Prediction models for mortality in tuberculous meningitis



Le Thi Phuong Thao

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2019

Abstract

Tuberculous meningitis (TBM) is the most severe manifestation of tuberculosis, killing or disabling nearly half of all sufferers despite proper anti-tuberculosis treatment. The British Medical Research Council (MRC) grade, which was first introduced in 1948, is the most widely used grading system for TBM severity. Despite its age and its lack of statistical derivation, improved and robust prediction models for poor outcomes in TBM based on large cohort studies and rigorous statistical methodology are still lacking. In addition, no studies have exploited repeated measurements to update the risk prediction.

The purpose of my DPhil thesis is to develop prognostic models for mortality in adult TBM patients and to address statistical issues that may arise during the modelling process. Firstly, I present prognostic models for 9-month mortality based on patient characteristics at diagnosis. Separate models were constructed for patients with and without HIV co-infection. The models were developed and validated based on a large pooled database of 748 HIV-infected and 951 HIV-uninfected subjects. This is the largest database of prospectively characterized TBM patients to date. The final models showed good performance and are clearly superior to the MRC grading system. However, predictions may be improved and updated once new information on the disease progress becomes available during follow-up. Therefore, in the second part of my thesis, I present prognostic models that dynamically predict death using time-updated Glasgow coma score and plasma sodium measurements, together with patient baseline characteristics. The models were carefully developed and validated based on 518 HIV-infected and 550 HIV-uninfected patients respectively. They outperformed the baseline models. I implemented the baseline models and the dynamic models in web-based calculators to facilitate their application in practice. These prognostic models can be used to assist clinicians in decision-making regarding treatment and patient

management. Lastly, I describe and compare strategies for incorporating variable selection with multiply imputed data in the procedure of building a prediction model. I evaluated the performance of the methods on simulated data, and on the same data set of 951 HIV-uninfected patients that I used in the first part. I provide insights on the strengths and weaknesses of the pragmatic approaches that are commonly used.

Acknowledgements

In this very early page of my thesis, I would like to express my sincere gratitude to many people without whom I would have not completed this challenging, yet rewarding, DPhil journey.

First and foremost, I would like to thank my primary supervisor, associate professor Ronald Geskus, for helping me get my thesis across the finishing line. I am grateful for his tremendous support, guidance, patience, and dedication that have empowered me to pursue my research ideas. He taught me to be critical, encouraged me to step out of my comfort zone, and to challenge myself every day in order to become a strong and independent researcher. I also wholeheartedly thank Dr. Marcel Wolbers, my co-supervisor, for his trust in giving me the great opportunity to do research at OUCRU, and for his invaluable guidance and encouragement in the early days when I was a very inexperienced statistician. His work ethic is a great example for me, not only during the DPhil period but also in my future career. Receiving a "happy birthday" recording from them, sent from a statistical conference, is one of my fondest memories. I am grateful to Professor Guy Thwaites, my co-supervisor, for his constructive and insightful inputs. He gave me a chance to see real world problems, reminded me about the meaning of my work, and made me believe that I could make a difference.

I would like to express my sincere thanks to my colleagues in the biostatistics group: Dr. Le Thanh Hoang Nhat, Dr. Phung Khanh Lam, and Dr. Nguyen Lam Vuong for their useful advice and suggestions, both in academia and in various aspects of life that helped me cope with difficult and stressful times through these past few years. A special thanks to Dr. Nguyen Thuy Thuong Thuong and Dr. Hannah Clapham for their advice, concern and encouragement.

I thank my younger sisters Le Hoang Phuong Nhi and Le Hoang Nguyen Phuong for being my constant sources of joy, and for taking care of my

parents whilst I was away. Finally, I would like to thank my parents for their unconditional love and trust.

List of publications during Dphil candidature:

As first author:

1. **Le Thi Phuong Thao** and Ronald Geskus. A comparison of model selection methods for prediction in the presence of multiply imputed data. *Biometrical Journal*, oct 2018. ISSN 03233847. doi: 10.1002/bimj.201700232. URL `http://doi.wiley.com/10.1002/bimj.201700232`.
2. **Le Thi Phuong Thao**, Ronald Geskus, and Guy Thwaites. Reply to Dian et al. *Clinical Infectious Diseases*, 67(12):1955, nov 2018a. ISSN 1058-4838. URL `http://dx.doi.org/10.1093/cid/ciy447`.
3. **Le Thi Phuong Thao**, A Dorothee Heemskerk, Ronald B Geskus, Nguyen Thi Hoang Mai, Dang Thi Minh Ha, Tran Thi Hong Chau, Nguyen Hoan Phu, Nguyen Van Vinh Chau, Maxine Caws, Nguyen Huu Lan, Do Dang Anh Thu, Nguyen Thuy Thuong Thuong, Jeremy Day, Jeremy J Farrar, M Estee Torok, Nguyen Duc Bang, Guy E Thwaites, and Marcel Wolbers. Prognostic Models for 9-Month Mortality in Tuberculous Meningitis. *Clinical Infectious Diseases*, 66 (4):523532, feb 2018b. ISSN 1058-4838. doi: 10.1093/cid/cix849. URL `https://academic.oup.com/cid/article/66/4/523/4237651`

As co-author:

1. Fiona V. Cresswell, Ananta S. Bangdiwala, David B. Meya, Nathan C. Bahr, Jose E. Vidal, M. Estee Torok, **Le Thi Phuong Thao**, Guy E. Thwaites, and David R. Boulware. Absence of cerebrospinal fluid pleocytosis in tuberculous meningitis is a common occurrence in HIV co-infection and a predictor of poor outcomes. *International Journal of Infectious Diseases*, 68:7778, mar 2018. ISSN 12019712. doi: 10.1016/j.ijid.2018.01.014. URL `https://linkinghub.elsevier.com/retrieve/pii/S1201971218300158`.
2. Joseph Donovan, Nguyen Hoan Phu, Nguyen Thi Hoang Mai, Le Tien Dung, Darma Imran, Erlina Burhan, Lam Hong Bao Ngoc, Nguyen Duc Bang, Do Chau Giang, Dang Thi Minh Ha, Jeremy Day, **Le Thi Phuong Thao**, Nguyen TT Thuong, Nguyen Nang Vien, Ronald B. Geskus, Marcel Wolbers, Raph L Hamers, Reinout

van Crevel, Mugi Nursaya, Kartika Maharani, Tran Tinh Hien, Kevin Baird, Nguyen Huu Lan, Evelyne Kestelyn, Nguyen Van Vinh Chau, and Guy E. Thwaites. Adjunctive dexamethasone for the treatment of HIV-infected adults with tuberculous meningitis (ACT HIV): Study protocol for a randomised controlled trial. *Wellcome Open Research*, 3:31, jun 2018a. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.14006.2. URL `https://wellcomeopenresearch.org/articles/3-31/v2`

3. Joseph Donovan, Nguyen Hoan Phu, **Le Thi Phuong Thao**, Nguyen Huu Lan, Nguyen Thi Hoang Mai, Nguyen Thi Mai Trang, Nguyen Thi Thu Hiep, Tran Bao Nhu, Bui Thi Bich Hanh, Vu Thi Phuong Mai, Nguyen Duc Bang, Do Chau Giang, Dang Thi Minh Ha, Jeremy Day, Nguyen TT Thuong, Nguyen Nang Vien, Ronald B. Geskus, Tran Tinh Hien, Evelyne Kestelyn, Marcel Wolbers, Nguyen Van Vinh Chau, and Guy E. Thwaites. Adjunctive dexamethasone for the treatment of HIV-uninfected adults with tuberculous meningitis stratified by Leukotriene A4 hydrolase genotype (LAST ACT): Study protocol for a randomised double blind placebo controlled non-inferiority trial. *Wellcome Open Research*, 3: 32, mar 2018b. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.14007.1. URL `https://wellcomeopenresearch.org/articles/3-32/v1`.

4. A. Dorothee Heemskerk, Mai Thi Hoang Nguyen, Ha Thi Minh Dang, Chau Van Vinh Nguyen, Lan Huu Nguyen, Thu Dang Anh Do, Thuong Thuy Thuong Nguyen, Marcel Wolbers, Jeremy Day, **Thao Thi Phuong Le**, Bang Duc Nguyen, Maxine Caws, and Guy E. Thwaites. Clinical Outcomes of Patients with Drug-Resistant Tuberculous Meningitis Treated with an Intensified Antituberculosis Regimen. *Clinical Infectious Diseases*, 65(1):2028, 2017. ISSN 15376591. doi: 10.1093/cid/cix230.

5. Nguyen TH Mai, Nicholas Dobbs, Nguyen Hoan Phu, Romain A Colas, **Le TP Thao**, Nguyen TT Thuong, Ho DT Nghia, Nguyen HH Hanh, Nguyen T Hang, A Dorothee Heemskerk, Jeremy N Day, Lucy Ly, Do DA Thu, Laura Merson, Evelyne Kestelyn, Marcel Wolbers, Ronald Geskus, David Summers, Nguyen VV Chau, Jesmond Dalli, and Guy E Thwaites. A randomised double blind placebo

controlled phase 2 trial of adjunctive aspirin for tuberculous meningitis in HIV uninfected adults. eLife, 7:120, 2018. ISSN 2050-084X. doi: 10.7554/eLife. 33478. URL http://www.ncbi.nlm.nih.gov/pubmed/29482717

6. Nguyen T T Thuong, Dorothee Heemskerk, Trinh T B Tram, **Le T P Thao**, Lalita Ramakrishnan, Vu T N Ha, Nguyen D Bang, Tran T H Chau, Nguyen H Lan, Maxine Caws, Sarah J Dunstan, Nguyen V V Chau, Marcel Wolbers, Nguyen T H Mai, and Guy E Thwaites. Leukotriene A4 Hydrolase Genotype and HIV Infection Influence Intracerebral Inflammation and Survival From Tuberculous Meningitis. *The Journal of Infectious Diseases*, 215(7):10201028, apr 2017. ISSN 0022-1899. URL http://dx.doi.org/10.1093/infdis/jix050.

