk factors [95]. However, before defining the Cox PH model, the hazard function must be described. The hazard function describes the probability of experiencing an event given that an individual has survived to a particular time. Specifically, the hazard function, $h(t)$, describes the conditional probability of the event occurring up to time t in the time interval t to $t+\Delta t$, given that the individual has survived up until that interval:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \left\{ \frac{Pr(t \leq T < t+\Delta t | T \geq t}{\Delta t} \right\} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

The cumulative hazard function $H(t)$ represents the risk of experiencing the event at the beginning of the observation period up until time t. Hence its relationship with the survival function is:

$$\begin{aligned} S(t) &= e^{-\int_0^t h(t)dt} \\ &= e^{-H(t)}. \end{aligned}$$

The Nelson-Aalen estimator is used to estimate the cumulative hazard function [96, 97]:

$$H(t) = \sum_{t_i < t} \frac{d(t_i)}{n(t_i)}.$$

The hazard function can be extended to any individual, i, based on the set of risk factors, $X_1, \cdots, X_p$ of that individual using the Cox PH model [95]. The Cox PH model is a semi-parametric regression model because it assumes a parametric relationship between the risk factors, $X_1, \cdots, X_p$, and the hazard of the event. However, no assumption is made about the functional form of the baseline hazard, $h_0(t)$. The baseline hazard function represents the hazard for an individual for which the risk factors are zero. The Cox PH model is defined as:

$$h_i(t) = h_0(t)e^{\beta_1 x_1 + B_2 x_2 + \cdots + \beta_p x_p}$$
$$\therefore \log\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The Cox PH model relies on a fundamental assumption that the hazard ratio, $\frac{h_i(t)}{h_0(t)}$, is constant over time, meaning that the effects of the risk factors on survival are time-independent [95]. However, if the assumption of proportional hazards is violated and the risk factors depend on time, models such as the stratified Cox model and parametric survival models can be used [93].

Maximum likelihood is used to estimate the regression parameters of the Cox PH model. If there are r unique, distinct times of the event, x_j is the vector of risk factors for the individual that experiences the event at time t_j, the individuals at risk for the occurrence of the event, (known as the risk set), at time j is described by: $R_j = \left\{i : t_i \geq t_j\right\}$ then the partial likelihood function for the regression coefficients, $\beta_1, \cdots, \beta_p$, of the Cox PH model can be estimated [98]:

$$l(\beta) = \prod_{j=1}^{r} \frac{e^{\beta x_j}}{\sum_{i \in R_j} e^{\beta x_i}}.$$

Since the Cox PH model is based on assumptions, diagnostic visualisations of the model residuals can be used to determine if the model is an adequate fit to the data and if the assumption of proportionality is valid [99]. The Cox-Snell residuals are used to assess the overall fit of the model to the data, the Martingale residuals are used to determine the functional form of the covariates in the model, the Deviance residuals are used to identify outliers, the Score residuals are used to identify influential observations and the Schoenfeld residuals are used to assess the proportional hazards assumption [99].

7.2.4 Addressing the violation of proportional hazards

The assumption of proportional hazards in the Cox regression models implies that the estimated hazards ratio for a covariate is constant over time for any two subjects. However, there are instances when this assumption is not met for one or more covariates. In the case when the

assumption of proportional hazards is violated, there are several statistical approaches that can be taken to address this violation which will be discussed briefly.

A stratified Cox model can be used to allow for a different baseline hazard for any covariate that does not meet the assumption of proportional hazards:

$$h_g(t) = h_{0g}(t)e^{\beta_1 x_1 + B_2 x_2 + \cdots + \beta_p x_p},$$

where g indicates the stratification level of the predictor that does not meet the assumption of proportional hazards. While this approach is simple to understand and implement, the effect of the stratified variable on the hazard of the outcome cannot be estimated. This is especially limiting in the instance when the effect of a randomised treatment violates the proportional hazards assumption.

Another approach to address the violation of the proportional hazards assumption is to fit a parametric survival model in which the baseline hazard function is assumed to follow a particular distribution thus relaxing the assumption of proportional hazards. Several distributions of the baseline hazard function are implementable including the Weibull, exponential, lognormal, log-logistic, Gompertz and generalised gamma distributions however, the implementation of these methods is beyond the scope of this thesis.

While parametric survival modelling is a common approach to overcome the assumption of proportional hazards for any particular covariate, it can be useful to explore statistically the implications of why the hazard ratio for a particular covariate is not proportional over time.

In the instance when there is clinical relevance to understanding why the hazard ratio for a particular covariate is not proportional over time, the covariates that violate the assumption of proportional hazards can be interacted with various functions of time. This method can be particularly useful when attempting to determine if there is a specific time point at which the effect of a covariate that is not constant over time changes.

7.3 Survival analysis for competing risks

Often in the clinical context, there is more than one event of interest that describes the disease or recovery process [100]. If there is more than one event of interest then these events are known as competing events because several, mutually exclusive events are competing to occur first [100]. Typically, we are concerned with one main event of interest however several other competing events can also occur.

The presence of competing events interferes with the crucial assumption of the Kaplan-Meier estimator that assumes the reason individuals are censored is independent of the survival time distribution of the event of interest [94]. However, if a competing event occurs that prevents the event of interest from occurring, such as death or hospitalisation, then we may never observe the event of interest and the assumption of independence of the Kaplan-estimator is violated [101]. Therefore the Kaplan-Meier estimator for standard survival analysis is an inappropriate choice for estimating survival distributions for competing risks data.

There are two main methods for analysing competing risks data: (1) cause-specific hazard modelling [102] and (2) the Fine and Gray method for modelling the subdistribution hazard [103]. Although these two methods are similar in that they both aim to estimate the survival function for the main event of interest in the presence of competing risks, they differ in their considerations of the risk set [100].

7.3.1 The cause-specific hazard function

Although for standard survival analysis we generally consider the survival function which is the probability of not having experienced the event up to a particular time, t, for cause-specific hazards, we consider the cause-specific cumulative incidence function for a particular event k which is the probability of experiencing the event k by a certain time point, t, given that other competing events can also occur [101]:

$$F_k(t) = P(T \leq t, K = k).$$

The risk set for the cause-specific hazard function differs from the risk set in the standard survival hazard function in that it excludes the individuals who experience competing events, by treating them as censored, while individuals who experience event k are included in the risk set. The cause-specific hazard function for a given cause k at time t is defined as the hazard of experiencing the event k in the presence of other competing events [101]:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t, K = k | T \geq t)}{\Delta t}.$$

The effects of the risk factors on the cause-specific hazard for cause k can be expressed in the proportional hazards framework:

$$\lambda_k(t|\boldsymbol{X}) = \lambda_{k,0}(t) exp(\boldsymbol{\beta}_k^T \boldsymbol{X}),$$

where λ_k is the cause-specific hazard for event k for a subject with a particular set of risk factors $\boldsymbol{X} = X_1, \cdots, X_p$. The baseline cause-specific hazard for event k is indicated by $\lambda_{k,0}$ and $\boldsymbol{\beta}_k$ represents the effects of risk factors on the hazard of cause k.

The partial likelihood function for the regression coefficients is maximised to determine the regression coefficients for the cause-specific PH model. The partial likelihood function considers the risk set $R_k = \{i : t_i \geq t_k\}$ which includes the individuals who have yet to experience any of the events of interest [93]:

$$l(\beta) = \prod_{k=1}^{n} \frac{e^{\beta x_k}}{\sum_{i \in R_k} e^{\beta x_i}}.$$

7.3.2 Results

The outcomes of CP, hospitalisation and death were analysed in the cause-specific proportional hazards framework with CP and hospitalisation as a set of competing events and CP and

death as a separate set of competing events. In the instances when participants were recorded as having experienced hospitalisation and CP at the same time, the participant's survival outcome was recorded as CP because it is reasonable to assume the hospitalisation event occurred because of CP.

For all the models, the effects of predicted risk class of CP and randomised treatment at baseline were considered as either associated with the hazard of the outcomes in an additive manner or as an interaction between the CP risk class and treatment. For some of the models, there was not enough evidence from the data to support a precise estimate for these interaction effects. For the models in which precise interaction terms were estimated, many of these effects did not have statistical significance or enough support from their Akaike information criterion (AIC) value to motivate their use over a simpler model. Therefore, for all the competing outcomes, the additive models are presented in the main body of the thesis while the models containing the interactions are presented in the appendix (Appendix C, Tables C.2 - C.5).

While the relative hazard of CP, with hospitalisation considered as an independent event, appeared to increase 3.50-fold for participants predicted to be at medium risk for CP relative to participants predicted to be at low risk for CP, this effect was not significant (95% CI = 0.706 - 17.325; p-value = 0.125) (Table 7.1). Unsurprisingly, there was evidence for a statistically significant increase in the hazard of CP in the cohort predicted to be at high risk for CP relative to the cohort predicted to be at low risk for CP (HR = 19.25; 95% CI = 4.707 - 78.723; p-value < 0.0001) (Table 7.1). Consistent with previous findings from the original analysis of the IMPI-1 RCT, prednisolone was associated with a decreased risk of the hazard of CP by approximately 49% relative to the placebo treatment (Table 7.1).

Similar trends and associations were observed between the predicted risk of CP, randomised treatment at baseline and the relative hazard of CP with death as an independent event (Appendix C, Table C.1).

Table 7.1: Summary of Cox PH estimates used to predict the relative hazard of CP with hospitalisation considered as an independent event. This model includes the predicted risk of CP and randomised treatment and the AIC = 914.38.

Effect	Hazard ratio	Standard error	Confidence interval	p-value
Medium vs low risk	3.50	0. 82	0.706 - 17.325	0.125
High vs low risk	19.25	0.72	4.707 - 78.732	<0.0001
Prednisolone vs placebo	0.51	0.25	0.313 - 0.835	0.007

Although the participants predicted to be at medium risk for CP had an increased relative hazard of hospitalisation, with CP considered as an independent event, relative to participants at low risk for CP, this effect was not significant (HR = 1.23; 95% CI = 0.883 - 1.726; p-value = 0.218). However, the relative hazard of hospitalisation, with CP considered as an independent event, appeared to increase 1.44-fold in participants predicted to be at high risk for CP relative to participants predicted to be at low risk for CP (95% CI = 1.080 - 1.919; p-value = 0.033) (Table 7.2).

For the hospitalisation outcome, the assumption of proportional hazards was violated for the effect of randomised treatment at baseline. Therefore, to address the assumption of proportional hazards and determine the specific time-points at which the effect of the randomised treatment on the hazard of hospitalisation changed, categorical time interactions with treatment were included. The AIC was used to determine the final time-points to include in the model.

From this analysis, prednisolone was associated with a statistically significant reduction in the hazard of hospitalisation by 42% relative to the placebo within the first 21 days of follow up (95% CI = 0.355 - 0.956; p-value = 0.033) (Table 7.2). However, the size of this reduction was reduced to between 22 days to 2 years of follow up and from 2 to 5 years of follow-up, prednisolone treatment was associated with a 2.72-fold increased risk of hospitalisation relative to the placebo. However this increase was not statistically significant (95% CI = 0.866 - 8.550) (Table 7.2).

Table 7.2: Summary of the Cox PH estimates used to predict the relative hazard of hospitalisation with CP considered as an independent event. This model includes the predicted risk of CP and randomised treatment and the AIC = 4019.87.

Effect	Hazard ratio	Standard error	Confidence interval	p-value
Medium vs low risk	1.23	0.17	0.883 - 1.726	0.218
High vs low risk	1.44	0.15	1.080 - 1.919	0.013
Prednisolone vs placebo 0 - 21 days	0.58	0.25	0.355 - 0.956	0.033
Prednisolone vs placebo 22 days - 2 years	0.84	0.14	0.637 - 1.097	0.197
Prednisolone vs placebo 2 - 5 years	2.72	0.58	0.866 - 8.550	0.087

Although there appeared to be an increased risk in the hazard of death, with CP considered as an independent event, for both participants predicted to be at medium or high risk of CP relative to participants predicted to be at low risk for CP, this effect was only significant in the high-risk cohort (Table 7.3). Prednisolone did not appear to be associated with the relative hazard of death (HR = 1.20; 95% CI = 0.920 - 1.560; p-value = 0.179) (Table 7.3).

Table 7.3: Summary of Cox PH estimates used to predict the relative hazard of death with CP considered as an independent event. This model includes the predicted risk of CP and randomised treatment and the AIC = 3083.44.

Effect	Hazard ratio	Standard error	Confidence interval	p-value
Medium vs low risk	1.11	0.20	0.747 - 1.650	0.606
High vs low risk	1.54	0.17	1.107 - 2.134	0.010
Prednisolone vs placebo	1.20	0.13	0.920 - 1.560	0.179

For all the cause-specific hazards models, the proportional hazards assumption was assessed using

the Schoenfeld test for each covariate (Appendix C, Figures C.1 and C.2). There was no evidence of violations of the assumption of proportional hazards for any of the covariates. The overall model fits were assessed using the plot of the Cox-Snell residuals and illustrated that the models were reasonable approximations of the data (Appendix C, Figure C.3).

However, the main problem with cause-specific hazards modelling is that the effects of the covariates on the events of interest such as CP, hospitalisation or death are not predictable as the cumulative incidence function is dependent on the cause-specific hazards of all the competing events considered in addition to cause k [101]. This implies that the hazard ratios for each covariate must be interpreted as the effect of that covariate on the hazard of the outcome considering that events experienced due to other outcomes are all independent.

Since there is no way to test if we can assume that events due to different outcomes are independent and in the cases of CP, hospitalisation and death which are unlikely to be independent of one another, a second approach to the analysis of competing risks survival data is the use of the subdistribution hazard model which was developed by Fine and Gray and allows us to directly model the cumulative incidence function for a particular event k [103].

7.3.3 The subdistribution hazard function

In contrast to the cause-specific scenario in which the risk set for cause k considers only the subjects who are at risk for event k at time t, the Fine and Gray risk set for a particular event k includes all the individuals regardless of if they experienced a competing event other than k [103]. Only the subjects who did not experience any of the events of interest after study termination are considered censored [101].

Similar to the cause-specific hazard function, the subdistribution hazard for cause k at time t is defined as:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, K = k | T \geq t \cup \{T < t, K \neq k\}}{\Delta t}.$$

The effects of the risk factors in the set $\boldsymbol{X} = X_1, \cdots, X_p$ on the subdistribution hazard of cause k can be defined within the PH framework [101]:

$$\bar{\lambda}_k(t|\boldsymbol{X}) = \bar{\lambda}_{k,0}(t) exp(\boldsymbol{\beta}_k^T \boldsymbol{X}),$$

where $\bar{\lambda}_k$ is the subdistribution hazard for event k:

$$\bar{\lambda}_k(t) = -\frac{\partial \log(1 - F_k(t))}{\partial t}.$$

Similar to the cause-specific PH model, the likelihood function is maximised to estimate the regression coefficients in the subdistribution PH model:

$$\bar{l}(\beta) = \prod_{k=1}^{r} \frac{e^{\beta x_k}}{\sum_{i \in \bar{R}_k} w_{ik}\, e^{\beta x_i}}.$$

However, the likelihood function for the subdistribution PH regression coefficients considers the risk set, $\bar{R}_k = \{i : (t_i \geq t_k) \cup (t_i \leq t_k)\}$, such that the risk set, $\bar{R}_k$, contains all the individuals who have not experienced the event of interest at the time t which includes the subjects who have experienced events other than event k [103].

7.3.4 Results

The outcomes of CP, hospitalisation death were analysed using a subdistribution PH framework similar to the cause-specific framework. However, in contrast to the cause-specific models in which competing events are assumed to be independent, the subdistribution PH function models the relative hazard of experiencing a particular competing event first given that other competing events can also occur.

In resemblance to the cause-specific hazards models, there was no additional benefit of including an interaction effect between the predicted risk of CP and randomised treatment at baseline in the Fine and Gray models (Appendix C, Tables C.7 - C.10).

The sizes and trends of the effects of the predicted risk of CP and randomised treatment on the subdistribution hazard of CP as a first event were similar to the cause-specific hazard models for CP (Table 7.4, Appendix C, Table C.6).

Table 7.4: Summary of subdistribution hazard estimates used to predict the relative hazard of CP given that hospitalisation can occur. This model includes the predicted risk of CP and randomised treatment and the AIC = 933.43.

Effect	Hazard ratio	Standard error	Confidence interval	p-value
Medium vs low risk	3.32	0.81	0.673 - 16.361	0.140
High vs low risk	17.89	0.72	4.377 - 73.144	<0.0001
Prednisolone vs placebo	0.54	0.25	0.334 - 0.886	0.015

In the Fine and Gray model for the hospitalisation as a first event given that CP could also occur, the effect sizes and trends of the predicted risk of CP and the subdistribution of hospitalisation were similar to the cause-specific hazard model (Table 7.5). However, the effect of treatment on the subdistribution hazard of hospitalisation was not proportional with time (Kolmogorov-Smirnov-test p-value = 0.006). Therefore a linear time interaction with treatment was included (Table 7.5) because categorical time interactions are not currently available in the software package used to model the subdistribution hazards [104].

From the Fine and Gray model for the hazard of hospitalisation given that CP could also occur, there was evidence for the effect of a 32% reduction in the hazard of hospitalisation as a first event for the participants who received the prednisolone treatment relative to the participants who received the placebo treatment (95% CI = 0.511 - 0.90). However, from the interaction with linear time, for every one day increase of follow-up, treatment with prednisolone appeared to be associated with a 1.001-fold increase in the subdistribution hazard of hospitalisation as a first event relative to treatment with placebo (95% CI = 1.00 - 1.001) (Table 7.5).

Table 7.5: Summary of subdistribution hazard estimates used to predict the relative hazard of hospitalisation given that CP can occur. This model includes the predicted risk of CP and randomised treatment and the AIC = 4039.75.

Effect	Hazard ratio	Standard error	Confidence interval	p-value
Medium vs low risk	1.22	0.17	0.875 - 1.710	0.240
High vs low risk	1.35	0.15	1.015 - 1.790	0.039
Prednisolone vs placebo	0.68	0.14	0.511 - 0.90	0.007
Prednisolone vs placebo:time	1.001	0.0006	1.00 - 1.001	0.016

Again, similar trends and sizes to the cause-specific hazard model for death were observed for the effects of the predicted risk of CP and randomised treatment on the subdistribution hazard of death as a first event given that CP could also occur (Table 7.6).

Table 7.6: Summary of subdistribution hazard estimates used to predict the relative hazard of death given that CP can occur. This model includes the predicted risk of CP and randomised treatment and the AIC = 3106.47.

Effect	Hazard ratio	Standard error	Confidence interval	p-value
Medium vs low risk	1.10	0.20	0.739 - 1.630	0.640
High vs low risk	1.42	0.17	1.024 - 1.980	0.036
Prednisolone vs placebo	1.24	0.14	0.951 - 1.610	0.110

Although an overall fit of the Fine and Gray models could not be obtained due to current software package limitations [104], we can assume that the Fine and Gray models were adequate fits to the data given that the subdistribution and cause-specific hazard estimates were similar and that the residual checks from the cause-specific hazard models suggested that the models fit the data well (Appendix C, C.3).

7.4 Discussion

This chapter aimed to address one of the secondary objectives of this thesis which was to determine if the risk score derived in the previous chapters was associated with the relative hazards of CP, hospitalisation and death in the IMPI-1 RCT.

Since the outcomes of CP, hospitalisation and death can be considered competing events, two survival analysis methods for competing risks namely (1) cause-specific hazards models and (2) Fine and Gray models were used to determine the potential associations between the predicted risk categories of CP and the competing outcomes of CP hospitalisation and death. Furthermore, the impact of prednisolone on the competing outcomes was also assessed in these models as prednisolone treatment was shown previously to be associated with a decreased relative hazard of both CP and hospitalisation in the IMPI-1 RCT [18].

While the cause-specific and Fine and Gray methods for competing risks have slightly different hazard ratio interpretations, for all the survival outcomes, these two methods resulted in similar estimates of the hazard ratios for both the effects of the predicted risk of CP and randomised treatment at baseline.

Although prednisolone was associated with a reduced relative hazard of CP and hospitalisation, there was no clear association between the effects of prednisolone on the relative hazard of death. Importantly, this research confirmed the results from the original IMPI-1 RCT analysis but additionally determined that the effect of prednisolone on the relative hazard of hospitalisation was not proportional over time. Specifically, these results indicate that while prednisolone was associated with a significant reduction in the hazard of hospitalisation within the first 21 days of follow-up relative to the placebo treatment, this effect decreased over time. Furthermore, there was some evidence that prednisolone treatment appeared to increase the long-term relative hazard of hospitalisation relative to the placebo treatment.

Most of the IMPI-1 trial participants were either hospitalised due to pericardial related outcomes

or for reasons related to TB infection. Therefore, prednisolone treatment may have reduced the relative hazard of the CP outcome early on in the trial follow-up thus potentially reducing the hazard of hospitalisation due to CP-related complications. However, prednisolone treatment was associated with an increased hazard of malignancies [18]. Therefore, one possible explanation for the increased relative hazard of hospitalisation in the long-term follow-up could potentially be due to hospitalisations resulting from the development of malignancies. However, further analysis would be required to determine the possible underlying effects of prednisolone on the hazard of hospitalisation.

Importantly, this research suggested that there was little evidence for an increased relative hazard of any of the competing outcomes of CP, hospitalisation or death in the cohort of participants predicted to be at medium risk for CP relative to the low-risk cohort. However, the participants who were predicted to be at high risk for CP were associated with an increased relative hazard of CP, hospitalisation and death relative to the low-risk cohort.

While these results provide interesting insights into the potential effects of prednisolone treatment and the predicted baseline risk of CP on the various outcomes of CP, hospitalisation and death, there are several extensions to competing risks analysis such as multi-state and semi-competing risks models [101, 105]. These models are used to describe the progression of the disease/recovery process as a series of transitional states through which participants can enter and exit which could lead to a better understanding of the effects of the baseline risk of CP and prednisolone on the outcomes of CP, hospitalisation and death.

Crucially, binary classification was used to develop the risk score since the CP outcome was categorised into two classes: either CP or no CP. However, the risk score could have initially been developed using survival analysis methods since the time to the CP outcome was also recorded. Importantly, the extension of several supervised learning techniques for classification to the field of survival analysis has allowed for increasingly flexible approaches to model a survival outcome. Some of these techniques include lasso regularisation for the Cox PH model [106], random survival forests [107] and neural networks for survival analysis [108, 109].

While binary supervised learning classification, unlike survival analysis, is limited in that it does not consider the censoring process, the use of binary classification for the development of a risk score for CP has several advantages over survival analysis. Firstly, the Cox PH model and its extensions are restricted to predicting the probability of an outcome as a function of time while binary classification is used to predict the probability of an outcome independent of time which is useful for risk classification when we are interested in quantifying the absolute risk of a clinical outcome. Consequently, survival analysis is well-suited to when we are interested in predicting the risk of an individual's survival up to and beyond a specific time-point however binary classification is more appropriate when we are interested in the classification of individuals into different outcomes or risk classes.

Additionally, the accuracy metrics commonly used in binary classification such as AUC ROC can only be extended to a survival model if the model predictions are evaluated at an arbitrarily selected time point which can lead to biased estimates of the survival model's performance [110]. Furthermore, the odds ratios of the covariates in the logistic regression model allow us to understand the baseline risk associated with a specific risk factor and the outcome. However, the hazard ratios in a Cox survival model, are used to quantify the rate at which specific risk factors change the instantaneous probability of the occurrence of an event. Collectively, these considerations motivate the construction of the risk score using binary classification as opposed to survival analysis.

Finally, similar to the discussion in Chapter 6, the results of the application of the artificial neural network model to address the objectives relating to the survival analysis of the IMPI-1 RCT are presented in Appendix D.

7.5 Summary

This chapter has provided a background to the concepts and notation used in standard survival analysis where the outcome is the time to a single event of interest and competing risks survival analysis where the outcome is the time to an event of interest in the presence of competing

risks. The standard survival function and its estimation were introduced and the Cox proportional hazards model was presented to describe the relationship between the hazard of a single event and a set of risk factors. Finally, the cause-specific and subdistribution hazard functions were described for the analysis of a survival outcome in the presence of competing risks.

These concepts were applied to the IMPI-1 RCT data to determine if the effects of the predicted risk of CP and randomised treatment at baseline were associated with the competing outcomes of either CP and hospitalisation or CP and death.

While the effect of prednisolone was shown to reduce the hazard of CP, it did not reduce the hazard of death. Importantly, the effect of prednisolone was also shown to be associated with a reduction in the hazard of hospitalisation early on in follow-up but not in the long-term follow-up. Importantly, these analyses also demonstrated that participants who were predicted to be at high risk for CP were associated with an increased hazard of CP, hospitalisation and death relative to participants who were predicted to be at low risk for CP.

Chapter 8

Conclusions

The primary aim of this thesis was to determine the baseline demographic, clinical and echocardiographic predictors that are associated with CP to construct a risk score used to predict if an individual with either a suspected or confirmed diagnosis of TBP is expected to be at either low, medium or high risk for CP. This was achieved using the data collected from 1400 participants enrolled in the IMPI-1 RCT.

While six different supervised learning classification techniques, ranging from simplistic, namely logistic regression and classification trees, to complex, namely random forests, boosted trees, support vector machines and artificial neural networks, were applied to the IMPI-1 RCT data, the final model used for the construction of the risk score for CP was a multiple logistic regression model with lasso regularisation used to determine the risk factors that were most influential in predicting CP.

The final logistic regression model confirmed the results from previous studies that HIV infection was associated with a significantly reduced odds of CP. This model also identified significant associations between CP and NYHA functional class, effusive-constrictive pericarditis and cardiac tamponade. Additionally, the final logistic regression model demonstrated associations between CP and numerous other risk factors including peripheral oedema, hypotension, tachycardia, anaemia, white cell count, renal impairment and pulmonary infiltrates that were not considered statistically significant by the model but were important in predicting the baseline risk of CP.

Furthermore, these analyses supported the conclusions from the original IMPI-1 RCT that prednisolone treatment was associated with a significant reduction in the risk of CP. However, this research further extended these findings by demonstrating that while prednisolone was effective in reducing the relative and absolute risk of CP in participants predicted to be at medium and high risk for CP, this effect was not observed for participants predicted to be at low risk for CP.

As in previous research, pericardiocentesis was not shown to be directly associated with a reduction in the relative risk of CP in any of the predicted CP risk stratifications. However, pericardiocentesis may have indirectly reduced the risk of CP when it was performed in individuals with cardiac tamponade.

The analysis of the competing survival outcomes of CP, hospitalisation and death confirmed previous findings that prednisolone treatment was associated with a decreased relative hazard of CP and hospitalisation but not death. However, these analyses further demonstrated that individuals who were predicted to be at high risk for CP were associated with an increased hazard of CP, hospitalisation and death relative to individuals who were predicted to be at low risk for CP.

In totality, these analyses confirmed the hypothesis that in individuals who are suspected or confirmed to have TBP, the baseline risk of progressing to CP is not uniform with several identified risk factors being able to predict an individual's baseline risk for CP. Furthermore, while the effects of interventions such as pericardiocentesis on individuals with different risks for CP are still unclear, prednisolone treatment was only associated with a benefit to individuals predicted to be at medium and high risk for CP. Therefore, the use of prednisolone treatment in reducing the risk of CP should only be recommended for individuals suspected to be at either medium or high risk of CP as they are the most likely to benefit while prednisolone treatment should potentially be avoided in treating individuals with TBP that are suspected to be at low risk for CP as they are the least likely to derive any benefit.

Limitations and future research

This research had several limitations. Firstly, while multiple imputation and SMOTE-NC were necessary preprocessing steps required for statistical modelling of the binary CP outcome, these techniques are not ideal as they result in synthetic observations that were not truly recorded. While the application of these preprocessing techniques was not the focus of this thesis, ideally several different imputation and data balancing methods should be applied and compared to determine the most appropriate currently available technique.

Additionally, although logistic regression was selected to stratify the cohort into the various risk classes of CP, the logistic regression model did not have the best performance when applied to the test data. Therefore a different classification method could have resulted in different results and interpretations. While the applications of the artificial neural network model, that had the best model performance, are presented to suggest that overall the general findings were similar across these two models, a rigorous sensitivity analysis should be conducted using all the trained models to validate the findings. Moreover, while internal validation techniques were used to determine the predictive performance of the classification models, the final classifier for CP needs to be externally validated in ideally many different diverse samples of individuals with suspected or probable TBP.