"All models are wrong, but some are useful"

- George Box, *1987*

Contents

1 Introduction **1**

1.1 Overview of tuberculous meningitis 1

1.1.1 Epidemiology 2

1.1.2 Pathology 2

1.1.3 Clinical presentation and complication 3

1.1.4 Diagnostic 5

1.1.5 Management of TBM 7

1.2 Introduction to prognostic models 10

1.2.1 What is a prognostic model and why is it important 10

1.2.2 Source of data and treatment use 13

1.2.3 Baseline and dynamic prediction 14

1.3 Variable selection in the presence of multiply imputed data 17

1.3.1 Missing data 17

1.3.2 Variable selection for fully observed data 18

1.3.3 Variable selection with multiply imputed data 19

1.3.4 Prognostic models in tuberculous meningitis: Literature review 20

1.3.4.1 Prognostic model for TBM 21

1.3.4.2 Other articles describing risk factors for poor outcome 22

1.4 Objectives of the thesis 26

2 Materials and common statistical methods in prediction **27**

2.1 Materials 27

2.1.1 Study populations for TBM prediction models 27

2.1.2 Study procedure 39

2.1.3 Laboratory investigations 41

2.1.4 Creation of the pooled database 41

2.2 Common statistical methods in prediction 41

2.2.1 Model development . . . 41
2.2.1.1 Treatment of missing data . . . 41
2.2.1.2 Statistical model . . . 43
2.2.2 Model validation . . . 43
2.2.2.1 Model performance . . . 43
2.2.2.2 Model validation . . . 45
2.2.2.3 Model presentation . . . 46
2.2.2.4 Statistical software . . . 47

3 Prognostic models for 9 month mortality in tuberculous meningitis 48
3.1 Introduction . . . 48
3.2 Methods . . . 49
3.2.1 Study population . . . 49
3.2.2 Primary outcome . . . 49
3.2.3 Candidate predictors . . . 49
3.2.4 Statistical analysis . . . 51
3.2.4.1 Transformation of covariates . . . 51
3.2.4.2 Treatment of missing data . . . 52
3.2.4.3 Model development, validation, and assessment . . . 53
3.3 Results . . . 54
3.3.1 Baseline characteristics of study participants . . . 54
3.3.2 Prognostic models . . . 59
3.3.3 Model validation . . . 67
3.3.4 Comparison of prognostic models with MRC grade and GCS alone . . . 71
3.3.5 Nomograms and web app for mortality prediction . . . 74
3.4 Discussion . . . 80
3.5 Appendix . . . 82

4 Dynamic prediction of death in patients with tuberculous meningitis using time-updated Glasgow coma score and plasma sodium measurements 87
4.1 Introduction . . . 87
4.2 Methods . . . 88
4.2.1 Study population . . . 88
4.2.1.1 Study participants . . . 88
4.2.1.2 Primary outcome . . . 88

4.2.1.3 Predictors . . . 89
4.2.2 Statistical analysis . . . 90
4.2.2.1 Transformation of covariates . . . 90
4.2.2.2 Statistical models for prediction . . . 90
4.2.2.3 Extra analyses and alternative modelling approach . . . 92
4.2.2.4 Model performance and validation . . . 93
4.2.2.5 Statistical software . . . 94
4.3 Results . . . 94
4.3.1 Baseline characteristics of study participants . . . 94
4.3.2 Longitudinal profile of GCS and plasma sodium . . . 98
4.3.3 Parameter estimates . . . 99
4.3.4 Predictive performance . . . 102
4.3.5 Model display and illustration . . . 106
4.4 Discussion . . . 110
4.5 Appendix . . . 112

5 A comparison of model selection methods for prediction in the presence of multiply imputed data 116
5.1 Introduction . . . 116
5.2 Model selection strategies . . . 117
5.2.1 Lasso models . . . 117
5.2.2 General model selection approaches . . . 119
5.3 Simulation study . . . 121
5.3.1 Simulation design . . . 121
5.3.2 Results . . . 123
5.3.2.1 Fully observed data . . . 123
5.3.2.2 Multiply imputed data . . . 126
5.4 Real data application . . . 134
5.4.1 Data description . . . 134
5.4.2 Results . . . 135
5.5 Discussion . . . 139
5.6 Appendix . . . 143

6 Conclusion **151**
6.1 Contributions of the thesis 151
6.1.1 Clinical contributions 151
6.1.2 Statistical contributions 153
6.2 Future research 154

Bibliography **157**

List of Figures

2.1 The TBM database. Annual recruitment into the six contributing studies by human immunodeficiency virus status. Red lines show estimated 9-month (8- month for the Aspirin trial) mortality rates (as percentages) for each calendar year, based on the Kaplan-Meier method. Abbreviations: ART, antiretroviral therapy; HIV, human immunodeficiency virus; TBM, tuberculous meningitis. 38

2.2 An algorithm to calculate optimism in the presence of multiply imputed data based on the classical enhanced bootstrap resampling procedure (Harrell, 2015). $*$ denotes the measurements in bootstrap samples. . . 47

3.1 Flow diagram 52

3.2 Kaplan Meier survival curves by human immunodeficiency virus (HIV) status 59

3.3 Non-linear effect of plasma sodium (mmol/L) in HIV-infected patients. The plot is generated based on one imputed data set. 69

3.4 Calibration plots of candidate models by HIV status in internal and temporal validation. The circles and lines display the predicted 9 month mortality (from the prognostic model) versus the observed 9-month month mortality (based on Kaplan-Meier estimation) in ten patient groups defined by quantiles of the predicted event probabilities. Each circle represents one quantile group. For internal validation, blue crosses display bootstrap bias-corrected Kaplan-Meier estimates. 70

3.5 Comparison between predicted survival of the final models (dashed lines) and observed Kaplan-Meier estimates (solid lines). Risk groups are defined using cut-points at the 16th, 50th, 84th percentiles of the prognostic index as generated by the final models (defining good, fairly good, fairly poor, and poor prognostic subgroups). 71

3.6 Cumulative/dynamic ROC curves (apparent estimate) for mortality evaluated at 9 months for the final prognostic model, GCS, and MRC Grade in the HIV-uninfected and HIV-infected TBM populations. . . 73
3.7 Nomograms for the prediction of 9-month mortality based on the final prognostic models for the HIV-uninfected and HIV-infected TBM populations, respectively. To derive a prediction, locate the value of each predictor on the corresponding variable line, read the corresponding points assigned on the 0-100 scale and sum all of these points to a total point score. Then read the result on the Total Points scale and its corresponding prediction below. For HIV-infected nomogram, the cohort variable was chosen as the most recent trial (Heemskerk et al., 2016) since this is most relevant for future prediction 76
3.8 Nomograms for the prediction of 9-month mortality based on the final prognostic models for the HIV-uninfected and HIV-infected TBM populations respectively including drug resistance information. To derive a prediction, locate the value of each predictor on the corresponding variable line, read the corresponding points assigned on the 0-100 scale and sum all of these points to a total point score. Then read the result on the Total Points scale and its corresponding prediction below. For HIV-infected nomogram, the cohort variable was chosen as the most recent trial (Heemskerk et al., 2016) since this is most relevant for future prediction 78
3.9 Screenshot for the web application 79

4.1 Kaplan Meier survival curves by human immunodeficiency virus (HIV) status 98
4.2 Subject specific profiles for Glasgow coma score (GCS) and plasma sodium over the first 30 days, stratified by observed survival status within 120 days. The dark blue line represents the median value per day, the blue area represents the interquartile range. Plasma sodium of HIV-uninfected patients is plotted as log2(plasma sodium - 100) with the original value as label. 99
4.3 Hazard ratio for plasma sodium at three landmark times: day 0, day 10 and day 30 after the start of TB treatment. A sodium value of 135 mmol/L was taken as the reference. The shaded areas represent the 95% confidence intervals. 100

4.4 Comparison between average predicted survival (dashed line) and Kaplan-Meier estimates (solid line) at day 0, day 10 and day 30. Groups are defined using cut-off points at the 16th, 50th, and 84th percentiles of the prognostic index as generated by the final models (defining "good", "fairly good", "fairly poor", and "poor" prognostic subgroups). . . . 104
4.5 Apparent AUC and Brier score of the landmark model (LM) and the joint model (JM) by human immunodeficiency virus (HIV) status. The prediction was carried out at day 10, 20 and 30 for mortality until day 60, 90 and 120 after diagnosis 105
4.6 Dynamic predictions of survival for patient A at four different days after the start of TB treatment. For each day, the left panels display the observed values of GCS and plasma sodium over time, the blue line represents the fitted value for sodium. The right panels represent the estimated survival probability in red with 95% confidence intervals in grey 108
4.7 Dynamic predictions of survival for patient B at four different days after the start of TB treatment. For each day, the left panels display the observed values of GCS and plasma sodium over time, the blue line represents the fitted value for sodium. The right panels represent the estimated survival probability in red with 95% confidence intervals in grey. 109
A4.1 Sensitivity analysis for sodium (HIV-uninfected): Hazard ratio for plasma sodium at three landmark times: day 0, day 10 and day 30 after the start of TB treatment. A sodium value of 135 mmol/L was taken as the reference. The shaded areas represent the 95% confidence intervals 113
A4.2 Sensitivity analysis for sodium (HIV-infected): Hazard ratio for plasma sodium at three landmark times: day 0, day 10 and day 30 after the start of TB treatment. A sodium value of 135 mmol/L was taken as the reference. The shaded areas represent the 95% confidence intervals 114
A4.3 Screenshot for the web-app 115

5.1 Fully observed data - first data generating mechanism (section 5.3.1a)): Mean AUC and Brier score over 500 fully observed generated data sets for lasso models with three loss functions for the selection of the penalty. DEV: deviance; BRIER: brier score; AUC: AUC as loss function. 124

5.2 Fully observed data - first data generating mechanism (section 5.3.1.a): Mean number of selected variables for the lasso models with three loss functions for the selection of the penalty. DEV: deviance; BRIER: brier score; AUC: AUC as loss function 125
5.3 First data generating mechanism (section 5.3.1.a): Mean Brier score (top figure) and AUC (bottom figure) over 500 generated data sets for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. 127
5.4 First data generating mechanism: The effect of the number of imputations on model performance in the two most extreme data scenarios. Brier score (top figure), AUC (bottom figure). Best case: 4% missing and 15 EPV; Worst case: 20% missing and 5 EPV 129
5.5 First data generating mechanism (section 5.3.1.a) : Mean number of selected variables for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. 130
5.6 First data generating mechanism: The effect of the number of imputations on the number of selected variables in the two most extreme data scenarios. Best case: 4% missing and 15 EPV; Worst case: 20% missing and 5 EPV 131
5.7 Second data generating mechanism (section 5.3.1.b): Mean Brier score (top figure) and AUC (bottom figure) over 500 generated data sets for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. The performance measures of BBeF in data with EPV of 3 are not available due to convergence problems in many generated data sets. 133
5.8 TBM data: Corrected AUC and Brier score via internal validation, with logistic regression models 136
5.9 TBM data: Corrected cumulative/dynamic AUC via internal validation, with Cox proportional hazard models 138
A5.1 Correlation matrix of the covariates 144
A5.2 Fully observed data - second data generating mechanism (section 5.3.1.b): Mean AUC and Brier score over 500 fully observed generated data sets for lasso models with three loss function for the selection of the penalty. DEV: deviance; BRIER: brier score; AUC: AUC as loss function. . . . 145

A5.3 First data generating mechanism: Mean Brier score and AUC over 500 generated data sets for all scenarios with 10 imputed data sets. Black bars represent the empirical standard deviation 146
A5.4 Second data generating mechanism: Number of selected variables for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. The values for BBeF in data with EPV of 3 are not available due to convergence problems in many generated data sets. . 148
A5.5 Second data generating mechanism: Mean Brier score (top figure) and AUC (bottom figure) over 500 generated data sets for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. Black bars represent the empirical standard deviation. 150

6.1 Illness-death model in TBM data. 155

List of Tables

1.1 Glasgow coma score 4
1.2 Modified British Medical Research Council (MRC) TBM severity grades 4
1.3 TBM diagnostic criteria 6
1.4 Standard regimen for adjunctive treatment with dexamethasone . . . 9
1.5 Literature review: risk factors for poor outcome in TBM 23

2.1 Summary of included studies 30
2.2 Schedule of procedures 40

3.1 Definition of candidate predictors 50
3.2 Baseline characteristics of patients included in the pooled database, overall and by HIV status 56
3.3 Univariable analyses for candidate predictors in the HIV-uninfected TBM population using Cox regression models 60
3.4 Univariable analyses for candidate predictors in the HIV-infected TBM population using Cox regression models 62
3.5 Final Cox regression models for 9-month survival in each HIV population. Estimates were pooled across multiply imputed datasets. 65
3.6 Discrimination of candidate models for TBM mortality by HIV population (with and without drug resistance information) measured by the area under the cumulative/dynamic ROC curve (AUC) at nine months. 68
A3.1 P-values for the interaction between age and other covariates. Of note, the p-values were not corrected for multiple testing. 83
A3.2 Models based on variable selection which were not chosen as final models in HIV-uninfected and HIV-infected TBM populations. 84
A3.3 Final Cox regression models for 9-month survival in each HIV population including drug resistance information. Estimates were pooled across multiply imputed datasets. 85

4.1 Baseline characteristics of patients with tuberculous meningitis included in the pooled database, overall and by human immunodeficiency virus status 96
4.2 Parameter estimates (with 95% confidence intervals) from landmark supermodel by HIV status 101
4.3 Optimism-corrected AUC and Brier score of the landmark model and the baseline model by human immunodeficiency virus (HIV) status. The prediction was carried out at day 10, 20 and 30 for mortality until day 60, 90 and 120 after diagnosis. 103
4.4 Baseline characteristic of patients A and B 106

5.1 Summary of considered model selection methods 120
5.2 Simulation scenarios 122
5.3 Variable selection in the TBM data set, all models 137
A5.1 Variable inclusion frequency (%) for the model selection on bootstrap data methods in the TBM data using logistic regression 143

Chapter 1

Introduction

Summary: I give a general introduction to tuberculous meningitis (TBM) including a review of risk factors for mortality, and survey the current status of predictive modelling in TBM. In addition, I also provide an overview of prediction models in general, and of statistical methods for model selection in the presence of multiply imputed data. I conclude this chapter by summarizing the objectives of the thesis.