Lastly, while classification, which considered the binary outcome of CP, was used to construct the risk score for CP, another approach could be to use the CP survival outcome to develop a risk score to predict the risk factors associated with the relative hazard of progressing to CP.

Appendix A

Data exploration and preprocessing

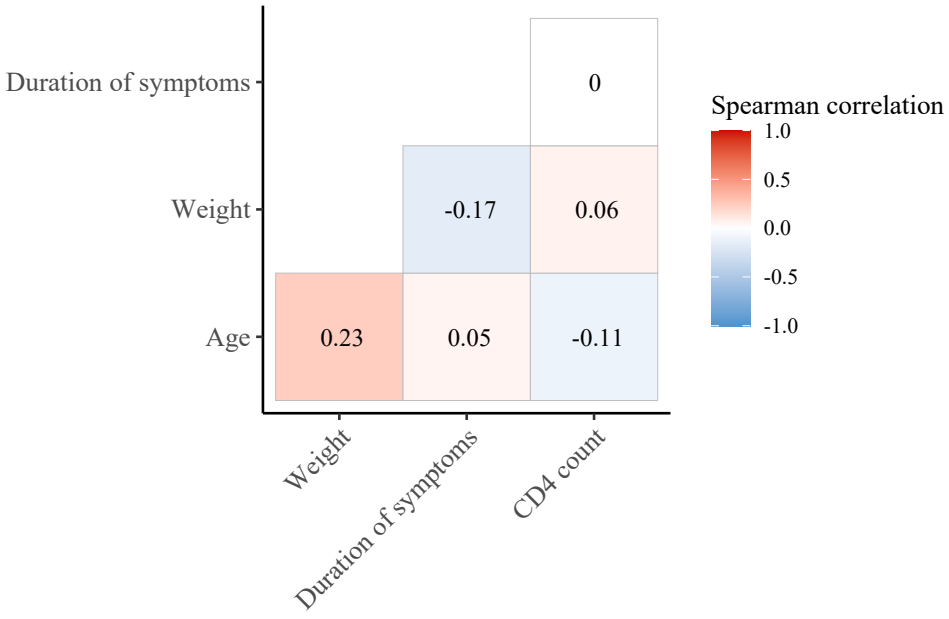

Figure A.1: Correlations between continuous variables at baseline. The numbers inside the squares are the Spearman correlation coefficients between two variables. Squares shaded in red represent a positive Spearman correlation while squares shaded in blue represent a negative Spearman correlation.

Table A.1: Baseline characteristics of HIV-positive patients stratified according to treatment. The categorical variables are presented as a count of patients with percentages indicated in brackets.

Characteristic	Overall (n = 939)	Treatment		p-value
		Prednisolone (n = 474)	Placebo (n = 465)	
CD4 count (cells/μL), n (%)				
0 - 50	114 (12.14)	63 (13.29)	51 (10.97)	
50 - 200	339 (36.10)	179 (37.76)	160 (34.41)	
>200	262 (27.90)	129 (27.22)	133 (28.60)	
Unknown	224 (23.86)	103 (21.73)	121 (26.02)	0.50
HIV opportunistic infections, n (%)				
Yes	530 (56.44)	258 (54.43)	272 (58.49)	
No	278 (29.61)	141 (29.75)	137 (29.46)	
Unknown	131 (13.95)	75 (15.82)	56 (12.04)	0.58
On anti-retroviral treatment at study entry, n (%)				
Yes	199 (21.19)	97 (20.46)	102 (21.94)	
No	713 (75.93)	365 (77)	348 (74.84)	
Unknown	27 (2.88)	12 (2.53)	15 (3.23)	0.68

Table A.2: Comparison of categorical variable proportions in original and imputed placebo training datasets.

Variable	Original placebo training dataset (%)	Imputed placebo training dataset (%)
On TB medication before randomisation		
Yes	93.58	95.06
No	6.42	4.94
NYHA functional class		
Class I	18.35	18.52
Class II	50.72	50.62
Class III	23.09	23.05
Class IV	7.84	7.82
Anaemia (haemoglobin $\leq$ 10 g/dL)		
Yes	53.01	52.88
No	46.99	47.12
Renal impairment (creatinine $>$ 105 μmol/L)		
Yes	12.44	11.11
No	87.56	88.89
White cell count $>$ 10x10^9/L		
Yes	8.26	8.23
No	91.74	91.77
Effusion size (cm)		
Small ($<$1cm)	8.65	8.64
Medium (1-2cm)	24.89	24.90
Large ($\geq$2cm)	66.46	66.46
Cardiac tamponade at presentation		
Yes	55.09	61.52
No	44.91	38.48

Table A.2 Continued: Comparison of categorical variable proportions in original and imputed placebo training datasets.

Variable	Original placebo training dataset (%)	Imputed placebo training dataset (%)
Effusive-constrictive pericarditis at presentation		
Yes	43.23	48.97
No	56.77	51.03
Pulmonary infiltrates on chest radiograph		
Yes	31.19	29.63
No	68.81	70.37
Atrial fibrillation on electrocardiogram		
Yes	3.57	2.67
No	96.43	97.33

Table A.3: Comparison of categorical variable proportions in original and imputed placebo test datasets.

Variable	Original placebo test dataset (%)	Imputed placebo test dataset (%)
On TB medication before randomisation		
Yes	94.44	95.19
No	5.56	4.81
NYHA functional class		
Class I	14.49	14.42
Class II	51.21	51.44
Class III	26.57	26.44
Class IV	7.73	7.69

Table A.3 Continued: Comparison of categorical variable proportions in original and imputed placebo test datasets.

Variable	Original placebo test dataset (%)	Imputed placebo test dataset (%)
Renal impairment (creatinine $>$ 105 μmol/L)		
Yes	13.40	13.94
No	86.60	86.06
Effusion size (cm)		
Small ($<$1cm)	7.46	7.21
Medium (1-2cm)	20.40	19.71
Large ($\geq$2cm)	72.14	73.08
Cardiac tamponade at presentation		
Yes	61.04	58.65
No	38.96	41.35
Effusive-constrictive pericarditis at presentation		
Yes	42.36	42.31
No	57.64	57.69
Pulmonary infiltrates on chest radiograph		
Yes	33.51	31.73
No	66.49	68.27
Atrial fibrillation on electrocardiogram		
Yes	6.88	5.29
No	93.13	94.71

Table A.4: Comparison of categorical variable proportions in original and imputed placebo full datasets.

Variable	Original placebo dataset (%)	Imputed placebo dataset (%)
On TB medication before randomisation		
Yes	93.84	95.24
No	6.16	4.76
NYHA functional class		
Class I	17.20	17.15
Class II	50.87	51.01
Class III	24.13	24.06
Class IV	7.80	7.78
Anaemia (haemoglobin $\leq$ 10 g/dL)		
Yes	52.69	52.59
No	47.31	47.41
Renal impairment (creatinine $>$ 105 μmol/L)		
Yes	12.74	12.10
No	87.26	87.90
White cell count $>$ 10x10^9/L		
Yes	8.24	8.21
No	91.76	91.79
Effusion size (cm)		
Small ($<$1cm)	8.39	8.50
Medium (1-2cm)	23.56	23.05
Large ($\geq$2cm)	68.15	68.44
Cardiac tamponade at presentation		
Yes	56.97	63.11
No	43.03	36.89

Table A.4 Continued: Comparison of categorical variable proportions in original and imputed placebo full datasets.

Variable	Original placebo dataset (%)	Imputed placebo dataset (%)
Effusive-constrictive pericarditis at presentation		
Yes	42.97	48.56
No	57.03	51.44
Pulmonary infiltrates on chest radiograph		
Yes	31.87	31.12
No	68.13	68.88
Atrial fibrillation on electrocardiogram		
Yes	4.58	3.46
No	95.42	96.54

Table A.5: Comparison of categorical variable proportions in original and imputed prednisolone full datasets.

Variable	Original prednisolone dataset (%)	Imputed prednisolone dataset (%)
On TB medication before randomisation		
Yes	95.43	96.46
No	4.57	3.54
NYHA functional class		
Class I	19.43	19.55
Class II	48.51	48.44
Class III	23.12	23.09
Class IV	8.94	8.92

Table A.5 Continued: Comparison of categorical variable proportions in original and imputed prednisolone full datasets.

Variable	Original prednisolone full dataset (%)	Imputed prednisolone dataset (%)
Anaemia (haemoglobin $\leq$ 10 g/dL)		
Yes	54.69	54.67
No	45.31	45.33
Renal impairment (creatinine $>$ 105 μmol/L)		
Yes	12.42	11.76
No	87.58	88.24
White cell count $>$ 10x10^9/L		
Yes	5.97	5.95
No	94.03	94.05
Effusion size (cm)		
Small ($<$1cm)	7.31	7.37
Medium (1-2cm)	25.15	25.07
Large ($\geq$2cm)	67.54	67.56
Cardiac tamponade at presentation		
Yes	56.62	57.65
No	43.38	42.35
Effusive-constrictive pericarditis at presentation		
Yes	45.19	49.01
No	54.81	50.99
Pulmonary infiltrates on chest radiograph		
Yes	33.96	34.14
No	66.04	65.86

Table A.5 Continued: Comparison of categorical variable proportions in original and imputed prednisolone full datasets.

Variable	Original prednisolone full dataset (%)	Imputed prednisolone dataset (%)
Atrial fibrillation on electrocardiogram		
Yes	6.18	4.67
No	93.82	95.33

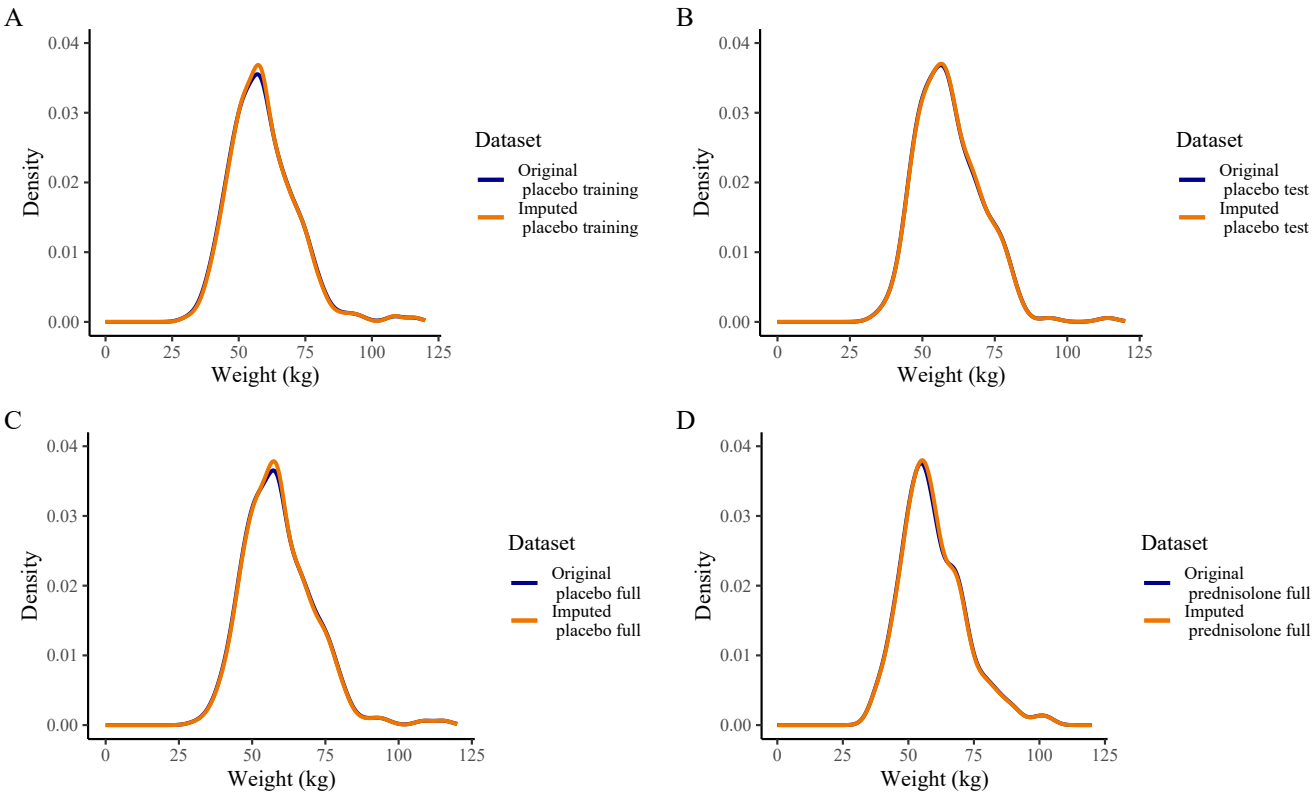

Figure A.2: Comparison of the original and imputed density distributions of participant weight at baseline. A: placebo training dataset, B: placebo test dataset, C: placebo full dataset, D: prednisolone full dataset.

Appendix B

Constrictive pericarditis classification

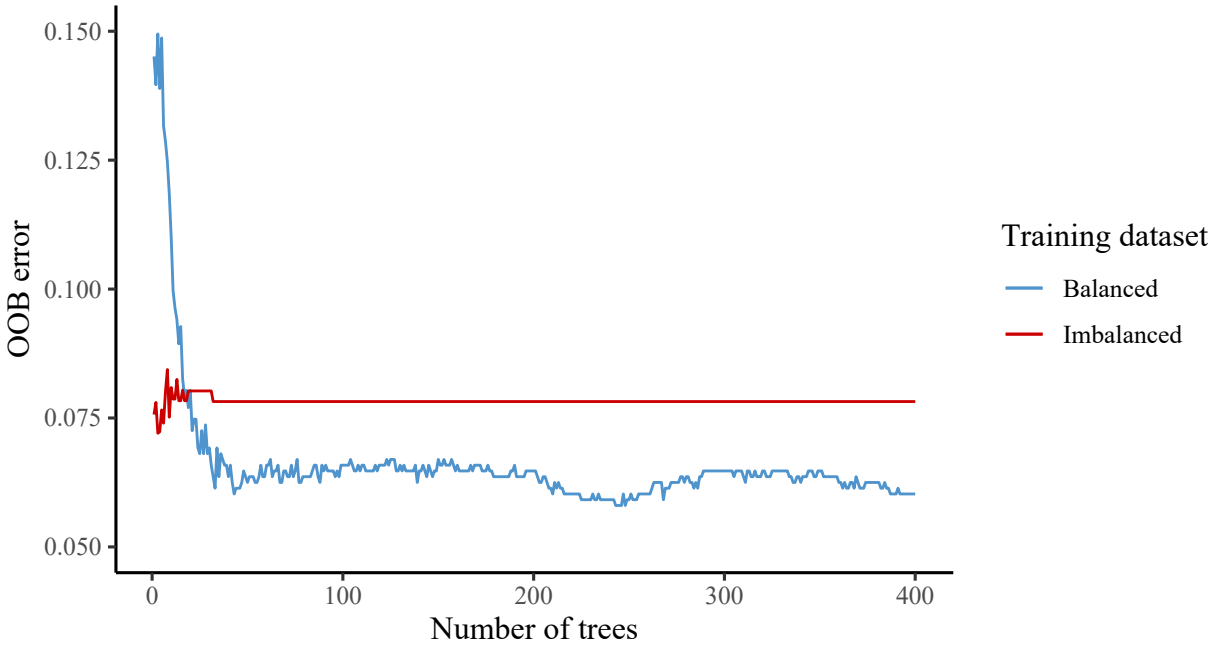

Figure B.1: Comparison of OOB errors of random forests trained using either the balanced or imbalanced placebo training dataset.