1.1 Overview of tuberculous meningitis

Robert Koch (1843-1910), a German scientist and physician, discovered *Mycobacterium tuberculosis* bacteria in March 1882 (Sakula, 1983). In the opening of his lecture, he said: "If the importance of a disease for mankind is measured by the number of fatalities it causes, then tuberculosis must be considered much more important than those most feared infectious diseases, plague, cholera and the like. One in seven of all human beings dies from tuberculosis. If one only considers the productive middle-age groups, tuberculosis carries away one-third and often more". Koch's innovative discovery was a major event in medical history, which paved a way for a new era of tuberculosis (TB) treatment and control. More than a century from the day Koch discovered the bacteria, TB still remains one of the top ten leading killers worldwide (World Health Organization, 2018). Every 18 seconds someone dies because of TB across the world. In 2016, an estimated of 10.4 million people were infected with TB, and 1.7 million died from the disease (World Health Organization, 2016).

Most often TB affects the lungs (pulmonary TB) but the bacteria can be located and cause disease in almost any organ (extrapulmonary TB). When the bacteria infect the meninges, which envelop the brain and the spinal cord, they cause tuberculous meningitis (TBM).

1.1.1 Epidemiology

The exact incidence and prevalence of TBM are unknown because its diagnosis is difficult. However, the estimated global burden of TBM is at least 100.000 cases per year (Wilkinson et al., 2017). The number depends on the age structure and the seroprevalence of the human immunodeficiency virus (HIV) within a population, and is normally proportional to the pulmonary TB burden (Wilkinson et al., 2017). In high TB burden populations, TBM is more prevalent in children under 5 years of age; whereas, in low TB burden populations, the disease is mostly seen in adults. HIV is the most important determinant for the development of TBM. The proportion of HIV infected patient with TBM can exceed 50% in highly TB endemic area (Wilkinson et al., 2017). HIV is a major risk factor for TBM in individuals with *Mycobacterium tuberculosis* infection. The risk increases in patients with a lower CD4 count (Thwaites et al., 2003, 2000). Other risk factors are alcoholism, diabetes, malignancy and recent corticosteroid use (Thwaites et al., 2003; Thwaites and Tran, 2005; Thwaites et al., 2000; Wilkinson et al., 2017).

TBM is the most severe manifestation of tuberculosis and is associated with substantial mortality and morbidity. Despite proper anti-tuberculosis treatment, the mortality continues to be as high as 30-75% (Bourgi et al., 2017; Heemskerk et al., 2016), especially in children and HIV-infected patients. Survivors are often left with irreversible neurological sequelae which severely affects their quality of life (Van and Farrar, 2014). Commonly reported neurological sequelae are hearing or vision impairment, stroke, seizes (Bourgi et al., 2017), motor deficit and cognitive impairment (van Well et al., 2009).

1.1.2 Pathology

Rich and McCordock explained the pathogenesis of TBM by the well-known two-step model (Rich, 1933; Thwaites and Tran, 2005). First, *Mycobacterium tuberculosis* bacteria invade the hosts alveolar macrophage through droplet inhalation, and then the bacteria can be disseminated to other organs in the body including the central

nervous system (CNS). At the meninges, they form subpial tubercles called Rich foci which can rupture and release bacteria into the subarachnoid space, heralding the onset of TBM.

1.1.3 Clinical presentation and complication

In the early stage of TBM, common clinical signs at presentation are headache, neck stiffness, fever, anorexia, lethargy, weight loss and sometimes altered consciousness. These non-specific symptoms can only be distinguished from other illnesses by their persistence and can be easily neglected if patients are not monitored regularly (Thwaites et al., 2013). In more advanced clinical stages, patients often present with meningeal irritation, seizures, cranial nerve palsies, hemiparesis, and abnormal movements (Garg, 1999).

The major complications of the disease are hydrocephalus, infarcts and tuberculomas which cause seizures and coma. Hydrocephalus is defined as the abnormal accumulation of cerebrospinal fluid (CSF) within the brain that leads to dilation of the ventricular system and increased intracranial pressure (Garfunkel et al., 2007) . Brain infarcts are dead tissue in the brain resulting from shortage in blood and oxygen supplies. Tuberculomas are focal TB lesions which can be single or multiple and can cause mass effect when becoming large (Choudhury, 1989). These complications usually occur within the first 3 months of treatment and can result in severe neurological sequelae in survivors. They can be fatal if not detected and treated quickly. Early detection and prompt treatment of TBM are therefore crucial in determining a patients outcome (Wilkinson et al., 2017).

The severity of the disease is classified by the British Medical Research Council grade (MRC grade) which was first constructed in 1948 from the streptomycin trial (Streptomycin in Tuberculosis Trials Commitee, 1948). Patients were subdivided into early (no clinical signs of meningitis or focal neurology and fully conscious), medium (patients condition falling between early and advanced) and advanced (extremely ill, in deep coma). This classification was based solely on clinical experience rather than statistical derivation. In 1974, the Glasgow coma score (GCS) was first described as an objective measurement of a patients conscious state following traumatic head injury (Teasdale and Jennett, 1974). The score has three components: eye opening, motor, and verbal responses (Table 1.1) which are added up to an overall score. GCS ranges from 3 to 15 with a higher score indicating better neurological condition.

Table 1.1: *Glasgow coma score*

Glasgow coma score	**Score**
Eye opening	
- Spontaneous	4
- To command	3
- To pain	2
- None	1
Motor	
- Obeys command	6
- Localizes pain	5
- Withdraws to pain	4
- Flexes to pain	3
- Extends to pain	2
- No response	1
Verbal	
- Orientated	5
- Confused	4
- Inappropriate words	3
- Incomprehensible	2
- None	1

With the introduction of the GCS, the MRC grade was modified to the following grading system: grade I (GCS 15; no focal neurological signs), grade II (GCS 11-14 or 15 with focal neurological signs), or grade III (GCS $\leq$ 10) (Thwaites and Tran, 2005). This grading system has become the most widely used classification for TBM severity up to now (Table 1.2).

Table 1.2: *Modified British Medical Research Council (MRC) TBM severity grades*

MRC grade	**Diagnostic criteria**
Grade I	Glasgow coma score 15, no focal neurological signs
Grade II	Glasgow coma score 11-14 OR Glasgow coma score 15 with focal neurological signs
Grade III	Glasgow coma score 10 with or without focal neurological signs

1.1.4 Diagnostic

TBM is diagnosed based on a combination of clinical features, laboratory and radiographic investigations.

Clinical features:
The clinical features of TBM are well described, however, they are non-specific and often delayed (Thwaites and Tran, 2005; Thwaites et al., 2007; Wilkinson et al., 2017). Therefore, early diagnosis is notoriously challenging. The diagnosis becomes apparent when patients present with advanced disease symptoms. However, at this state, patients often suffer irreversible neurological sequelae.
The duration of the symptoms is the most predictive feature of TBM (Thwaites and Tran, 2005). Patients are clinically diagnosed with TBM if they have more than 5 days of meningitis symptoms, neck stiffness and cerebrospinal fluid abnormalities (Thwaites and Tran, 2005).

Laboratory diagnosis
Clinically suspected TBM can be confirmed either by CSF microscopy with Ziehl-Neelsen (ZN) staining, mycobacterial culture or polymerase chain reaction (PCR). Despite recent advances, these laboratory diagnostics still lack sensitivity and an accurate early diagnostic test of TBM remains elusive. Thus, the decision to start anti-tuberculosis treatment is often based on a clinical features only (Nhu et al., 2014; Patel et al., 2014; Thwaites et al., 2004b; Thwaites and Tran, 2005).

Radiographic investigations
Neuroimaging using computed tomography (CT) and/or magnetic resonance imaging (MRI) has been used to assist the diagnosis of TBM. Images demonstrate basal meningeal enhancement, infarcts, hydrocephalus, and tuberculomas which normally present at the advanced stage of the disease (Wilkinson et al., 2017). These features are more commonly seen in children than in adults and also in people with HIV co-infection. However, about 30% of patients at an early stage of TBM will show a normal brain CT, and about 15% will show a normal MRI scan (Wilkinson et al., 2017).
A uniform case definition based on clinical expert judgement has been introduced in 2010 to help standardize clinical research in TBM (Table 1.3) (Marais et al., 2010). The criteria sub-divide patients into diagnostic classification of definite, probable, and possible TBM, and those with a confirmed alternative diagnosis. The classification is

retrospectively performed when all the laboratory and mycobacterial culture results are available.

Table 1.3: *TBM diagnostic criteria*

		Diagnostic score
Clinical criteria	(Maximum category score = 6)	
	Symptom duration of more than 5 days	4
	Systemic symptoms suggestive of tuberculosis (one or more of the following): weight loss (or poor weight gain in children), night sweats, or persistent cough for more than 2 weeks	2
	History of recent (within past year) close contact with an individual with pulmonary tuberculosis or a positive TST or IGRA (only in children < 10 years of age)	2
	Focal neurological deficit (excluding cranial nerve palsies)	1
	Cranial nerve palsy	1
	Altered consciousness	1
CSF criteria	(Maximum category score=4)	
	Clear appearance	1
	Cells: 10-500 per μl	1
	Lymphocytic predominance (>50%)	1
	Protein concentration > 1 g/L	1
	CSF to plasma glucose ratio of less than 50% or an absolute CSF glucose concentration less than 2.2mmol/L	1
Cerebral imaging criteria	(Maximum category score=6)	
	Hydrocephalus	1
	Basal meningeal enhancement	2
	Tuberculoma	2
	Infarct	1
	Pre-contrast basal hyperdensity	2
Evidence of tuberculosis elsewhere	(Maximum category score=4)	

Continued on next page

Table 1.3 – *Continued from previous page*

		Diagnostic score
Evidence of tuberculosis elsewhere	(Maximum category score=4)	
	Chest radiograph suggestive of active tuberculosis: signs of tuberculosis=2; miliary tuberculosis=4	2 or 4
	CT/ MRI/ ultrasound evidence for tuberculosis outside the CNS	2
	AFB identified or *Mycobacterium tuberculosis* cultured from another source-i.e., sputum, lymph node, gastric washing, urine, blood culture	4
	Positive commercial *Mycobacterium tuberculosis* NAAT from extra-neural specimen	4

Abbreviation: TST: Mantoux tuberculin skin; IGRA : Interferon Gamma Release Assay; NAAT: Nucleic acid amplification tests; AFB: Acid-Fast Bacillus
Diagnostic criteria based on total score:
Possible TBM: score 6-9 (if no brain imaging) or 6-11 (if brain imaging)
Probable TBM: score >9 (if no brain imaging) or >11 (if brain imaging)
Definite TBM: acid-fast bacilli seen in CSF or *M. tuberculosis* cultured or detected by commercial NAAT in CSF

1.1.5 Management of TBM

The management of TBM involves a combination of anti-tuberculosis chemotherapy, adjunctive anti-inflammatory treatment, supportive care, and management of complications that arise after the start of treatment.