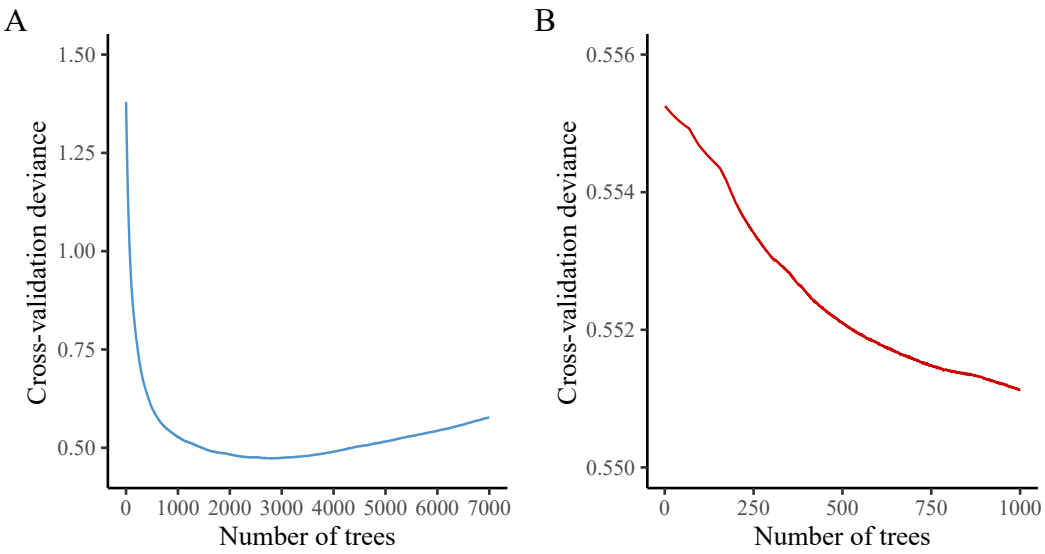

Figure B.2: Comparison of cross-validation reduction in deviance of boosted trees built using either the balanced or imbalanced placebo training dataset. A: SMOTE-NC balanced placebo training dataset, B: imbalanced placebo training dataset.

Table B.1: Summary of SVM tuned hyperparameters that resulted in lowest cross-validation error using different kernel functions for both the balanced and imbalanced placebo training datasets.

	Placebo training dataset			
	SMOTE-NC balanced		Imbalanced	
Kernel function	Tuned hyperparameters	CV error	Tuned hyperparameters	CV error
Linear	$C = 10$	0.1417	$C = 0.0001$	0.0783
Polynomial	$C = 0.001$ $\sigma = 0.5$ $d = 5$	0.0714	$C = 0.0001$ $\sigma = 0.005$ $d = 2$	0.0783
Radial basis	$C = 1$ $\sigma = 0.5$	0.0535	$C = 0.0001$ $\sigma = 0.005$	0.0783
Sigmoid	$C = 10$ $\sigma = 0.01$	0.1518	$C = 0.0001$ $\sigma = 0.005$	0.0783

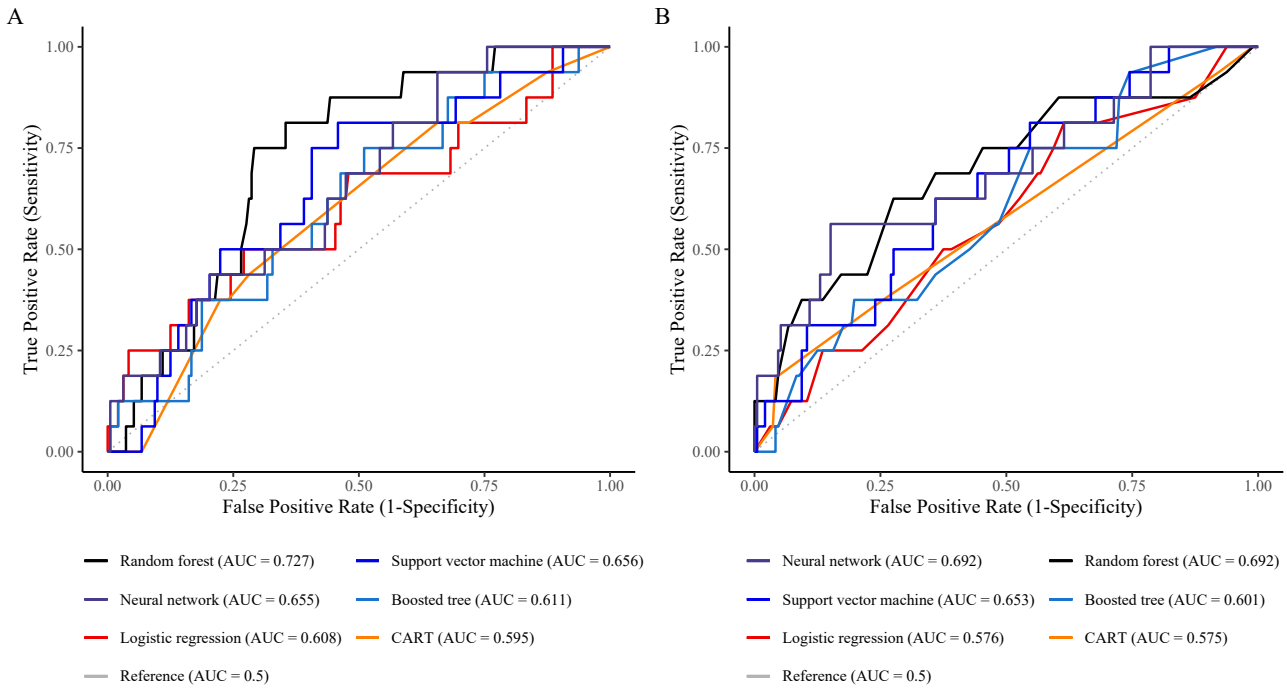

Figure B.3: ROC curves of the different classification models used to predict CP in the placebo test dataset. A: Final models built using the SMOTE-NC balanced placebo training dataset, B: final models built using the imbalanced placebo training dataset.

Table B.2: Variance inflation factor of predictors in the final multiple logistic regression model used to predict the probability of CP.

Predictor	Variance inflation factor
Sex	1.07
Country	1.25
NYHA functional class	1.29
On TB medication	1.08
Peripheral oedema	1.27
Hypotension	1.04
Tachycardia	1.13
Anaemia	1.26
White cell count	1.11
Renal impairment	1.13
Cardiac tamponade at presentation	1.22
Effusive-constrictive pericarditis at presentation	1.41
Pulmonary infiltrates	1.09
Definite TBP	1.16
Clinical signs of HIV infection	1.28

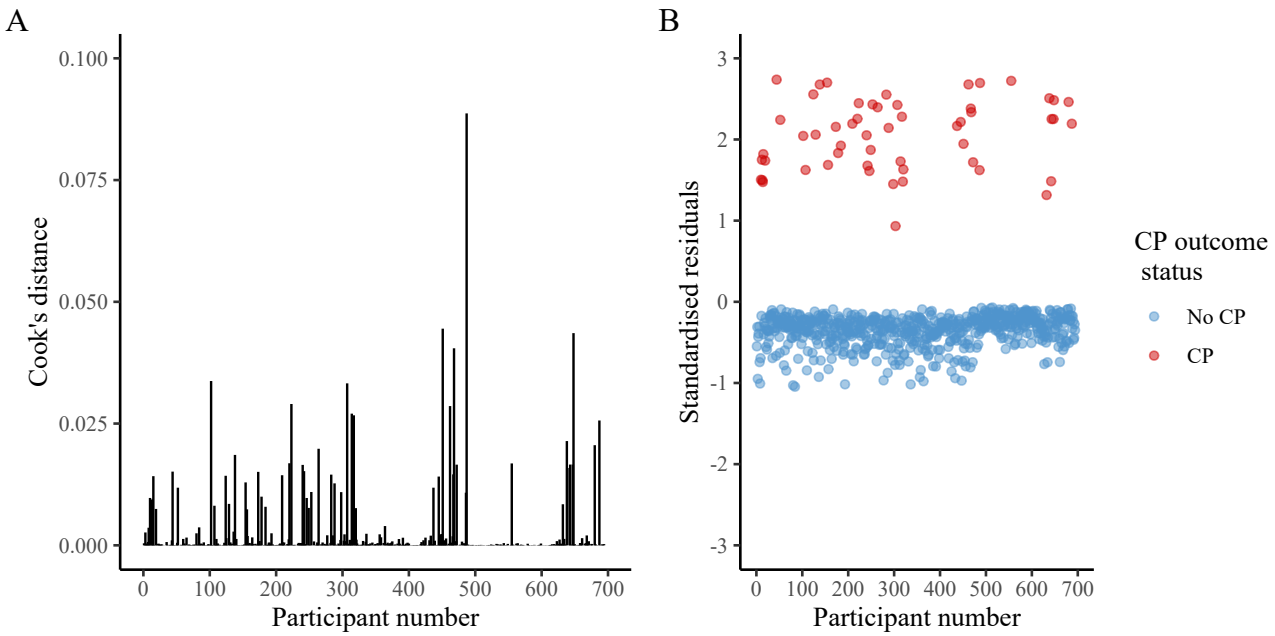

Figure B.4: Residual diagnostics of the final multiple logistic regression model used to predict the probability of CP. A: Cook's distance plot to indicate influential observation, B: standardised residuals plot to indicate outlying observations.

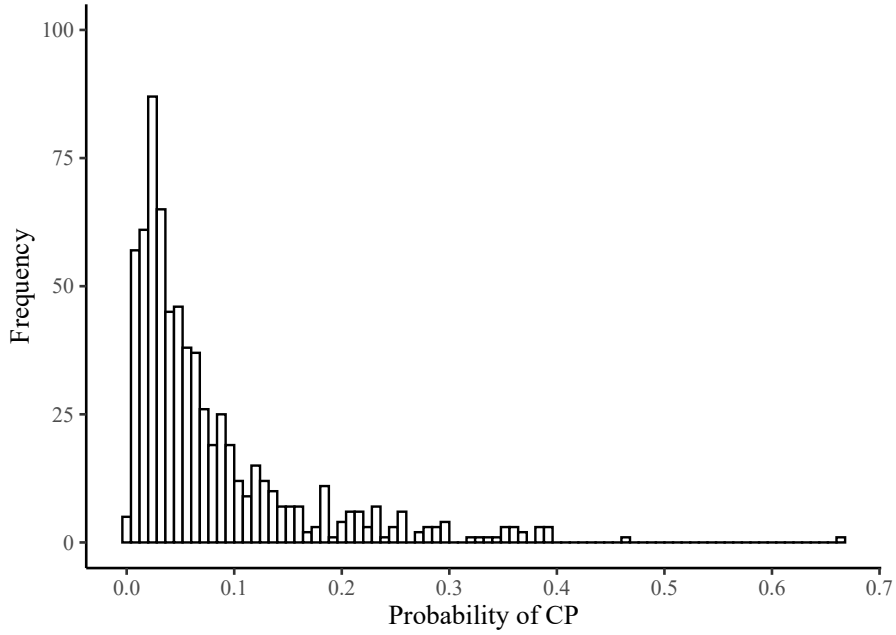

Figure B.5: Density distribution of the probabilities of CP in the IMPI-1 cohort who did not receive the prednisolone treatment determined by the final logistic regression model.

Appendix C

Competing risks analysis

Table C.1: Summary of Cox PH estimates used to predict the relative hazard of CP with death considered as an independent event. This model includes the predicted risk of CP and randomised treatment and the AIC = 1150.64.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	3.34	0.67	0.905 - 12.343	0.070
High vs low risk	14.52	0.59	4.577 - 46.073	$<$0.0001
Prednisolone vs placebo	0.55	0.23	0.353 - 0.855	0.008

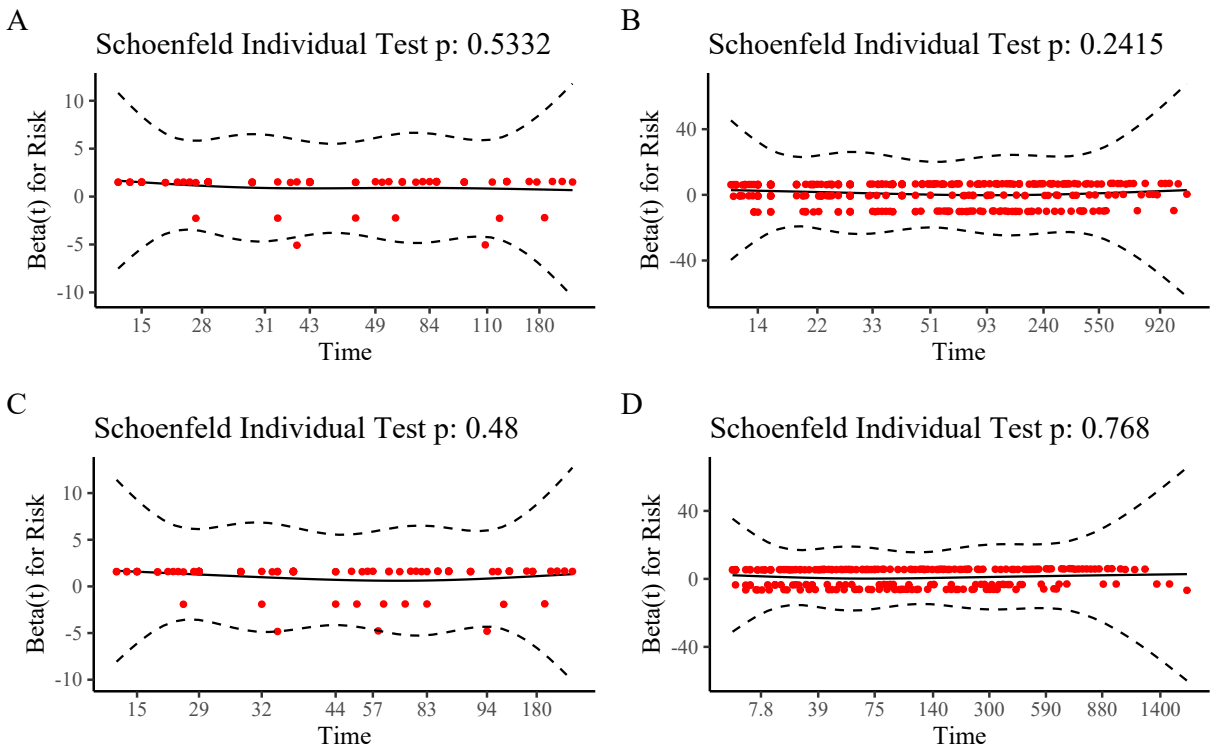

Figure C.1: Test for proportional hazards of risk of CP using Schoenfeld residuals as a function of time in the final models for each cause-specific outcome. A: CP with hospitalisation as an independent event, B: hospitalisation with CP as an independent event, C: CP with death as an independent event, D: death with CP as an independent event.

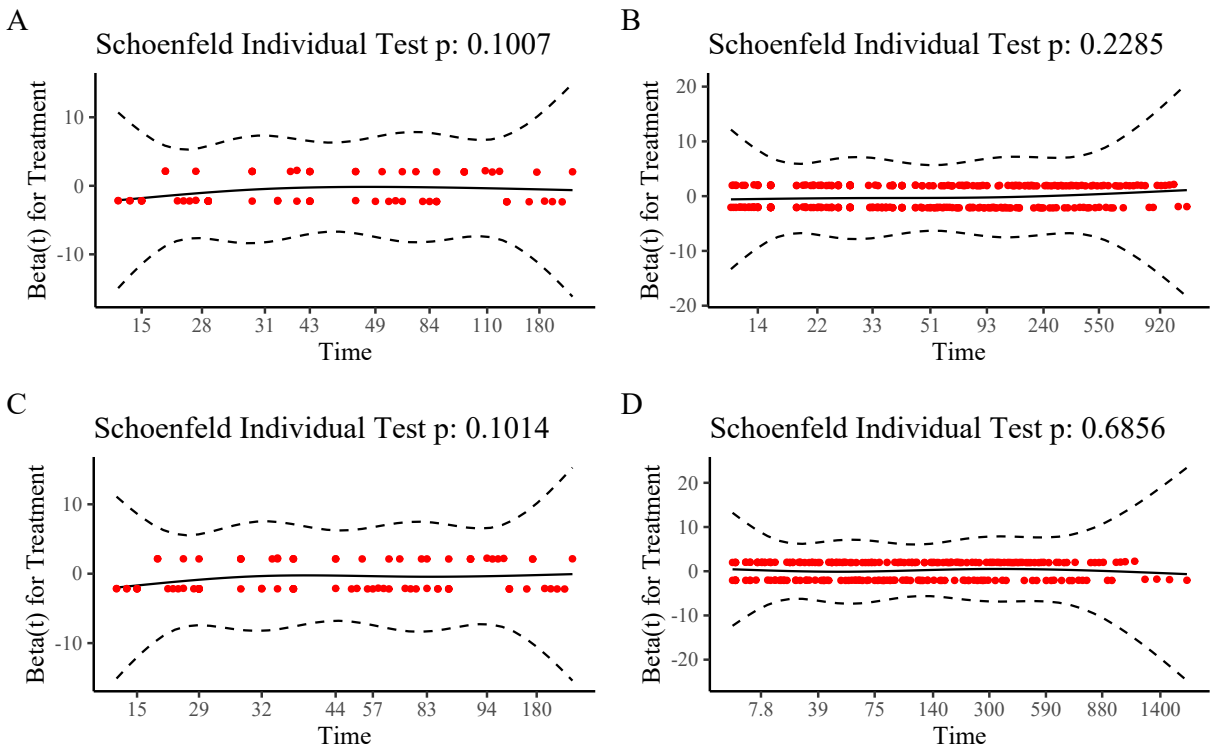

Figure C.2: Test for proportional hazards of treatment using Schoenfeld residuals as a function of time in the final models for each cause-specific outcome. A: CP with hospitalisation as an independent event, B: hospitalisation with CP as an independent event, C: CP with death as an independent event, D: death with CP as an independent event.

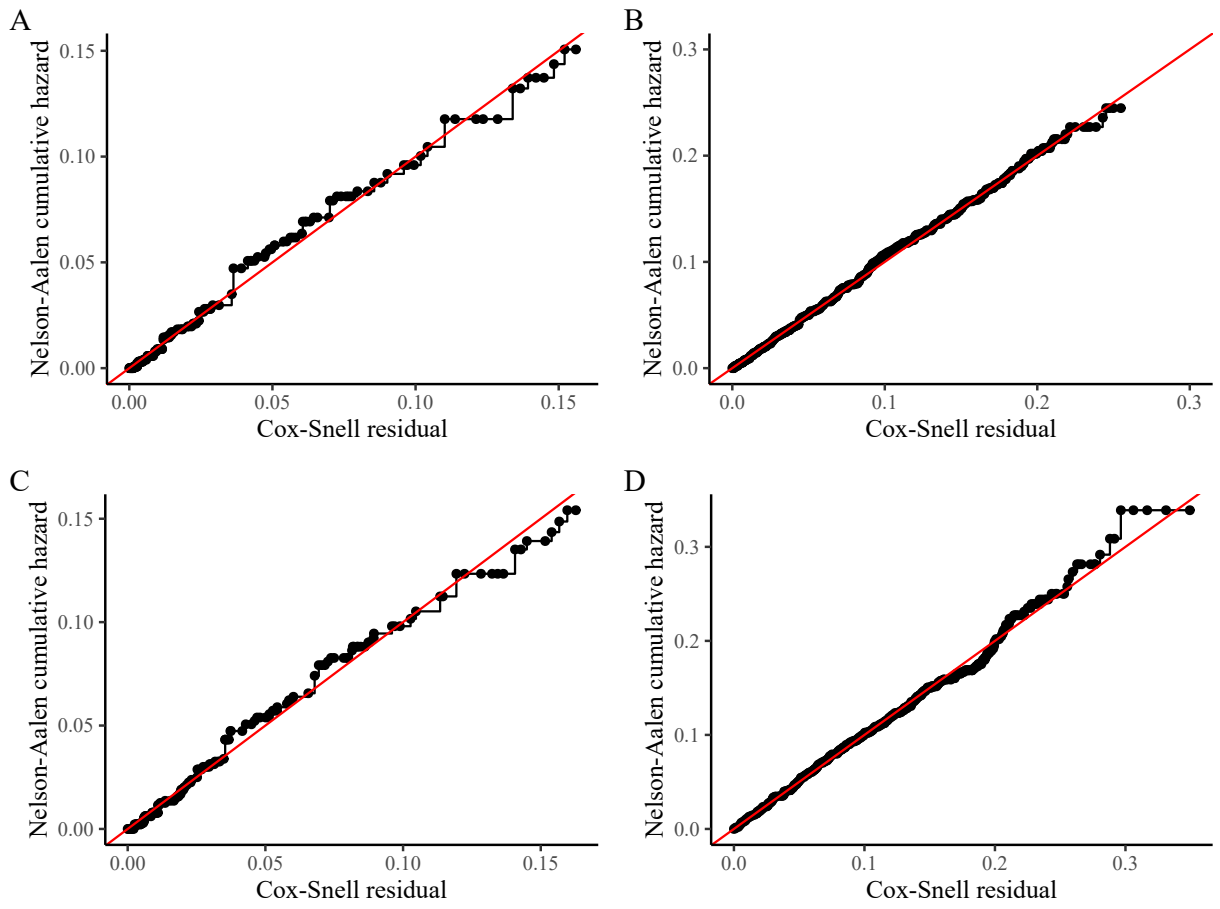

Figure C.3: Assessment of the overall fit of the cause-specific Cox proportional hazard models using Cox-Snell residuals for each cause-specific outcome. A: CP with hospitalisation as an independent event, B: hospitalisation with CP as an independent event, C: CP with death as an independent event, D: death with CP as an independent event.

Table C.2: Summary of Cox PH estimates used to predict the relative hazard of CP with hospitalisation considered as a competing event. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 913.39.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	2.5×10^{7}	2.11×10^{3}	-	0.994
Placebo effect in high risk patients	1.12×10^{8}	2.11×10^{3}	-	0.993
Prednisolone effect in low risk patients	8.32×10^{6}	2.11×10^{3}	-	0.994
Change in prednisolone effect in medium risk patients	2.49×10^{-8}	2.11×10^{3}	-	0.993
Change in prednisolone effect in high risk patients	6×10^{-8}	2.11×10^{3}	-	0.994

Table C.3: Summary of Cox PH estimates used to predict the relative hazard of hospitalisation with CP considered as a competing event. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 4022.59.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	1.62	0.24	1.008 - 2.610	0.046
Placebo effect in high risk patients	1.93	0.22	1.263 - 2.937	0.002
Prednisolone effect in low risk patients	1.25	0.36	0.773 - 2.033	0.359
Change in prednisolone effect in medium risk patients	0.58	0.35	0.296 - 1.150	0.120
Change in prednisolone effect in high risk patients	0.57	0.30	0.317 - 1.011	0.054

Table C.4: Summary of Cox PH estimates used to predict the relative hazard of CP with death considered as a competing event. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 1151.05.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	8.45	1.06	1.057 - 67.538	0.044
Placebo effect in high risk patients	26.79	1.01	3.693 - 194.378	0.001
Prednisolone effect in low risk patients	1.87	1.22	0.170 - 20.671	0.608
Change in prednisolone effect in medium risk patients	0.07	1.62	0.003 - 1.741	0.106
Change in prednisolone effect in high risk patients	0.31	1.25	0.027 - 3.614	0.352

Table C.5: Summary of Cox PH estimates used to predict the relative hazard of death with CP considered as a competing event. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 3085.84.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	0.94	0.31	0.508 - 1.734	0.840
Placebo effect in high risk patients	1.66	0.25	1.016 - 2.706	0.043
Prednisolone effect in low risk patients	1.22	0.28	0.696 - 2.126	0.492
Change in prednisolone effect in medium risk patients	1.34	0.41	0.602 - 3.004	0.471
Change in prednisolone effect in high risk patients	0.871	0.34	0.450 - 1.684	0.681

Table C.6: Summary of subdistribution hazard estimates used to predict the relative hazard of CP given that death can occur. This model includes the predicted risk of CP and randomised treatment and the AIC = 1162.72.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	3.30	0.67	0.897 - 12.135	0.072
High vs low risk	14.05	0.59	4.430 - 44.535	$<$0.0001
Prednisolone vs placebo	0.55	0.22	0.352 - 0.846	0.007

Table C.7: Summary of subdistribution hazard estimates used to predict the relative hazard of CP given that hospitalisation can occur. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 932.61.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	1.52×10^{5}	0.45	6.25×10^{4} - 3.68×10^{5}	<0.0001
Placebo effect in high risk patients	6.62×10^{5}	0.18	4.70×10^{5} - 9.34×10^{5}	<0.0001
Prednisolone effect in low risk patients	5.43×10^{4}	0.71	1.35×10^{4} - 2.18×10^{5}	<0.0001
Change in prednisolone effect in medium risk patients	4.07×10^{-6}	1.30	3.16×10^{-7} - 5.23×10^{-5}	<0.0001
Change in prednisolone effect in high risk patients	9.83×10^{-6}	0.76	2.23×10^{-6} - 4.34×10^{-5}	<0.0001

Table C.8: Summary of subdistribution hazard estimates used to predict the relative hazard of hospitalisation given that CP can occur. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 4045.95.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	1.59	0.24	0.985 - 2.560	0.058
Placebo effect in high risk patients	1.76	0.22	1.152 - 2.680	0.009
Prednisolone effect in low risk patients	1.25	0.24	0.772 - 2.010	0.370
Change in prednisolone effect in medium risk patients	0.60	0.34	0.305 - 1.170	0.140
Change in prednisolone effect in high risk patients	0.60	0.29	0.338 - 1.070	0.081

Table C.9: Summary of subdistribution hazard estimates used to predict the relative hazard of CP given that death can occur. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 1163.13.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	8.33	1.06	1.043 - 66.470	0.046
Placebo effect in high risk patients	25.87	1.01	3.571 - 187.430	0.001
Prednisolone effect in low risk patients	1.85	1.22	0.168 - 20.410	0.610
Change in prednisolone effect in medium risk patients	0.07	1.62	0.003 - 1.740	0.110
Change in prednisolone effect in high risk patients	0.31	1.25	0.027 - 3.620	0.350

Table C.10: Summary of subdistribution hazard estimates used to predict the relative hazard of death given that CP can occur. This model includes the predicted risk of CP and randomised treatment as an interaction effect and the AIC = 3108.97.