Anti-tuberculosis chemotherapy:
Treatment guidelines for TBM recommend the same standard anti-tuberculosis regimen as for pulmonary TB but for a longer duration, usually between 9-12 months. Treatment starts with a combination of four first-line drugs: isoniazid (5mg/kg/d; maximum, 300 mg/d), rifampin (10 mg/kg/d; maximum, 600 mg/d), pyrazinamide (25 mg/kg/d; maximum, 2 g/d), and ethambutol (20 mg/kg/d; maximum, 1.2 g/d) for the first 3 months. After this intensive phase, patients continue to receive rifampin

and isoniazid at the same doses for another 6 to 9 months. The above regimen is used when drug-resistant TB is not suspected.

To achieve the best therapeutic effect, treatment should start early, preferably before the onset of coma to increase the likelihood of disability-free survival. It is also recommended to avoid interruptions of any first-line anti-tuberculosis chemotherapy. Isoniazid and rifampin are the two crucial components of the chemotherapy and using them whenever possible is one of the key principles of a successful treatment (Thwaites, 2017).

Unfortunately, isoniazid and rifampin resistance has become increasingly common (Heemskerk et al., 2017; Wilkinson et al., 2017). In Vietnam, 20% of TBM cases are caused by bacteria resistant to isoniazid, and around 5% are resistant to at least rifampin and isoniazid (multi-drug resistant TB (MDR)) (Heemskerk et al., 2016). Isoniazid resistance has been shown to be associated with a significant increase in the risk of death from TBM (Heemskerk et al., 2017), especially in HIV-infected individuals. Meanwhile, outcomes for patients with MDR TBM are dismal, especially in those with HIV co-infection where mortality can reach nearly 100% (Marais et al., 2011; Seddon et al., 2012; Thwaites et al., 2005). Currently, the optimal treatment for TBM patient with drug resistance is still unknown. Two possible treatment options for isoniazid resistant-TB include:

- Option 1: Replace isoniazid with levofloxacin (20mg/kg/24 hrs; maximum 1000 mg/day) in the standard treatment. Patients with more severe disease and slow response can use pyrazinamide during the whole treatment period.

- Option 2: Stop isoniazid and treat with rifampin, ethambutol, and pyrazinamide for the entire 9-12 months of treatment. Patients with confirmed ethambutol resistance should follow option 1.

For patients with MDR, second-line anti-tuberculosis treatment should be given as soon as possible (Davis et al., 2018; Thwaites et al., 2013). According to the national guideline, the second-line agents consist of injectable drugs (kanamycin, amikacin, capreomycin), fluoroquinolones (levofloxacin, moxifloxacin, gatifloxacin) and oral drugs (ethionamide, prothionamide and cycloserine).

Because in vitro drug susceptibility testing takes weeks, it is very challenging for clinicians to promptly adjust treatment to overcome resistance. Rapid detection of rifampin resistance by GeneXpert has substantially facilitated the regimen tailoring. GeneXpert is a commercial real time PCR-based assay that can rapidly detect the bacteria and rifampin susceptibility in tandem. However, outcomes of patients with

MDR remain poor due to the low sensitivity of the method, and the low effectiveness of the second line antibiotics.

Adjunctive treatment:
One of the most important determinants of death from TBM is an excessive intracerebral inflammatory response. Adjunctive corticosteroids help reduce inflammation; therefore they have long been hypothesized to improve outcome. This hypothesis has been tested in a number of trials in the past few decades. In 2016, a systematic review and meta-analysis by the Cochrane collaboration involving 1140 patients from seven randomized clinical trials (RCT) (Prasad and Singh, 2008) concluded that corticosteroids increase short-term survival in HIV-uninfected adult and children. However, the benefit in HIV-infected patients still remains uncertain. Nonetheless, the current recommendation is to administer adjunctive corticosteroids for the first 6-8 weeks of treatment in all patients with TBM (Table 1.4).

Table 1.4: *Standard regimen for adjunctive treatment with dexamethasone*

	MRC grade I Daily dexamethasone dose/route	**MRC Grade II and III** Daily dexamethasone dose/route
Week 1	0.3 mg/kg/24 hrs IV	0.4 mg/kg/24 hrs IV
Week 2	0.2 mg/kg/24 hrs IV	0.3 mg/kg/24 hrs IV
Week 3	0.1 mg/kg/24 hrs IV	0.2 mg/kg/24 hrs IV
Week 4	3 mg/24 hrs Oral	0.1 mg/kg/24 hrs IV
Week 5	2 mg/24 hrs Oral	4 mg/24 hrs Oral
Week 6	1 mg/24 hrs Oral	3 mg/24 hrs Oral
Week 7	Stop	2 mg/24 hrs Oral
Week 8		1 mg/24 hrs Oral

Abbreviations: Hrs: hour; IV: intravenous

Supportive care and management of complications:
Supportive care plays an important role in TBM management, especially in the early stages of the disease. It includes fluid and electrolyte management, temperature regulation and mechanical ventilation. Special attention is paid to hyponatraemia (plasma sodium $<$ 135 mmol/l) management. Hyponatraemia is prevalent in patients with TBM (40-50%), and has been shown to be a risk factor for poor outcome (Anderson et al., 2010; Dass et al., 2003). Neurological complications occurring during the disease course such as hydrocephalus, stroke, raised intracranial pressure, tuberculoma also require active management. Unfortunately, little is known about how to best manage these complications (Wilkinson et al., 2017).

1.2 Introduction to prognostic models

This thesis focuses on the development of prognostic models for patients with TBM and this section provides a short overview of the topic. In a medical context, prognosis simply refers to the forecast of future clinical outcomes in individuals with a certain health condition (Hemingway et al., 2013). Medical doctors routinely make prognostic assessments as part of their daily clinical practice. They are interested in knowing the likely future course of the present illness, or whether the patient will recover from the current disease. This information can be used to suggest a suitable intervention and management plan for patients. Prognostic model research investigates the relationship between outcome and patients characteristics systematically in order to improve healthcare (Hemingway et al., 2013). It includes three main steps:

- Model development: building a statistical model that provides prediction for a future outcome based on a (ideally) large high-quality dataset and assessing its predictive performance. Internal validation techniques (e.g. cross- or bootstrap-validation) are necessary to provide a fair assessment of the predictive performance.
- Model validation: assessing the predictive performance of a prediction model on new subjects data that were not used in the model development step;
- Impact studies: quantifying the impact of using a prediction model on clinical decision making in practice.

1.2.1 What is a prognostic model and why is it important

Using one single predictor in isolation for prediction rarely gives an adequate prognostic estimate given the variability in disease aetiology, and patients characteristics. Hence, prognostic studies generally require the inclusion of multiple predictors, i.e. a multivariable approach (Moons et al., 2009). A multivariable statistical model can then be used to translate a patient's individual risk factors into an overall risk prediction for a certain outcome. Such a tool is normally referred to as a prognostic model, prediction model, or prognostic index (Steyerberg et al., 2013).

Prognostic models can be used to identify important risk factors and provide clinicians with objective, reproducible and reliable risk estimates, provided that the analysis has been performed properly. Hence, they assist clinicians in informing patients about their future disease outcomes, and facilitate medical decision making

regarding treatment and patient management (Steyerberg et al., 2013). Prognostic models are also used to select relevant patients for a clinical trial where the intervention focuses on a targeted subpopulation. In addition, they can be used to perform stratified analysis according to predicted risk groups, or to improve the power of statistical analyses in a clinical trial (Steyerberg et al., 2013). The latter is achieved by targeting the analysis on groups of patients selected according to their predicted risk (Roozenbeek et al., 2009), or adjusting the analysis by prognostic factors (Hernández et al., 2004). Another application of prognostic models is to support decisions regarding treatment initiation (Hingorani et al., 2013). Specifically, whether a patient receives a certain intervention or not may depend on their predicted risk. However, if the prediction model was developed based on data from untreated individuals, the predicted outcome risk are not necessarily transportable to differential response when treatment is present. Hence, before implementing any treatment tailoring according to a prognostic model, the efficacy of the intervention on any risk group of patients should be validated, preferably in a clinical trial.

A prognostic model consists of three different parts: outcome, predictors and a statistical model. An outcome could be death, treatment response, disease progression, or quality of life during a well-defined period of follow-up (Moons et al., 2009). Candidate prognostic predictors can be clinical or non-clinical profiles of a patient such as patient demographics, clinical history, physical examination, disease severity, laboratory results and any previous treatment (Moons et al., 2009).

When choosing a prognostic predictor, the focus lies in the association between the predictor and the outcome. This is markedly different from aetiological research where the aim is to validate an established causal relationship. In prediction, the aim is to predict as accurately as possible. Thus, all variables with potential associations with the outcome of interest can be included as a predictor (Moons et al., 2009). Additional criteria include the availability of the data at prediction time, and data quality which refers to the completeness, validity and accuracy of the data (Shmueli, 2010).

The relationship between outcome and predictors are captured in a statistical model. Because quite often the outcome of interest is binary (event), possibly in combination with an event time, the statistical model of choice is often a logistic or Cox regression, respectively. The analysis of a prognostic model requires multivariable regression techniques, which have been extensively described and discussed in the literature (Harrell, 2015; Royston and Sauerbrei, 2008; Steyerberg, 2009).

Statistical aspects of developing a prognostic model are covered in Chapter 2. In brief, primary statistical issues include handling of missing data, performing variable selection, modeling continuous variable, and assessing performance. Other considerations may include exploring the interaction between covariates, assessing the robustness of the model to outliers and influential measurements, and checking model assumptions (Royston et al., 2009; Steyerberg et al., 2013). These issues are, preferably, clearly addressed via pre-definition of the chosen analytical approach before the actual development of the prognostic model.

One of the most important problems of prognostic models is over-fitting. The term describes the situation that the prediction model fits the data under study very well but fails to predict well for future subjects (Steyerberg, 2009). This phenomenon normally happens when the prediction model is constructed based on a small sample size with many included covariates. Special care should be given when building a prediction model on limited data, especially when stepwise variable selection is involved (Steyerberg et al., 2000). The major concern is the biased estimation of regression coefficients (Steyerberg et al., 1999). To avoid over-fitting, a common recommendation is that at least 10 events are needed for each parameter included in a model to obtain accurate regression coefficients (Peduzzi et al., 1996). Of note, for logistic regression, the number of events is the smaller value when comparing the number of subjects who have the outcome versus the number of subjects who do not (Austin and Steyerberg, 2017). The one in ten rule is however controversial and has received a lot of criticism (Courvoisier et al., 2011; Ogundimu et al., 2016; van Smeden et al., 2016). Several authors suggested that it could be relaxed, even up to EPV = 2 for linear regression (Vittinghoff and McCulloch, 2007; Austin and Steyerberg, 2015). Others suggested that a higher minimum number of EPV is required. For example, (Harrell, 2015) recommended at least 15 EPV, and Steyerberg et al. (2000) recommended at least 50 EPV is needed to perform stepwise selection with $\alpha = 0.05$. Also, a number of studies critiqued the lack of theoretical and empirical foundations of the rule, and its weak relation to the model predictive performance(Ogundimu et al., 2016; van Smeden et al., 2016, 2018). They suggested that considering the rule alone is insufficient, and other factors should be taken into account when determining an adequate sample size for developing a prediction model. For example, Courvoisier et al. (2011) proposed two quantities: the correlation between covariates, and the anticipated magnitude of regression coefficients based on previous publication.