Effect	Hazard ratio	Standard error	95% CI	p-value
Placebo effect in medium risk patients	0.91	0.32	0.491 - 1.690	0.770
Placebo effect in high risk patients	1.49	0.25	0.911 - 2.440	0.110
Prednisolone effect in low risk patients	1.21	0.29	0.691 - 2.120	0.50
Change in prednisolone effect in medium risk patients	1.39	0.41	0.623 - 3.120	0.42
Change in prednisolone effect in high risk patients	0.92	0.34	0.473 - 1.780	0.80

Appendix D

Artificial neural network model application

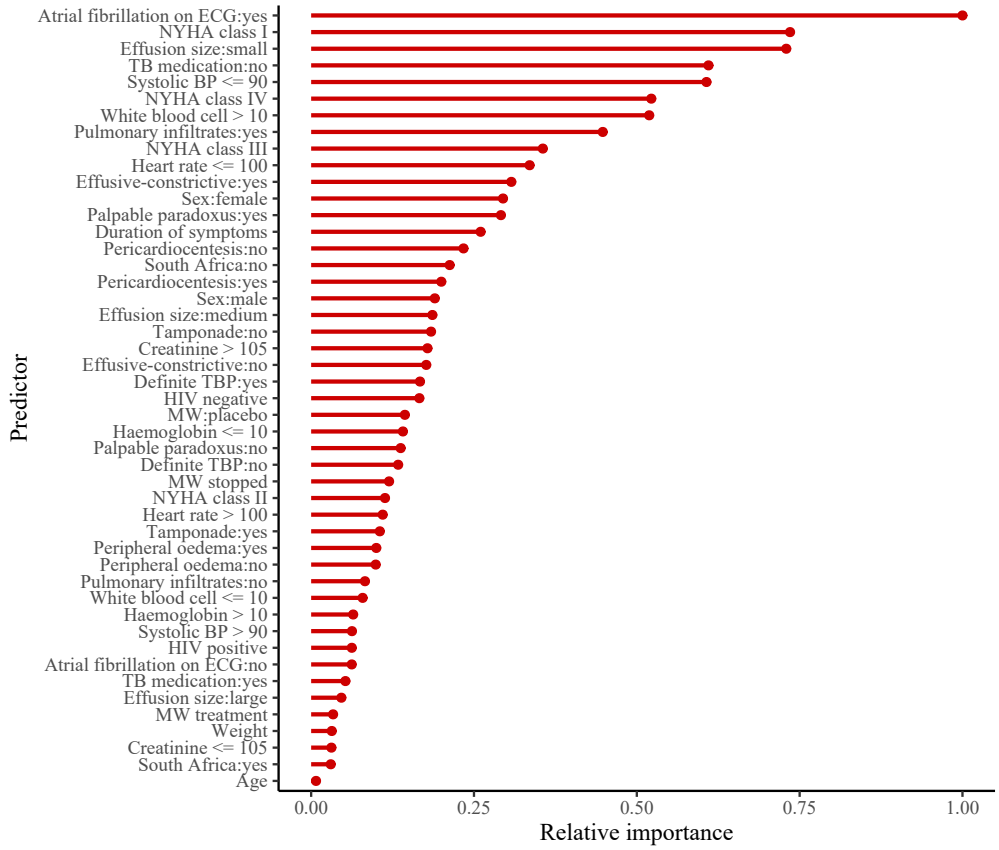

Figure D.1: The relative importance of the variables in the artificial neural network model used to predict CP. The features are ranked according to their importance in predicting CP relative to the other features included in the model.

Table D.1: Summary of the effect of prednisolone on the CP outcome in participants categorised as either low, medium or high risk for CP.

	Placebo			Prednisolone			
	CP outcome			CP outcome			
Risk	No CP	CP	Total	No CP	CP	Total	Risk ratio (95% CI)
Low	91	0	91	88	2	90	2.04 (-)
Medium	99	0	99	109	1	110	0.91 (-)
High	450	54	504	478	28	506	0.51 (0.327 - 0.788)
Total	640	54	694	675	31	706	0.55 (0.361 - 0.852)

Table D.2: Summary of the effect of prednisolone on absolute risk reduction and number needed to treat in participants categorised as either low, medium or high risk for CP.

	Risk score		
Risk measures, % (95% CI)	Low	Medium	High
Absolute risk with placebo	0	0	10.71
Absolute risk with prednisolone	2.22	0.91	5.53
Absolute risk reduction	2.22 (1.874 - 7.745) with no benefit from prednisolone	0.91 (0.865 - 2.683) with no benefit from prednisolone	5.18 (1.847 - 8.655)
The number needed to treat to benefit	No prednisolone benefit	No prednisolone benefit	19 (12 - 54)

Table D.3: Summary of the pericardiocentesis effect on the CP outcome in the patients categorised as either low, medium or high risk for CP.

Placebo							
	Pericardiocentesis not performed			Pericardiocentesis performed			
	CP outcome			CP outcome			
Risk	No CP	CP	Total	No CP	CP	Total	Risk ratio (95% CI)
Low	49	0	49	42	0	42	-
Medium	43	0	43	56	0	56	-
High	164	19	183	286	35	321	1 (0.592 - 1.701)
Total	256	19	275	384	35	419	1.15 (0.674 - 1.973)
Prednisolone							
	Pericardiocentesis not performed			Pericardiocentesis performed			
	CP outcome			CP outcome			
Risk	No CP	CP	Total	No CP	CP	Total	Risk ratio (95% CI)
Low	40	1	41	48	1	49	0.43 (0.028 - 6.641)
Medium	52	0	52	57	1	58	0.91 (-)
High	175	11	186	303	17	320	0.83 (0.396 - 1.729)
Total	267	12	279	408	19	427	0.96 (0.473 - 1.943)

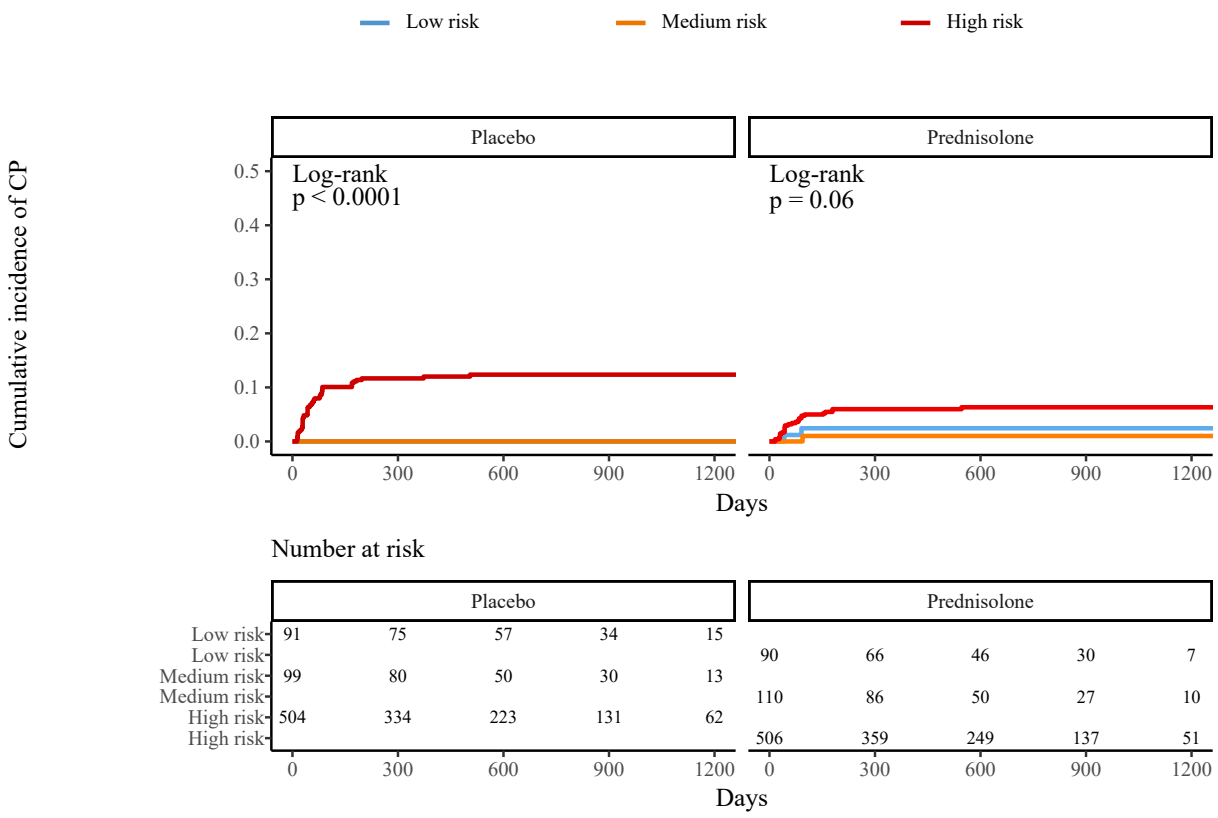

Figure D.2: Kaplan-Meier estimates of the cumulative incidence of CP as a first event stratified according to the risk of CP and treatment.

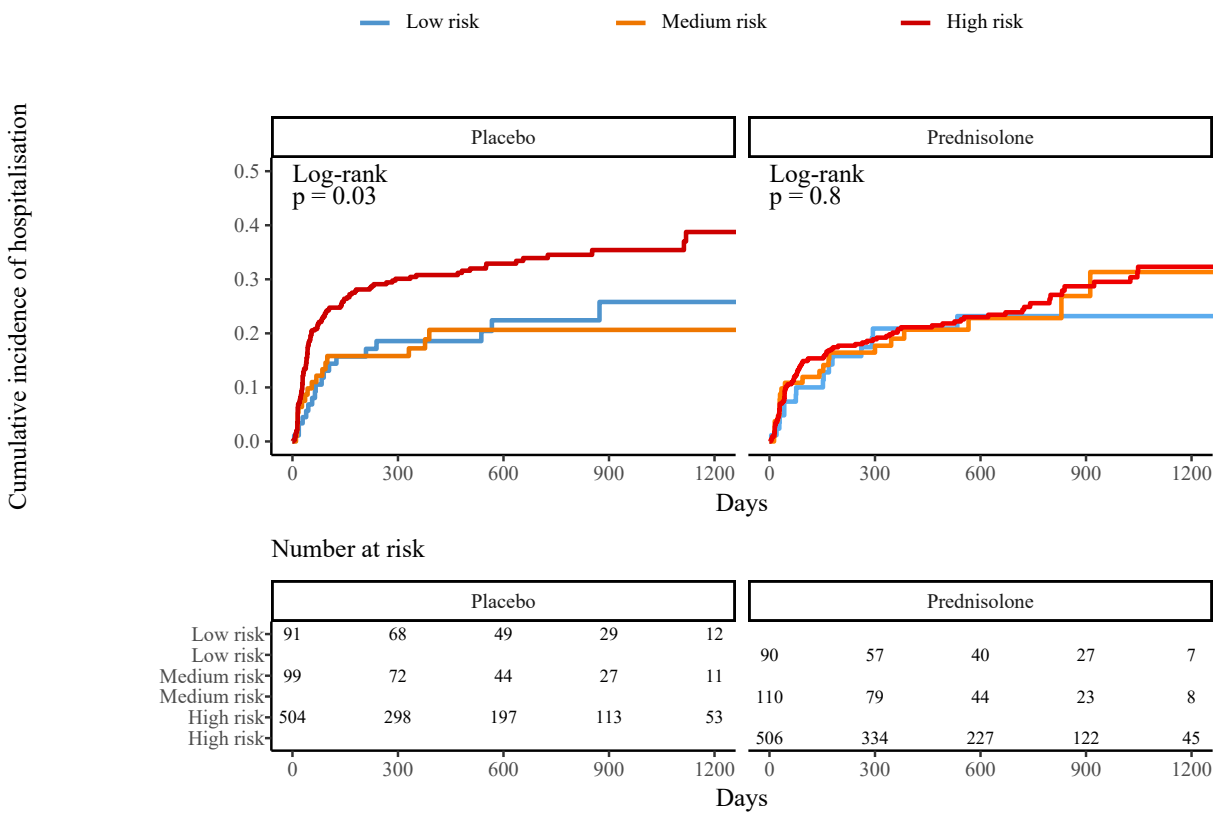

Figure D.3: Kaplan-Meier estimates of the cumulative incidence of hospitalisation as a first event stratified according to the risk of CP and treatment.

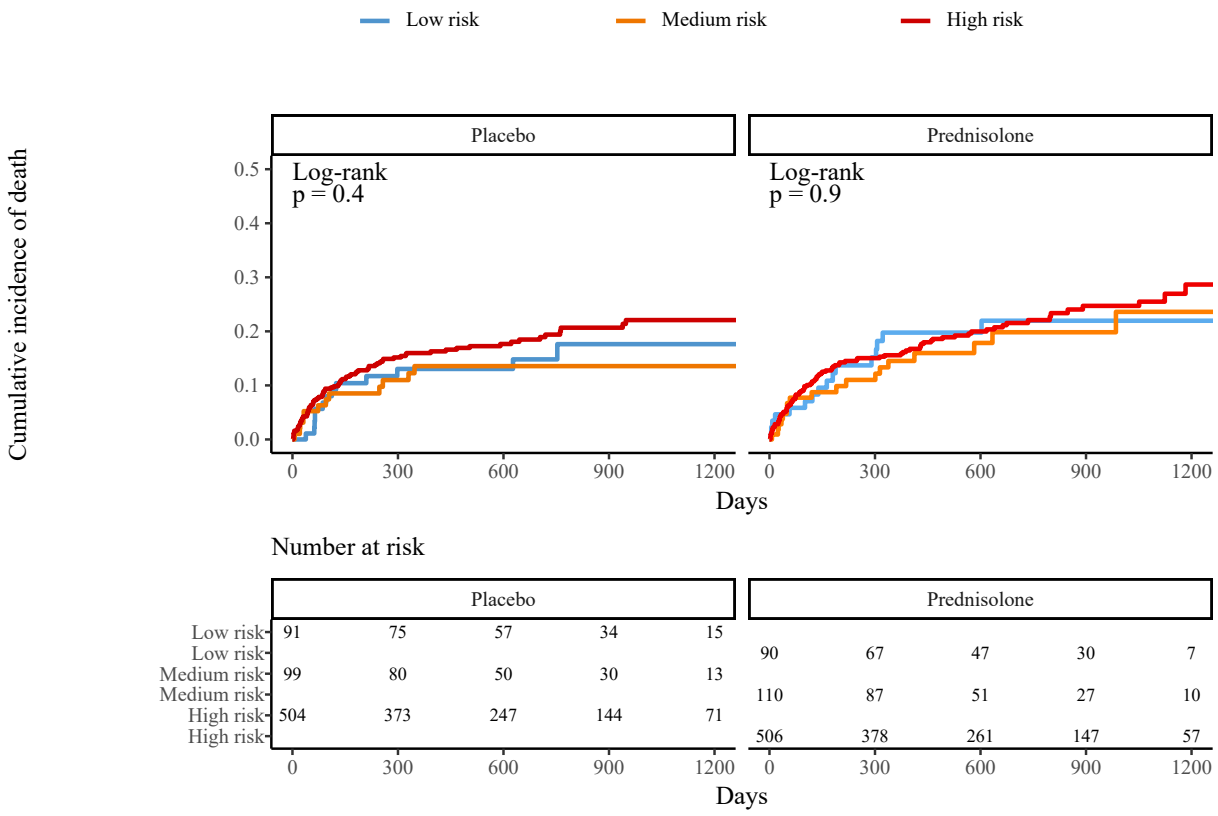

Figure D.4: Kaplan-Meier estimates of the cumulative incidence of death as a first event stratified according to the risk of CP and treatment.