Over-fitting causes over-optimism in the predictive performance (Royston et al., 2009). In other words, we are likely to overestimate the predictive performance of a

model if we build and assess it on the same data set. The amount of over-optimism can be quantified and corrected by validating the model internally using resampling techniques (Harrell, 2015; Hastie et al., 2009) or, ideally, by validating it on an external independent dataset.

In summary, a prognostic model is reliable if it is developed on a large and high-quality dataset, analyzed using sound and robust statistical modeling method, and validated in an independent sample (Steyerberg et al., 2013).

1.2.2 Source of data and treatment use

Ideally, the most suitable design for building a prediction model is a prospective cohort study (Moons et al., 2009). Patients enter a cohort study if they fulfill certain criteria, for example, being diagnosed with TBM. After that, they are followed until they develop the outcomes of interest or, if they do not develop the outcome, until a well-defined target follow-up duration. In a prospective cohort study, usually all subjects receive the same treatment, i.e. standard of care, hence it enables us to measure the predictors and outcome in the absence of prognosis modifying novel treatments (Moons et al., 2009). In addition, prospective cohort studies have a pre-defined schedule of predictor and outcome assessments, which minimises the number of missing values and loss to follow-up when collecting data. Alternatively, data from randomized clinical trial can also be used to develop prediction models. It can be seen as a special case of a prospective cohort study with the presence of one or multiple interventions. If the trial is effective, one could restrict the analysis to the subset of untreated or treated patients. However, this approach could result in a model with larger prediction uncertainty and higher risk to be over-fitted because of the reduction in sample size and the number of events (Groenwold et al., 2016). Alternatively, treatment could be included as a predictor in the prediction model (Collins et al., 2015; Groenwold et al., 2016; Moons et al., 2009). If the trial is ineffective, one can simply combine data from treatment and control arms together.

The major concern of a prediction model that is based on data from one or multiple RCTs is its lack of generalisability. The reasons include the strict recruitment criteria for the trial, and the difference in the quality of clinical care for patients in the trial and outside the trial (Collins et al., 2015; Moons et al., 2009).

It is quite common for patients in prognostic studies to receive risk-lowering treatment during follow-up. For example, the standard anti-tuberculosis chemotherapy is given to all patients with clinically suspected TBM. The intervention, if effective,

will reduce the predicted risk. This, in turn, will hamper the generalisability of the model in populations where the treatment is not available. However, if the effect of treatment use is consistent across populations, the prognostic model still provides valid prediction. Such treatment or intervention is called background treatment, and it does not affect the validity of a prediction model (Pajouheshnia et al., 2017).

1.2.3 Baseline and dynamic prediction

The majority of prognostic models in the medical literature use patient information collected at a single baseline time point to predict the future occurrence of an outcome. The single time point is often chosen as a clinically relevant baseline, which could be time of diagnosis, or time of treatment initiation. Baseline models to predict death for patients with TBM are presented in Chapter 3. In general, baseline prognostic models are easy to implement in clinical practice and are useful. However, they do not capture the change in patients clinical profiles due to treatment response and disease progression. They also miss complications arising both from drug uses or the disease, which could change the likelihood of outcome.
Nowadays, it is common in much medical research with long-term follow-up that several biomarkers are repeatedly measured during the disease course. The changes in biomarkers trajectory over time could potentially provide a better understanding of the disease mechanism and progression (Rizopoulos and Ghosh, 2011). It is clinically relevant to investigate whether such information could allow initial prognosis to be updated and improved. This question is studied in Chapter 4.

Building dynamic prediction models that provide updated predicted risk as new information is collected is an active research area in statistics. Two of the main statistical approaches are landmarking and the joint modeling of longitudinal and time to event data. In the landmarking approach, prediction is based on a time-to-event model with the value of the longitudinal covariates determined at the landmark time t_{LM}. The data consist of the subjects who are still alive and under follow up at t_{LM}, and their longitudinal information recorded up to t_{LM}. Typically, the survival model is a Cox proportional hazard model with an administrative censoring at a fixed prediction time window. The censoring is imposed to reduce possible bias due to the violation of the proportional hazard assumption. Of note, the prediction is therefore only valid for chosen prediction window.
The longitudinal markers are normally measured with error, and their complete path is unknown because they are not measured at all time points (Papageorgiou et al.,

2018). The classical landmarking approach however assumes that the longitudinal marker is measured without error. In addition, by only looking at the last observed longitudinal value, it ignores the previous trajectory of the marker before the landmark time, and neglects the time elapsed between the measurement time and the landmark time (Ferrer et al., 2017; Maziarz et al., 2017; Rizopoulos et al., 2017). A better alternative is the two-stage partly conditional method (Maziarz et al., 2017). It requires fitting a mixed-effects model for the longitudinal marker on the patients who are at risk at t_{LM}. The fitted marker value at t_{LM} as obtained from the mixed-effects model replaces the last observed value of the longitudinal marker in the landmark model.

Prediction for a wide range of landmark times can be obtained by constructing the so-called landmarking super model. It involves three main steps: i) creating landmarking data sets for a series of time points, ii) stacking them together into one prediction data set, with a variable that quantifies the landmark time, and iii) fitting a Cox model on the stacked data in which the regression coefficients are allowed to depend on landmark time in a smooth way (Van Houwelingen, 2007; van Houwelingen and Putter, 2011). The Cox model is a clock-forward model, i.e. the hazard depends on time since the baseline, to simplify the derivation of the statistical properties (van Houwelingen and Putter, 2011).

The landmarking approach is straightforward, easy to explain and computationally feasible as it only requires fitting a simple model with baseline covariates. However, it does not satisfy the consistency condition for a valid prediction function as formulated by Jewell and Nielsen (1993). Loosely speaking, the condition implies that the prediction obtained from the model at each landmark time should be consistent with those from earlier time points (van Houwelingen and Putter, 2011). To satisfy the condition, the longitudinal process and the hazard function must be linked at all time points, and a specification for the distribution of the former one is necessary (Ferrer et al., 2017; Suresh et al., 2017).

In the joint modelling approach, the prediction is derived from the joint distribution of the longitudinal response and the time-to-event outcome. Joint models may fall under the umbrella of joint multivariate random effects models, or depending on the factorization of the joint distribution, they can be shared random effect models, or pattern-mixture models. Of those, the shared random effect models have received the most attention in the literature. I only consider this type of joint model in my thesis. In this approach, the longitudinal marker is treated as a time-varying covariate instead of a time-fixed variable as in the landmarking approach.

The shared random effects model consists of two linked submodels. Typically, these are a mixed-effects model that describes the true longitudinal process and a Cox model for the survival process. The two submodels are linked by using a function of shared random effects, which can be specified in various ways (Henderson et al., 2000; Rizopoulos, 2011; Tsiatis and Davidian, 2004). Under this joint model, the random effects account for two sources of associations: between repeated measurement of biomarkers in the longitudinal process, and between the longitudinal and the time to event processes. The estimation of the parameters can be based on maximum likelihood (Guedj et al., 2011; Rizopoulos, 2012; Thiébaut et al., 2005; Wulfsohn and Tsiatis, 1997) or a Bayesian approach, typically using the Markov chain Monte Carlo algorithm (R. Brown and G. Ibrahim, 2003; Rizopoulos, 2016; Wang and Taylor, 2001; Xu and Zeger, 2001). Computing the conditional survival probability from a shared random effects model involves a fairly complicated integral over the distribution of the longitudinal process. Thus it requires special software. Several extensions for the classical shared random effects model have been proposed, including modelling multiple longitudinal markers (Lin et al., 2002; Rizopoulos and Ghosh, 2011), multiple recurrent outcomes(Elashoff et al., 2008), and complex association structures (Mauff et al., 2017; Rizopoulos, 2012).

The motivations behind choosing a joint modeling over a landmarking approach for dynamic prediction for time-to-event outcomes include:

- Its ability to handle the endogeneity of a time-varying covariate in relation to the survival outcome. A variable is endogenous if its value can only be determined as long as the outcome of interest has not occurred (Kalbfleisch and Prentice, 2002).

- Its ability to generate a consistent estimate of the risk provided that the joint distribution of the longitudinal and time to event outcomes is correctly specified (Rizopoulos et al., 2017).

However, a joint model is more difficult to explain to clinicians, and often requires intensive computation. Hence, their application in practice is still limited despite of the aforementioned advantages.
In terms of prediction performance, the two approaches have been compared under several settings (Ferrer et al., 2017; Huang et al., 2016; Maziarz et al., 2017; Rizopoulos et al., 2017; Suresh et al., 2017). In general, the superiority of joint models over landmarking is only guaranteed if the model for the longitudinal process is correctly

specified (Ferrer et al., 2017; Rizopoulos et al., 2017). The two methods are contrasted in our setting in Chapter 4.

1.3 Variable selection in the presence of multiply imputed data

1.3.1 Missing data

Missing data is common in clinical research, yet poorly discussed in most prediction studies (Bouwmeester et al., 2012). Choosing a statistical approach to handle missing data depends on the missing data mechanism and the proportion of missingness in the data.

Data is classified as missing completely at random (MCAR) if the probability of an observation being missing does not depend on the observed variables or outcome. Data is missing at random (MAR) if the probability of being missing depends on observed values. Finally, data is missing not at random (MNAR) if the missiningness is classified as neither MCAR or MAR (Little and Rubin, 2002; White et al., 2011).

The most common approach to deal with missing data is to restrict the analysis to individuals with complete information, or the so-called complete case analysis. This approach is widely used because of its simplicity. Complete case analysis can be used when the percentage of cases with at least one value missing in the data is less than 5%, or when the MCAR assumption is reasonable (Little and Rubin, 2002). On the other hand, as a rule of thumb, if the proportion of missing values is larger than 5%, implementing multiple imputation is recommended (Collins et al., 2015; Little and Rubin, 2002; White et al., 2011). This method creates multiple copies of data by replacing missing values with plausible ones. When correctly implemented, multiple imputation provides unbiased estimates under MAR mechanism. Regarding the required number of imputed datasets, classic recommendations state that 3-5 imputations are adequate to yield reliable estimates (Little and Rubin, 2002). More recently, White et al. (2011) showed that a larger value is often needed and suggested a rule of thumb of one imputation per percentage of incomplete cases. Estimates from multiply imputed (MI) data sets are then combined using Rubins rule (Little and Rubin, 2002).

1.3.2 Variable selection for fully observed data

In a prediction model, subject matter knowledge and expert opinion should be the most important rationale for selecting a variable (Harrell, 2015). However, this information is not always available, and expert opinion might also introduce bias (Steyerberg, 2009). An alternative option is to start with a broad set of variables based on subject matter knowledge, and then perform some form of automated variable selection to remove variables that do not improve the prediction. For fully observed data, such methods have been extensively studied. The most straightforward, and therefore the most popular variable selection method is backward stepwise selection based on p-value or AIC. The critical shortcomings of this method include instability in selected variables and inflated parameter estimates (Harrell, 2015). Of note, p-values are used in hypothesis testing for the population correlation between variables and outcomes. In prediction, the relationship on the others hand is studied with the aim of providing an accurate prediction at the individual level. Hence, although variables with small p-values are likely to be good predictors, the decision of whether to include a variable in the model should be made based on predictive value rather than on p-value. Several alternative approaches have been proposed that show more robust performance. Austin and Tu (2004) suggested performing backward elimination on bootstrap samples from the original data. A final model is fitted based on variables with an inclusion frequency over all bootstrap samples above some threshold. Using bootstrap resampling allows us to better separate true predictors from the noise ones. Therefore, compared to using stepwise selection on a single data set, it addresses the instability in variable selection and enhances model performance. In their case study, a threshold of 60% showed the best performance. However, the author only considered a limited number of inclusion frequencies. Another attractive method is the least absolute shrinkage and selection operator (lasso) selection technique. It combines variable selection and shrinkage via a penalized likelihood technique (Tibshirani, 1996). Shrinkage can improve predictive performance, with some bias in the parameter estimates. Lasso and its related methods such as ridge regression and elastic net are termed regularization methods (Hastie et al., 2009). Details of the lasso method are described in Chapter 5. Using a simulation study, van Houwelingen and Sauerbrei (2013) compared the performance of lasso with standard backward elimination followed by some form of shrinkage. The two methods yielded similar performance, with a more parsimonious model obtained by the backward elimination method. The authors considered three fairly strict significant levels of $\alpha = 0.157, 0.05$ and 0.01 for backward elimination models, while they only used the "optimal" penalty in the lasso

model. A more parsimonious model is obtained by choosing a stronger penalty, for instance, based on the 1-se rule (Hastie et al., 2009) or the "tolerance" rule (Kuhn and Johnson, 2013). However, a stronger penalty may over-shrink parameter estimates. The impact of a stronger penalty on model performance is investigated in Chapter 5 of the thesis. In a recent simulation study, Hastie et al. (2017) compared best subset selection, forward selection and the lasso for linear regression. The authors observed that there was no uniformly best method in all considered scenarios. Best subset selection and forward stepwise selection performed similarly throughout, and better than lasso in high signal-to-noise cases. On the other hand, lasso outperformed the two methods in low signal-to-noise scenarios.

1.3.3 Variable selection with multiply imputed data

Variable selection in the presence of multiply imputed data needs special attention. With stepwise backward elimination, the gold standard approach entails fitting the model under consideration to each MI data set, and combine results via Rubins rule (Little and Rubin, 2002) to obtain Wald tests for all variables. The variable with the largest associated p-value is removed from the model before moving on to the next step. However, it may require intensive computation (Wood et al., 2008). Numerous approaches have been proposed that are less computationally demanding and can be applied with both stepwise backward elimination and shrinkage selection methods. The most straightforward approach is to perform variable selection on each imputed data set separately and combine results afterwards. The set of selected variables is likely to vary by MI data set. One option is to refit a final model based on the most frequently ($\geq$50 %) selected variables (Wood et al., 2008). The second approach is an extension of the first one where variable selection is used in combination with bootstrapping to stabilize the model (Heymans et al., 2007; Long and Johnson, 2014). The variable inclusion frequency is calculated over all bootstrap and MI data sets. This method takes into account the uncertainty caused both by missing data and by sampling variability. The third approach is to stack all MI data sets and perform variable selection on that single data set (Wood et al., 2008). Unlike the previous two approaches, the stacked approach does not lead to a different set of selected variables. When computing p-values in the stacked backward elimination procedure, Wood et al. (2008) additionally proposed a weighting scheme to account for the fraction of missingness per variable, and the repeated occurrence of individuals. Using elastic net as a penalized likelihood variable selection method, Wan et al. (2015)

adopted the same idea but their proposed weighting scheme is per observation rather than per variable. The weighting scheme is explained in Chapter 5. Chen and Wang (2013) proposed an alternative that applies to lasso selection called MI-lasso. It is a type of grouped lasso, in which a variable is either not selected at all, or selected in each MI data set, just like in the stacked approach. Contrary to the stacked approach, it may give different coefficient estimates per MI data set.

A review of current strategies for variable selection in the presence of missing data was given in Zhao and Long (2017). Despite significant contributions, it is unclear which approach performs best and in which circumstances one approach is preferable to the others. These questions are addressed in Chapter 5 of my thesis.

1.3.4 Prognostic models in tuberculous meningitis: Literature review

As mentioned above, treatment for TBM is the same with standard anti-tuberculosis regimes for pulmonary TB but it is given for a longer duration. Apart from the treatment, supportive care and management of complications are vital in improving patients condition and reducing neurological sequelae. Hence, it is important to identify patients with poor prognosis at an early stage of the disease so that prompt action could be implemented. This process can be facilitated by a reliable prediction model.

I reviewed all articles published from September, 1996 - September, 2016 indexed on PubMed satisfying the following search terms: "tuberculous meningitis"[Title] AND (score[All Fields] OR "prediction"[All Fields] OR "prognostic"[All Fields] OR "risk factors"[All Fields] OR "clinical trial"[All Fields]) AND ("1996/09/01"[PDAT] : "2016/09/01"[PDAT]) NOT ("children"[All Fields]) . Amongst the results, there are 19 articles studying risk factors for mortality or a composite of death and neurological outcome in TBM patients (Table 1.5). However, studies in which prognostic models were developed and validated are scarce. There is only one article that presents a prediction model for TBM using information that was available at baseline (Erdem et al., 2015), and no articles have been published that take into account changing clinical variables after the start of TBM treatment.

1.3.4.1 Prognostic model for TBM

When initiating this thesis, the only publication on a prediction model for unfavorable outcome in TBM is a multi-center study in Europe conducted on 507 TBM confirmed adult patients (Erdem et al., 2015). A single model was presented for both patients with and without HIV-coinfection. HIV-infection regardless of CD4 cell count level, together with alcoholism and the use of immunosuppressive drugs such as dexamethasone was aggregated into one predictor called immunosuppression. Unfavorable outcome (165/507) was defined as a composite of either mortality (86/507) or sequela at the end of antibiotic treatment (79/507). Sequela was defined as any central nervous system (CNS) dysfunction that persisted during the follow-up period. The author came up with a full model with 13 variables selected based on their significant association with the outcome in univariable analyses. The authors then employed binary logistic regression with stepwise backward selection in model selection, the bootstrap resampling method in model validation, and handled missing data by multiple imputation. In the presence of MI data, the variable selection was performed in combination with bootstrap resampling. The chosen variable inclusion threshold was 15%. Eight predictors were selected based on their inclusion frequency. The author refitted the model based on these eight predictors. Amongst them, age was not significant so it was removed from the final model. The final model identified seven prognostic factors. Altered consciousness, diabetes mellitus, immunosuppression, neurological deficits, hydrocephalus, and cerebral vasculitis were associated with an unfavorable outcome, whereas, nausea on presentation linked with a better outcome. Finally, they constructed a severity index, called Haydarpasa Meningitis Severity Index (HAMSI), for use in practice and compared its predictive performance with the classical MRC staging system via a receiver operating characteristic (ROC) curve. The HAMSI score demonstrated a better performance compared to the classical MRC grade.

Despite the great effort in building the first prognostic model in TBM, the paper suffered several limitations. Regarding the selection of predictors, the authors did not consider GCS as a prognostic factor due to missingness in their database. They also did not consider drug resistance, or any laboratory parameters such as CSF lymphocyte count, or sodium level despite their established prognostic characteristics (Rock et al., 2008). Regarding the statistical analysis, the author selected predictors based on their statistical significance in univariable analysis. This type of model selection is not recommended by many authors (Harrell, 2015; Steyerberg, 2009). In addition, they removed age due to its statistical insignificant effect in the final

model. It is unclear whether the author repeated this extra variable selection step in the validation process. Furthermore, they claimed that they failed to establish the association between the classical MRC staging system (Streptomycin in Tuberculosis Trials Commitee, 1948) and the disease outcome. However, univariable analysis for this variable gave a p-value < 0.001. The authors did not report the value of the area under the receiver operating characteristic (ROC) curve (AUC) and c-index even though they were mentioned in the statistical section. Finally, it is unclear which data set (i.e. the complete case data set, or multiply imputed data sets) was used to calculate the reported ROC curve and whether it was validated or not.

1.3.4.2 Other articles describing risk factors for poor outcome

I summarized the other 19 studies describing risk factors for poor outcome in TBM in Table 1.5. Poor outcome was defined as mortality or a composite of mortality and neurological sequelae. Amongst 19 studies, five explored risk factors for TBM patients with HIV infection, 11 studies are for HIV-uninfected patients, and three studies are for all patients regardless of HIV status. The majority of the studies were conducted in Asia (15/19); the others were from USA, Brazil, Europe and South Africa. Data was mainly from retrospective cohort studies. There are three studies utilizing data from RCTs. Most studies considered a wide range of candidate risk factors, including demographic information, clinical features, complications at presentation, and laboratory tests. The most commonly used statistical modeling method was multivariable logistic regression. It was used in 11/19 studies, while Cox regression was used in 6/19 studies. Stepwise selection was the only considered variable selection method, and was performed in 2/19 studies. Missing data were handled by multiple imputation in only one study (Vinnard et al., 2010).

In summary, the above review shows the paucity of prediction models for mortality in TBM, and the lack of a prognostic tool that could be implemented in clinical practice. Indeed, there is only one study that developed and validated a prognostic model for mortality using patients characteristics information at diagnosis. Moreover, the utilization of repeated measurements to update the risk prediction is completely absent in the TBM literature. Most studies on risk factors were based on a small sample size and failed to take into account the importance of handling missing values, and variable selection. Despite its age and the lack of statistical derivation, the MRC grade is still the most widely used classification for TBM severity.

Table 1.5: *Literature review: risk factors for poor outcome in TBM*

Paper	Sample size	Place	Study design	Outcome	Risk factor	Model
Salekeen et al. (2013)	52	Pakistan	Retrospective cohort study	Mortality within one year	Older age, higher MRC grade, hyponatraemia, total leucocyte count ($> 9000/\mu$L), hydrocephalus, and use of mechanical ventilation	Univariate logistic regression
George et al. (2012)**	98	India	Retrospective cohort study	In-hospital mortality	Age (> 40 years), CSF protein (≤ 60 mg%)	Multinomial logistic regression
Yasar et al. (2010)	160	Turkey	Retrospective cohort study	Mortality within 6 months	older age, altered sensorium , underlying comorbidities , and higher leukocytosis	Multivariable logistic regression
Shaw et al. (2010)	135	USA	Retrospective cohort study	Mortality	older age	Multivariable Cox regression
Croda et al. (2010) *	108	Brazil	Retrospective cohort study	Mortality within 9 months	Tachycardia, prior HAART	Multivariable Cox regression
Hosoglu et al. (2002)	434	Turkey	Retrospective cohort study	In-hospital mortality	Convulsion, comatose mental status, delayed or interrupted treatment	Multivariable logistic regression
Efsen et al. (2013) *	100	EU and Argentina	Retrospective cohort study	Mortality within 12 months	Prior AIDS, being on cART	Multivariable Poisson regression
Tho et al. (2012)*	186	Vietnam	RCT	Mortality within 9 months	Drug resistance	Multivariable Cox regression
Marais et al. (2011)*	120	South Africa	Retrospective cohort study	In-hospital mortality	lower CD4, higher MRC grade	- Stepwise selection - Multivariable logistic regression