Table D.4: Summary of Cox PH estimates used to predict the relative hazard of CP with hospitalisation considered as an independent event. This model includes the predicted risk of CP and randomised treatment and the AIC = 942.77.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	0.45	1.22	0.041 - 4.987	0.517
High vs low risk	6.58	0.72	1.611 - 26.856	0.009
Prednisolone vs placebo	0.52	0.25	0.320 - 0.855	0.010

Table D.5: Summary of the Cox PH estimates used to predict the relative hazard of hospitalisation with CP considered as an independent event. This model includes the predicted risk of CP and randomised treatment and the AIC = 4023.24.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	1	0.24	0.631 - 1.593	0.999
High vs low risk	1.27	0.18	0.883 - 1.820	0.198
Prednisolone vs placebo 0 - 21 days	0.58	0.25	0.355 - 0.956	0.033
Prednisolone vs placebo 22 days - 2 years	0.84	0.14	0.637 - 1.099	0.20
Prednisolone vs placebo 2 - 5 years	2.74	0.58	0.873 - 8.624	0.084

Table D.6: Summary of Cox PH estimates used to predict the relative hazard of CP with death considered as a competing event. This model includes the predicted risk of CP and randomised treatment and the AIC = 1173.13.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	0.45	1.22	0.041 - 4.932	0.511
High vs low risk	7.86	0.72	1.934 - 31.977	0.004
Prednisolone vs placebo	0.56	0.23	0.358 - 0.867	0.010

Table D.7: Summary of Cox PH estimates used to predict the relative hazard of death with CP considered as an independent event. This model includes the predicted risk of CP and randomised treatment and the AIC = 3090.93.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	0.97	0.26	0.584 - 1.623	0.917
High vs low risk	1.13	0.20	0.754 - 1.681	0.564
Prednisolone vs placebo	1.20	0.13	0.923 - 1.566	0.172

Table D.8: Summary of subdistribution hazard estimates used to predict the relative hazard of CP given that hospitalisation can occur. This model includes the predicted risk of CP and randomised treatment and the AIC = 960.41.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	0.44	1.22	0.040 - 4.877	0.510
High vs low risk	6.20	0.72	1.511 - 25.401	0.011
Prednisolone vs placebo	0.56	0.25	0.340 - 0.907	0.019

Table D.9: Summary of subdistribution hazard estimates used to predict the relative hazard of hospitalisation given that CP can occur. This model includes the predicted risk of CP, randomised treatment and the AIC = 4042.06.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	1	0.23	0.634 - 1.588	0.990
High vs low risk	1.21	0.18	0.849 - 1.731	0.290
Prednisolone vs placebo	0.68	0.14	0.510 - 0.899	0.007
Prednisolone vs placebo:time	1.001	0.0006	1 - 1.003	0.016

Table D.10: Summary of subdistribution hazard estimates used to predict the relative hazard of CP given that death can occur. This model includes the predicted risk of CP and randomised treatment and the AIC = 1184.06.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	0.44	1.23	0.040 - 4.886	0.510
High vs low risk	7.691	0.72	1.885 - 31.381	0.005
Prednisolone vs placebo	0.56	0.23	0.357 - 0.862	0.009

Table D.11: Summary of subdistribution hazard estimates used to predict the relative hazard of death given that CP can occur. This model includes the predicted risk of CP and randomised treatment and the AIC = 3111.64.

Effect	Hazard ratio	Standard error	95% CI	p-value
Medium vs low risk	0.98	0.26	0.586 - 1.630	0.930
High vs low risk	1.07	0.21	0.714 - 1.590	0.750
Prednisolone vs placebo	1.24	0.14	0.951 - 1.610	0.110

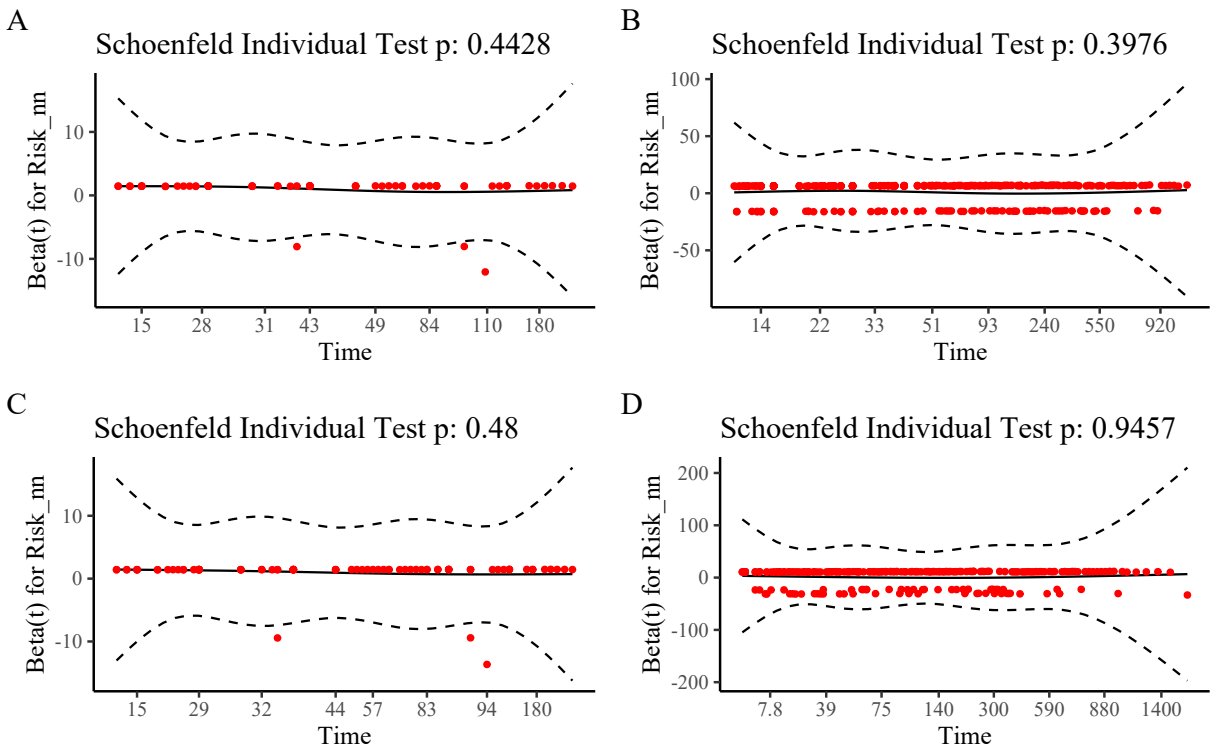

Figure D.5: Test for proportional hazards of risk of CP using Schoenfeld residuals as a function of time in the final models for each cause-specific outcome. A: CP with hospitalisation as an independent event, B: hospitalisation CP as an independent event, C: CP with death an independent event, D: death with CP as an independent event.

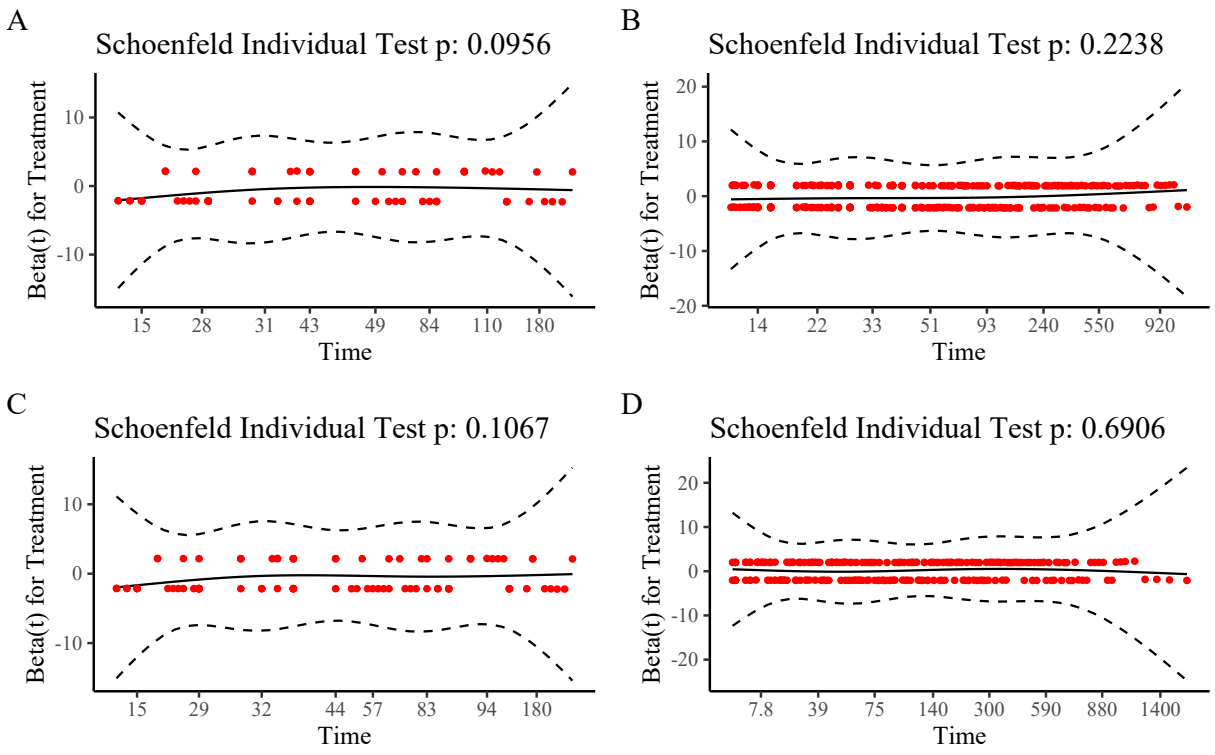

Figure D.6: Test for proportional hazards of treatment using Schoenfeld residuals as a function of time in the final models for each cause-specific outcome. A: CP with hospitalisation as an independent event, B: hospitalisation CP as an independent event, C: CP with death an independent event, D: death with CP as an independent event.

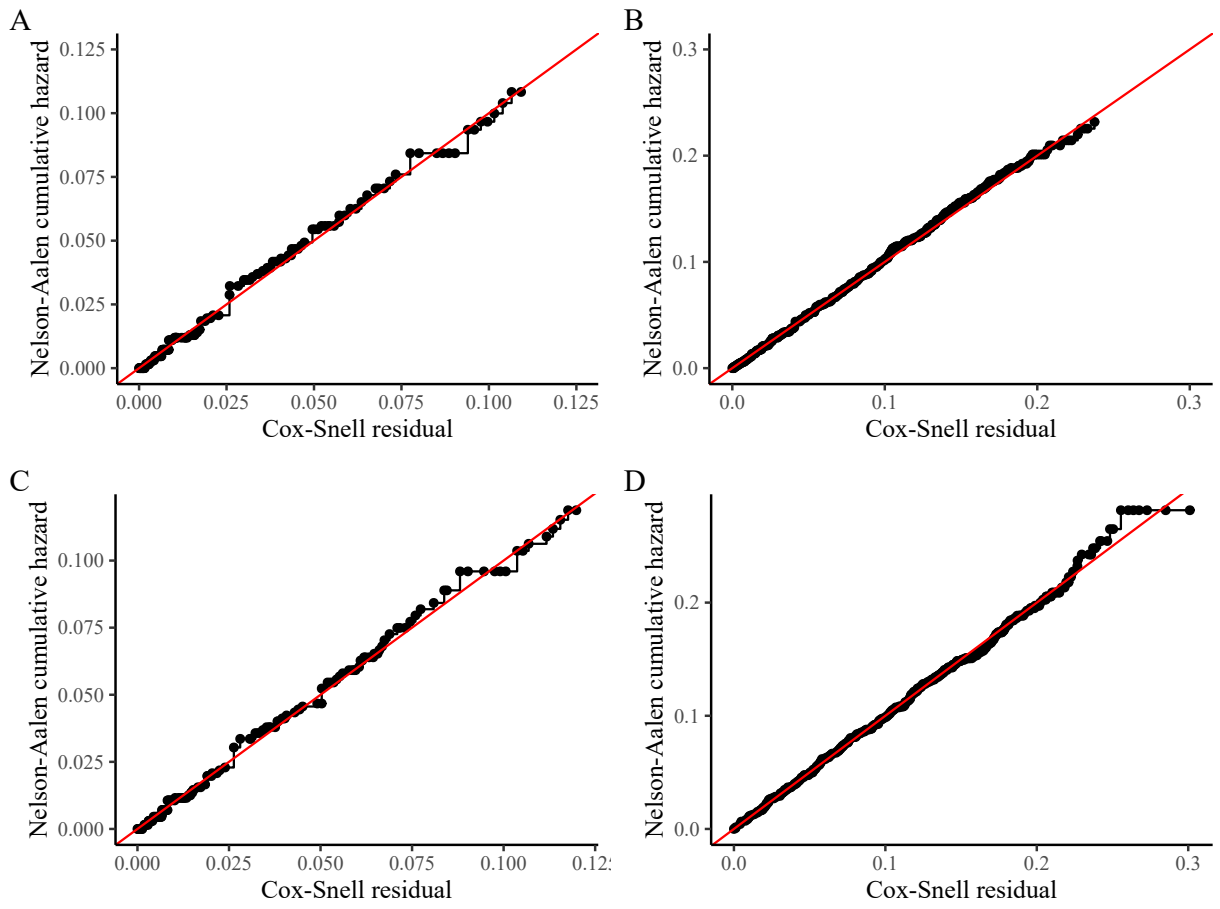

Figure D.7: Assessment of the overall fit of the cause-specific Cox proportional hazard model using Cox-Snell residuals for each cause-specific outcome. A: CP with hospitalisation as an independent event, B: hospitalisation CP as an independent event, C: CP with death an independent event, D: death with CP as an independent event.

Bibliography

1. Murray, J. F. A century of tuberculosis. *American journal of respiratory and critical care medicine* **169,** 1181–1186 (2004).

2. *P0309.3 - Mortality and causes of death in South Africa: Findings from death notification, 2018* http://www.statssa.gov.za/?page_id=1854&PPN=P0309.3&SCH=7923. (Accessed: 2021-11-28 15:45:22).