Continued on next page

Table 1.5 – *Continued from previous page*

Paper	**Sample size**	**Place**	**Study design**	**Outcome**	**Risk factor**	**Model**
Török et al. (2011a) *	253	Vietnam	RCT	Mortality within 9 months	higher MRC grade	Multivariable Cox regression
Vinnard et al. (2010)**	1614	USA	Retrospective cohort study	Mortality during anti-tuberculous treatment.	Isoniazid resistance, older age, race, HIV positive	- Multiple imputation - Multivariable logistic regression
Lau et al. (2005)	166	Hong Kong	Retrospective cohort study	Mortality within one year	older age, lower CSF lymphocyte, poorer conscious baseline level	Multivariable Cox regression
Thwaites et al. (2004a)**	545	Vietnam	RCT	Mortality within 9 months	GCS ($\leq$ 10), hemiparesis, previous TB, extraneural or pulmonary TB, HIV infection, lower hematocrit, lower CSFleukocyte count, lower ratio of CSF to plasma glucose, alteration of the antituberculosis regimen, Dexamethason	Multivariable Cox regression
Wang et al. (2002)	160	Taiwan	Retrospective cohort study	Mortality within 6 months	older age, altered sensorium, underlying co-morbidities and leukocytosis	Multivariable logistic regression
Gu et al. (2015)	156	China	Retrospective cohort study	Severe sequelae (quadriplegia, severe intellectual impairment, blindness and/or deafness) or death	older age, change in consciousness, lower GCS and concomitant hydrocephalus, use of prednisone	Multivariable logistic regression

Continued on next page

Table 1.5 – *Continued from previous page*

Paper	**Sample size**	**Place**	**Study design**	**Outcome**	**Risk factor**	**Model**
Jha et al. (2015)	118	India	Prospective cohort study	Mortality or modified BI score $\leq$12 within 6 months.	higher MRC grade, modified BI ($<$ 12), positive tuberculosis PCR	Multivariable logistic regression
Hsu et al. (2010)	108	Taiwan	Retrospective cohort study	Mortality within 9 months	older age, hydrocephalus, steroids for 2 weeks	Multivariable logistic regression
Sütla et al. (2003)	61	Turkey	Retrospective cohort study	Mortality or 12 month BI score $<$12	MRC grade	- Stepwise variable selection; - Multivariable logistic regression
Lu et al. (2001)	36	Taiwan	Retrospective cohort studys	Mortality or 12 month BI score $<$12	Hydrocephalus, higher MRC grade	

* Study includes only HIV-infected patients;
** Study includes both HIV-infected and HIV-uninfected patients
Abbreviations: BI: Barthel index; PCR: polymerase chain reaction; cARV: combination antiretroviral therapy; HAART: highly active antiretroviral therapy ; TBM: tuberculous meningitis; TB: Tuberculosis; GCS: Glasgow coma score; MRC: Medical research council; CSF: Cerebrospinal fluid; HIV: human immunodeficiency virus; AIDS: acquired immune deficiency syndrome

1.4 Objectives of the thesis

A prediction model is an important instrument to assess an individuals risk and facilitate informed medical decisions. However, the current prognosis research in TBM is still limited. Between 2001 and 2017, the Oxford University Clinical Research Unit (OUCRU) Vietnam has enrolled around 1800 patients with TBM from five RCTs (Heemskerk et al., 2016; Mai et al., 2018; Thwaites et al., 2011; Török et al., 2011a) and one observational study (Torok et al., 2008). This is one of the largest sources of prospectively collected clinical data for TBM worldwide. Given the availability of this well-characterized database, my DPhil thesis set out to develop robust prediction models for mortality, and to address statistical issues arising during the modeling process. These include the following objectives:

1. To develop and validate prognostic models for mortality in patients with TBM using baseline information
2. To develop and validate prognostic models for mortality in patients with TBM using longitudinal information and patients characteristic at baseline
3. To compare methods in performing model selection for prediction in the presence of multiply imputed data

The structure of the thesis is as follows:

- Chapter 2 describes the pooled database, and common statistical methods that are used in the subsequent chapters.
- Chapter 3 presents the development and validation of prognostic model for 9-month mortality in TBM, using information at baseline.
- Chapter 4 presents the development and validation of dynamic prediction model for TBM using time-updated Glasgow coma score and plasma sodium, together with patients' characteristics at baseline.
- Chapter 5 compares two classes of model selection in the presence of multiply imputed data.
- Chapter 6 summaries the conclusion from the previous chapters and proposes several avenues for future research in the field.

Chapter 2

Materials and common statistical methods in prediction

Summary: In this chapter, I describe the pooled TBM database that was used throughout my thesis, together with the data collection procedure and diagnostic laboratory investigations. I additionally specify common statistical methods used in the subsequent chapters including handling of missing values and the main steps in building and validating a prediction model.

2.1 Materials

2.1.1 Study populations for TBM prediction models

The study population comprises data from six TBM studies conducted by the Oxford University Clinical Research Unit (OUCRU) in Vietnam between 2001 and 2017 (Heemskerk et al., 2017; Mai et al., 2018; Thwaites et al., 2004a, 2011; Torok et al., 2008; Török et al., 2011a). These studies were performed in partnership with two recruiting hospitals, Pham Ngoc Thach hospital for Tuberculosis and Lung disease and the Hospital for Tropical Disease (HTD), both in Ho Chi Minh City, Vietnam. The two hospitals are respective tertiary referral centers for tuberculosis and infectious diseases in southern Vietnam. Details of the six studies are summarized as follows and in Table 2.1.

Dexamethasone trial

The Dexamethasone trial (Thwaites et al., 2004a) was a RCT of standard anti-tuberculosis chemotherapy combined with adjunctive dexamethasone or placebo, conducted between 2001 and 2004. The trial included 545 adults (436 HIV-uninfected, 98 HIV-infected, and 11 with unknown HIV status) with a clinical diagnosis of TBM. None of the HIV infected patients in this trial received antiretroviral therapy (ART). The trial demonstrated that adjunctive dexamethasone significantly improved 9-month survival. Following this trial, adjunctive dexamethasone became the standard of care in Vietnam and the rest of the world.

Fluoroquinolone trial

The Fluoroquinolone trial (Thwaites et al., 2011) was a RCT of standard TB treatment alone or in combination with ciprofloxacin, levofloxacin or gatifloxacin during the first 2 months of therapy. The trial recruited 61 patients (58 HIV-uninfected, 3 HIV infected) with suspected TBM from 2003-2004. The trial revealed that levofloxacin had higher cerebrospinal fluid (CSF) penetration, measured by the ratio of plasma to cerebrospinal fluid concentrations, compared to the other fluoroquinolones.

TBM HIV cohort

This study (Torok et al., 2008) was a prospective observational study conducted from 2004-2005 to investigate clinical features, microbiological characteristics and potential prognostic factors for survival in HIV-infected patients with clinically suspected TBM (N = 58). ART was not routinely available when the study was carried out. The study identified three baseline risk factors that associated with mortality namely: advanced MRC grade, low plasma sodium, and low CSF lymphocyte percentage.

ART timing trial

This study (Török et al., 2011a) was a RCT of immediate versus deferred ART conducted between 2005-2007. It randomized 253 ART-naive patients with HIV-associated TBM to immediate or delayed (by 2 months) ART initiation. The trial could not demonstrate the superiority of immediate ART in 9-month survival and patients in the immediate ART arm had more grade 4 adverse events. All patients enrolled in this study were ART naive at diagnosis.

Intensified treatment trial

This study (Heemskerk et al., 2016) was a RCT of a standard versus an intensified anti-tuberculosis regimen, which consisted of a higher dose of rifampin and additional

levofloxacin for the first 8 weeks of treatment. The trial included 817 adults (349 with HIV-infection) with a clinical diagnosis of TBM recruited from 2011-2014. The trial could not demonstrate that intensified anti-tuberculosis treatment improved 9-month survival and showed a very similar efficacy in both arms.

Aspirin trial

This study (Mai et al., 2018) was a phase 2b RCT of a standard TB treatment alone or in combination with Aspirin 81 mg/day or Aspirin1000 mg/day added to the first 60 days of the therapy. The trial enrolled 120 adult patients with TBM between 2017 and 2018. The primary safety endpoint was defined as the development of gastro-intestinal or cerebral bleeding by 60 days and both Aspirin doses appeared to be safe. The trial could not demonstrate the efficacy of aspirin in reducing the number of brain infarction or death by 60 days for all patients. However, the authors concluded that results were sufficiently encouraging to warrant investigation of adjunctive aspirin in a large phase 3 trial.

Table 2.1: *Summary of included studies*

Cohort	Dexamethasone trial	Fluoroquinolone trial	TBM HIV cohort	ART timing trial	Intensified treatment trial	Aspirin trial
Main publication	Dexamethasone for the treatment of tuberculous meningitis in adolescents and adults (Thwaites et al., 2004a)	A randomized pharmacokinetic and pharmacodynamics comparison of fluoroquinolones for tuberculous meningitis (Thwaites et al., 2011)	Clinical and microbiological features of HIV-associated tuberculous meningitis in Vietnamese adults (Torok et al., 2008)	Timing of initiation of antiretroviral therapy in human immunodeficiency virus (HIV) associated tuberculous meningitis (Török et al., 2011a)	Intensified anti-tuberculosis chemotherapy for the treatment of Vietnamese adults with tuberculous meningitis (Heemskerk et al., 2016)	A randomised double blind placebo controlled phase 2 trial of adjunctive aspirin for tuberculous meningitis in HIV-uninfected adults (Mai et al., 2018)
Recruitment period	2001-2004	2003-2004	2004-2005	2005-2007	2011-2014	2014-2016
Sample size (HIV/no HIV)	545 (98, 436)[1]	61 (3, 53) [2]	58 (58, 0)	253 (253, 0)	817 (349, 468)	120(0,120)
Study design	RCT	RCT	Observational study	RCT	RCT	RCT

Continued on next page

Table 2.1 – *Continued from previous page*

Cohort	**Dexamethasone trial**	**Fluoroquinolone trial**	**TBM HIV cohort**	**ART timing trial**	**Intensified treatment trial**	**Aspirin trial**
Randomized interventions	Two arms, double-blind - Standard antituberculosis treatment - Standard antituberculosis treatment plus adjunctive treatment with dexamethasone [3]	Four arms, open label - Standard antituberculosis treatment - Standard antituberculosis treatment with oral dose of + ciprofloxacin (750mg/12 hours), + levofloxacin (500mg/12 hours), or + gatifloxacin (400mg/24 hours) for the first 60 days of treatment		Two arms, double-blind - Immediate ART (initiated within 7 day of commencing tuberculosis treatment) - Deferred ART (initiated after 2 months of tuberculosis treatment)	Two arms, double-blind - Standard antituberculosis treatment - Standard antituberculosis treatment intensified with a higher dose of rifampin (15mg/kg) and levofloxacin (20mg/kg) for the first 8 weeks of treatment	Three arms, double-blind - Standard antituberculosis treatment - Standard antituberculosis treatment with Aspirin 81mg/day - Standard antituberculosis treatment with Aspirin 1000 mg/day
Inclusion criteria	- Age $\geq$ 14 years - Clinical diagnosis of TBM	- Age $\geq$ 14 years - Clinical diagnosis of TBM	- Age $\geq$ 15 years - Clinical diagnosis of TBM - HIV infected	- Age $\geq$ 15 years - Clinical diagnosis of TBM - HIV infected	- Age $\geq$ 18 years - Clinical diagnosis of TBM	- Age $\geq$ 18 years - Clinical diagnosis of TBM - HIV uninfected