3. Mayosi, B. M. Contemporary trends in the epidemiology and management of cardiomyopathy and pericarditis in sub-Saharan Africa. *Heart* **93,** 1176–1183 (2007).

4. Strang, J. *et al.* Controlled clinical trial of complete open surgical drainage and of prednisolone in treatment of tuberculous pericardial effusion in Transkei. *The Lancet* **332,** 759–764 (1988).

5. Mayosi, B. M. *et al.* Mortality in patients treated for tuberculous pericarditis in sub-Saharan Africa. *South African Medical Journal* **98,** 36–40 (2008).

6. *South Africa* https://www.unaids.org/en/regionscountries/countries/southafrica. (Accessed: 2020-12-28 14:16:57).

7. Mayosi, B. M., Burgess, L. J. & Doubell, A. F. Tuberculous pericarditis. *Circulation* **112,** 3608–3616 (2005).

8. Strang, J. *et al.* Management of tuberculous constrictive pericarditis and tuberculous pericardial effusion in Transkei: results at 10 years follow-up. *Qjm* **97,** 525–535 (2004).

9. Mayosi, B. M. Interventions for treating tuberculous pericarditis. *Cochrane database of systematic reviews* (2002).

10. Tsang, T. S. *et al. Consecutive 1127 therapeutic echocardiographically guided pericardiocenteses: clinical profile, practice patterns, and outcomes spanning 21 years* in *Mayo Clinic Proceedings* **77** (2002), 429–436.

11. Nguyen, C. T., Lee, E., Luo, H. & Siegel, R. J. Echocardiographic guidance for diagnostic and therapeutic percutaneous procedures. *Cardiovascular diagnosis and therapy* **1,** 11–36 (2011).

12. Reuter, H., Burgess, L. J., Louw, V. J. & Doubell, A. F. The management of tuberculous pericardial effusion: experience in 233 consecutive patients (2007).

13. Isiguzo, G., Du Bruyn, E., Howlett, P. & Ntsekhe, M. Diagnosis and Management of Tuberculous Pericarditis: What Is New? *Current Cardiology Reports* **22,** 1–8 (2020).

14. Kim, K. H. *et al.* Effusive-constrictive pericarditis after pericardiocentesis: incidence, associated findings, and natural history. *JACC: Cardiovascular Imaging* **11,** 534–541 (2018).

15. Syed, F. F., Ntsekhe, M., Mayosi, B. M. & Oh, J. K. Effusive-constrictive pericarditis. *Heart failure reviews* **18,** 277–287 (2013).

16. Strang, J. Tuberculous pericarditis. *Journal of Infection* **35,** 215–219 (1997).

17. Strang, J. *et al.* Controlled trial of prednisolone as adjuvant in treatment of tuberculous constrictive pericarditis in Transkei. *The Lancet* **330,** 1418–1422 (1987).

18. Mayosi, B. M. *et al.* Prednisolone and Mycobacterium indicus pranii in tuberculous pericarditis. *New England Journal of Medicine* **371,** 1121–1130 (2014).

19. Ntsekhe, M. *et al.* HIV infection is associated with a lower incidence of constriction in presumed tuberculous pericarditis: a prospective observational study. *PLoS One* **3,** e2253 (2008).

20. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning* (Springer, 2013).

21. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **6,** 1–21 (2012).

22. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118 (2012).

23. Nakagawa, S. & Freckleton, R. P. Missing inaction: the dangers of ignoring missing data. *Trends in ecology & evolution* **23,** 592–596 (2008).

24. Little, R. J. & Rubin, D. B. *Statistical analysis with missing data* (John Wiley & Sons, 2019).

25. Breiman, L. Random forests. *Machine learning* **45,** 5–32 (2001).

26. Sun, Y., Wong, A. K. & Kamel, M. S. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence* **23,** 687–719 (2009).

27. Liu, Y., Yu, X., Huang, J. X. & An, A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing & Management* **47,** 617–631 (2011).

28. Al-Rifaie, M. M. & Alhakbani, H. A. *Handling class imbalance in direct marketing dataset using a hybrid data and algorithmic level solutions* in *2016 SAI Computing Conference (SAI)* (2016), 446–451.

29. Rahman, M. M. & Davis, D. N. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing* **3,** 224 (2013).

30. Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H. & Santos, J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. *ieee ComputatioNal iNtelligeNCe magaziNe* **13,** 59–76 (2018).

31. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16,** 321–357 (2002).

32. Steyerberg, E. W. *et al. Clinical prediction models* (Springer, 2019).

33. Moons, K. G. *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart* **98,** 683–690 (2012).

34. Van Diepen, M., Ramspek, C. L., Jager, K. J., Zoccali, C. & Dekker, F. W. Prediction versus aetiology: common pitfalls and how to avoid them. *Nephrology Dialysis Transplantation* **32,** ii1–ii5 (2017).

35. Bellazzi, R. & Zupan, B. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics* **77,** 81–97 (2008).

36. Wyatt, J. C. & Altman, D. G. Commentary: Prognostic models: clinically useful or quickly forgotten? *Bmj* **311,** 1539–1541 (1995).

37. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* **21,** 128 (2010).

38. Wijeysundera, D. N. Predicting outcomes: is there utility in risk scores? *Canadian Journal of Anesthesia/Journal canadien d'anesthésie* **63,** 148–158 (2016).

39. Zhang, L., Wang, Y., Niu, M., Wang, C. & Wang, Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. *Scientific reports* **10,** 1–10 (2020).

40. Mostavi, M., Chiu, Y.-C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression. *BMC medical genomics* **13,** 1–13 (2020).

41. O'Brien, T. R. *et al.* An IL28B genotype-based clinical prediction model for treatment of chronic hepatitis C. *PloS one* **6,** e20904 (2011).

42. Bouwmeester, W. *et al.* Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* **9,** e1001221 (2012).

43. Siontis, G. C., Tzoulaki, I., Castaldi, P. J. & Ioannidis, J. P. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of clinical epidemiology* **68,** 25–34 (2015).

44. Brehaut, J. C., Stiell, I. G. & Graham, I. D. Will a new clinical decision rule be widely used? The case of the Canadian C-spine rule. *Academic emergency medicine* **13,** 413–420 (2006).

45. Dekker, F. W., Ramspek, C. L. & van Diepen, M. Con: Most clinical risk scores are useless. *Nephrology Dialysis Transplantation* **32,** 752–755 (2017).

46. Truett, J., Cornfield, J. & Kannel, W. A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of chronic diseases* **20,** 511–524 (1967).

47. *Cardiovascular Diseases* `https://www.who.int/westernpacific/health-topics/cardiovascular-diseases`. (Accessed: 2021-05-02 12:20:17).

48. Gordon, T. & Kannel, W. B. Multiple risk functions for predicting coronary heart disease: the concept, accuracy, and application. *American heart journal* **103,** 1031–1039 (1982).

49. Brindle, P., Beswick, A., Fahey, T. & Ebrahim, S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart* **92,** 1752–1759 (2006).

50. Levy, W. C. *et al.* The Seattle heart failure model. *Circulation* **113,** 1424–1433 (2006).

51. Lloyd-Jones, D. M. Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation* **121,** 1768–1777 (2010).

52. Pocock, S. J. *et al.* Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *European heart journal* **34,** 1404–1413 (2013).

53. Brindle, P. *et al.* Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. *BMJ* **327,** 1267 (2003).

54. Unnikrishnan, P. *et al.* Development of health parameter model for risk prediction of CVD using SVM. *Computational and mathematical methods in medicine* **2016** (2016).

55. Austin, P. C., Tu, J. V., Ho, J. E., Levy, D. & Lee, D. S. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology* **66,** 398–407 (2013).

56. Dimopoulos, A. C. *et al.* Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC medical research methodology* **18,** 1–11 (2018).

57. Dai, W. *et al.* Prediction of hospitalization due to heart diseases by supervised learning methods. *International journal of medical informatics* **84,** 189–197 (2015).

58. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one* **12,** e0174944 (2017).

59. Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D. & Rakowski, W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine* **26,** 172–181 (2003).

60. Austin, P. C., Lee, D. S., Steyerberg, E. W. & Tu, J. V. Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biometrical journal* **54,** 657–673 (2012).

61. Wu, J., Roy, J. & Stewart, W. F. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care,* S106–S113 (2010).

62. Carugo, O., Eisenhaber, F. & Carugo. *Data mining techniques for the life sciences* (Springer, 2010).

63. Maroco, J. *et al.* Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes* **4,** 1–14 (2011).

64. Panahiazar, M., Taslimitehrani, V., Pereira, N. & Pathak, J. Using EHRs and machine learning for heart failure survival analysis. *Studies in health technology and informatics* **216,** 40 (2015).

65. Gibson, W. J. *et al.* Machine learning versus traditional risk stratification methods in acute coronary syndrome: a pooled randomized clinical trial analysis. *Journal of thrombosis and thrombolysis* **49,** 1–9 (2020).

66. Schperberg, A. V., Boichard, A., Tsigelny, I. F., Richard, S. B. & Kurzrock, R. Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials. *International journal of cancer* **147,** 2537–2549 (2020).

67. Venkatesan, P. & Yamuna, N. Treatment response classification in randomized clinical trials: a decision tree approach. *Indian Journal of Science and Technology* **6,** 3912–3917 (2013).

68. Lo, A. W., Siah, K. W. & Wong, C. H. Machine learning with statistical imputation for predicting drug approvals. *Available at SSRN 2973611* (2018).

69. DiMasi, J. *et al.* A tool for predicting regulatory approval after phase II testing of new oncology compounds. *Clinical Pharmacology & Therapeutics* **98,** 506–513 (2015).

70. Lacson, R. C. *et al.* Use of machine-learning algorithms to determine features of systolic blood pressure variability that predict poor outcomes in hypertensive patients. *Clinical kidney journal* **12,** 206–212 (2019).

71. Dobson, A. J. & Barnett, A. G. *An introduction to generalized linear models* (CRC press, 2018).

72. Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied logistic regression* (John Wiley & Sons, 2013).

73. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (1996).

74. Friedman, J., Hastie, T., Tibshirani, R., *et al. The elements of statistical learning* **10** (Springer series in statistics New York, 2001).

75. Han, J., Kamber, M. & Pei, J. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems* **5,** 83–124 (2011).

76. Goeman, J., Meijer, R. & Chaturvedi, N. L1 and L2 penalized regression models. *cran. r-project. or* (2012).

77. Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and regression trees* (CRC press, 1984).

78. Breiman, L. Bagging predictors. *Machine learning* **24,** 123–140 (1996).

79. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *Journal of Animal Ecology* **77,** 802–813 (2008).

80. Schapire, R. E. The strength of weak learnability. *Machine learning* **5,** 197–227 (1990).

81. Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20,** 273–297 (1995).

82. Boser, B. E., Guyon, I. M. & Vapnik, V. N. *A training algorithm for optimal margin classifiers* in *Proceedings of the fifth annual workshop on Computational learning theory* (1992), 144–152.

83. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65,** 386 (1958).

84. Basheer, I. A. & Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods* **43,** 3–31 (2000).

85. Nielsen, M. A. *Neural networks and deep learning* (Determination press San Francisco, CA, 2015).

86. Candel, A., Parmar, V., LeDell, E. & Arora, A. Deep learning with H2O. *H2O. ai Inc,* 1–21 (2016).

87. Maxim, L. D., Niebo, R. & Utell, M. J. Screening tests: a review with examples. *Inhalation toxicology* **26,** 811–828 (2014).

88. Davis, J. & Goadrich, M. *The relationship between Precision-Recall and ROC curves* in *Proceedings of the 23rd international conference on Machine learning* (2006), 233–240.

89. Boyd, K., Costa, V. S., Davis, J. & Page, C. D. *Unachievable region in precision-recall space and its effect on empirical evaluation* in *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning* **2012** (2012), 349.

90. Masaisa, F., Gahutu, J. B., Mukiibi, J., Delanghe, J. & Philippé, J. Anemia in human immunodeficiency virus–infected and uninfected women in Rwanda. *The American journal of tropical medicine and hygiene* **84,** 456 (2011).

91. Nurminen, M., Nurminen, T. & Corvalan, C. F. Methodologic issues in epidemiologic risk assessment. *Epidemiology,* 585–593 (1999).

92. Geskus, R. B. *Data analysis with competing risks and intermediate states* (CRC Press, 2015).

93. Kleinbaum, D. G. & Klein, M. *Survival analysis* (Springer, 2010).

94. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53,** 457–481 (1958).

95. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202 (1972).

96. Aalen, O. Nonparametric inference for a family of counting processes. *The Annals of Statistics,* 701–726 (1978).

97. Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics* **14,** 945–966 (1972).

98. Efron, B. The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association* **72,** 557–565 (1977).

99. Lemeshow, S. & May, S. *Applied survival analysis: regression modeling of time-to-event data* (Wiley, 2008).

100. Beyersmann, J., Allignol, A. & Schumacher, M. *Competing risks and multistate models with R* (Springer Science & Business Media, 2011).

101. Putter, H., Fiocco, M. & Geskus, R. B. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine* **26,** 2389–2430 (2007).

102. Prentice, R. L. *et al.* The analysis of failure times in the presence of competing risks. *Biometrics,* 541–554 (1978).

103. Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* **94,** 496–509 (1999).

104. Gray, B. *cmprsk: Subdistribution Analysis of Competing Risks* R package version 2.2-10 (2020). `https://CRAN.R-project.org/package=cmprsk`.

105. Alvares, D., Haneuse, S., Lee, C. & Lee, K. H. SemiCompRisks: an R package for independent and Cluster-Correlated analyses of Semi-Competing risks data. *arXiv preprint arXiv:1801.03567* (2018).

106. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software* **39,** 1–13 (2011).

107. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The annals of applied statistics* **2,** 841–860 (2008).

108. Biganzoli, E., Boracchi, P., Mariani, L. & Marubini, E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine* **17,** 1169–1186 (1998).

109. Lee, C., Zame, W. R., Yoon, J. & van der Schaar, M. *Deephit: A deep learning approach to survival analysis with competing risks* in *Thirty-second AAAI conference on artificial intelligence* (2018).

110. Chen, H.-C., Kodell, R. L., Cheng, K. F. & Chen, J. J. Assessment of performance of survival prediction models for cancer prognosis. *BMC medical research methodology* **12,** 1–11 (2012).