Continued on next page

Table 2.1 – *Continued from previous page*

Cohort	**Dexamethasone trial**	**Fluoroquinolone trial**	**TBM HIV cohort**	**ART timing trial**	**Intensified treatment trial**	**Aspirin trial**
Exclusion criteria	- Contraindication to corticosteroid use - More than one dose of any corticosteroids prior to recruitment - More than 30 days of antituberculosis chemotherapy immediately prior to recruitment - Lack of informed consent	- Contraindication to fluoroquinolone use - Pregnancy - Lack of informed consent	- Proven alternative diagnosis	- Proven alternative diagnosis - Contraindication to ART or antituberculosis treatment - Previous ART - Antituberculosis treatment 8-30 days immediately prior to recruitment - Pregnancy - Lack of informed consent	- Contraindication to fluoroquinolones of rifampin use - More than 7 days of antituberculosis treatment prior to recruitment - Known or suspected MDR[4]-TBM - Creatinine >3x ULN - Bilirubin >2.5x ULN - AST/ALT >5x ULN - Pregnancy - Lack of informed consent	- Contraindication to aspirin, dexamethasone or antituberculosis treatment - Unable to have an MRI brain scan within 5 days (120 hours) of randomization - Known or suspected MDR-TBM - History of diagnosed peptic ulceration or gastro-intestinal bleeding - Active gastro-intestinal bleeding is suspected - Pregnancy - Lack of informed consent

Continued on next page

Table 2.1 – *Continued from previous page*

Cohort	**Dexamethasone trial**	**Fluoroquinolone trial**	**TBM HIV cohort**	**ART timing trial**	**Intensified treatment trial**	**Aspirin trial**
TBM diagnosis		Antituberculosis treatment for TBM initiated on decision of the treating clinician				
- Definite TBM	AFB seen or *M.tb* cultured from CSF		AFB seen or *M.tb* cultured from CSF	AFB seen or *M.tb* cultured from CSF	AFB seen or *M.tb* cultured from CSF	AFB seen or *M.tb* cultured from CSF
- Probable TBM	One or more of the following: - Suspected active TB on Xray - AFB from other specimen - Clinical evidence of other EPTB		One or more of the following: - Suspected active TB on Xray - AFB from other specimen - Clinical evidence of other EPTB	One or more of the following: - Suspected active TB on Xray - AFB from other specimen - Clinical evidence of other EPTB	One or more of the following: - Suspected active TB on Xray - AFB from other specimen - Clinical evidence of other EPTB	One or more of the following: - Suspected active TB on Xray - AFB from other specimen - Clinical evidence of other EPTB

Continued on next page

Table 2.1 – *Continued from previous page*

Cohort	**Dexamethasone trial**	**Fluoroquinolone trial**	**TBM HIV cohort**	**ART timing trial**	**Intensified treatment trial**	**Aspirin trial**
- Possible TBM	At least four of the following: - History of TB - Duration of illness >5 days - Altered consciousness - Focal neurologic signs. - Yellow CSF - >50% lymphocytes in the CSF - CSF/blood glucose ratio <0.5			At least two of the following - History of TB - Duration of illness >5 days - GCS <15 - Focal neurological signs And at least two of the following - Yellow CSF - >50% lymphocytes in the CSF - CSF/blood glucose ratio <0.5	At least two of the following - History of TB - Duration of illness >5 days - GCS <15 - Focal neurological signs And at least two of the following - Yellow CSF - >50% lymphocytes in the CSF - CSF/blood glucose ratio <0.5	At least two of the following - History of TB - Duration of illness >5 days - GCS <15 - Focal neurological signs And at least two of the following - Yellow CSF - >50% lymphocytes in the CSF - CSF/blood glucose ratio <0.5
- No TBM	Another diagnosis confirmed by microbiologic or histopathological evaluation.			Another diagnosis confirmed by microbiologic or histopathological evaluation	Another diagnosis confirmed by microbiologic or histopathological evaluation	Another diagnosis confirmed by microbiologic or histopathological evaluation
Confirmed TBM cases	170(31.2%)	39 (63.9%)	54 (93.1%)	158 (62.5%)	407 (49.8%)	92 (76.7%)
Follow up	9 months	9 months	9 months	12 months	9 months	8 months

Continued on next page

Table 2.1 – *Continued from previous page*

Cohort	**Dexamethasone trial**	**Fluoroquinolone trial**	**TBM HIV cohort**	**ART timing trial**	**Intensified treatment trial**	**Aspirin trial**
Primary outcome	9-month mortality	Fluoroquinolone CSF penetration	Not pre-defined (observational study)	9-month mortality	9-month mortality	- Safety endpoint: gastro-intestinal or cerebral bleeding by 60 days - Efficacy endpoint: new brain infarction confirmed or death by 60 days
Main finding	Treatment with dexamethasone was associated with a reduced risk of death (RR 0.69; 95%CI 0.52 to 0.92; P=0.01)	Levofloxacin had highest CSF penetration (median AUC_{0-24} 0.74; range, 0.58 to 1.03)	HIV associated TBM had high mortality (67.2%) and high drug resistance rates. Risk factors were higher TBM grade, low serum sodium and low CSF lymphocyte percentage	Immediate ART did not improve survival (HR, 1.12; 95% CI, 0.87-1.55, P = 0.31). Early ART was associated with more grade 4 adverse events (102 vs. 87 P=0.04)	Intensified anti-TB treatment did not improve 9-month survival (HR, 0.94; 95% CI, 0.73-1.22, P=0.66)	- High Aspirin dose did not associate with an increase in adverse events. Aspirin 1000mg may benefit those with definite TBM (11/32 (34.4%) events in placebo vs. 4/27 (14.8%) in aspirin 81 mg vs. 3/28 (10.7%) in aspirin 1000 mg; P=0.06)

Abbreviations: RCT= randomised controlled clinical trial, ART=antiretroviral therapy, MDR=multidrug resistance, ULN=upper limit of normal, AFB=acid-fast bacilli, *M.tb*= *Mycobacterium tuberculosis*, EPTB=extra pulmonary tuberculosis, GCS=Glasgow coma score, RR= relative risk, CI= confidence interval, HR= hazard ratio.

Continued on next page

Table 2.1 – *Continued from previous page*

Cohort	Dexamethasone trial	Fluoroquinolone trial	TBM HIV cohort	ART timing trial	Intensified treatment trial	Aspirin trial

[1]11 patients had an unknown HIV status
[2]5 patients had an unknown HIV status
[3]The detailed dexamethasone dosing schedule is given in the primary publication (and see Table 1.4)
[4]MDR (multidrug resistance) is defined as resistance of *M.tb* to at least rifampin and isoniazid

The number of recruited patients per year together with the annual mortality risk by HIV status are presented in Figure 2.1. The figure displays a decline in mortality over time, which is particularly pronounced in HIV-infected subjects. In this sub-population, the estimated 9-month mortality dropped from 52-92% during 2001-2007 to 35-50% during 2011-2015. Of note, the number of study participants on ART at enrolment in the first period was only 7/407 (1.7%), and increased to 117/341 (34.3%) in the second period.

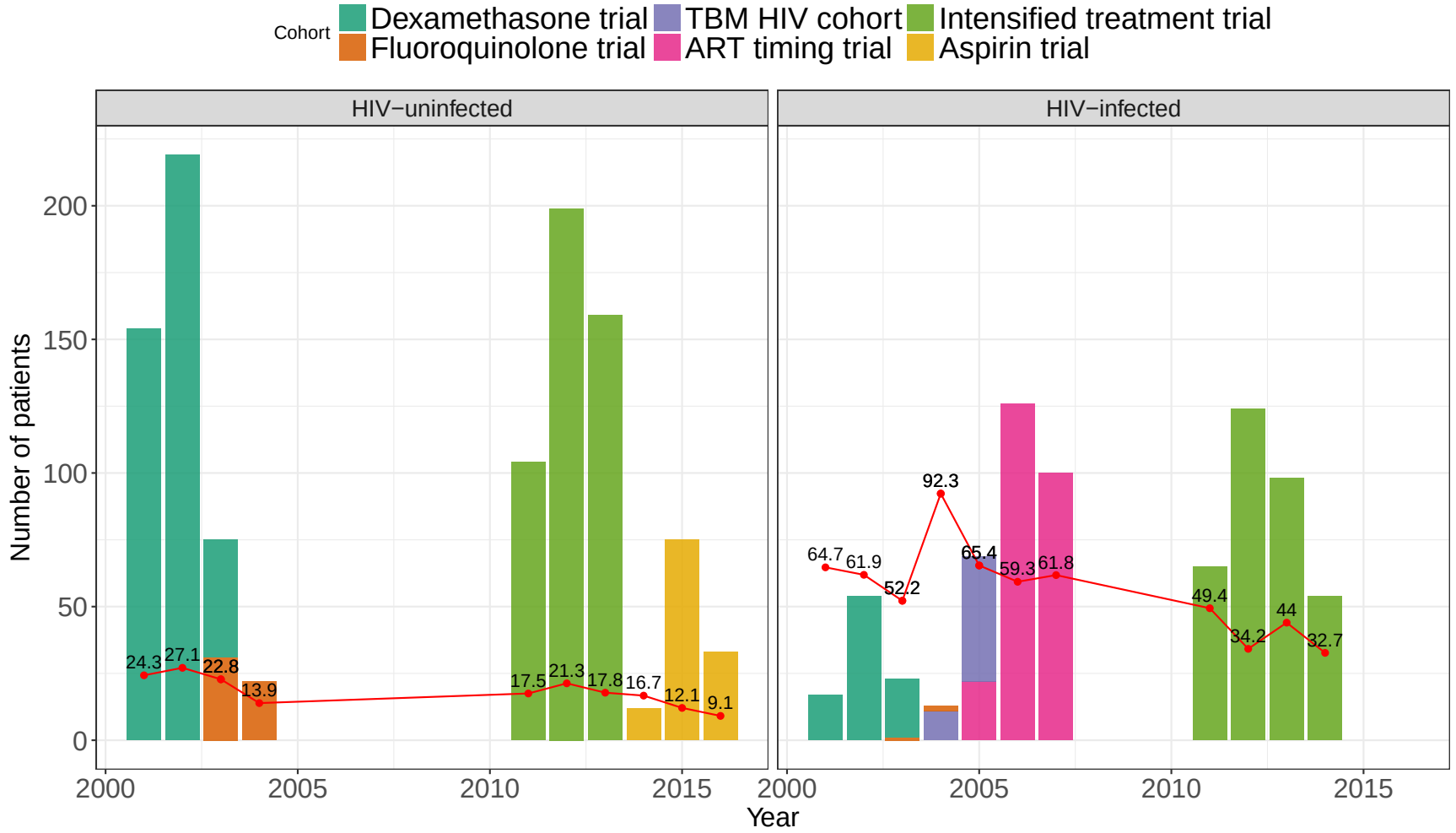


Figure 2.1: *The TBM database. Annual recruitment into the six contributing studies by human immunodeficiency virus status. Red lines show estimated 9-month (8- month for the Aspirin trial) mortality rates (as percentages) for each calendar year, based on the Kaplan-Meier method. Abbreviations: ART, antiretroviral therapy; HIV, human immunodeficiency virus; TBM, tuberculous meningitis.*

2.1.2 Study procedure

At the participating sites, the ward doctors identified patients with suspected TBM for screening. After that, informed consent was sought for those who met the study inclusion criteria. On admission, information on the patients demographic characteristics, clinical examination and laboratory investigation were collected. The clinical assessment determined patient's consciousness status, TBM MRC grade, and any neurological signs and symptoms. The laboratory evaluation included the following tests: haematology (full blood count), biochemistry (sodium, potassium, total protein, albumin, creatinine and liver function tests), cerebrospinal fluid (cell count, protein, glucose, lactate, Gram stain, Ziehl Neelsen (ZN) stain, bacterial and mycobacterial culture), and HIV test. Additional tests for HIV-infected patients included immunology (CD4 count) and virology tests (confirmatory HIV test, plasma HIV-1 RNA, HbsAg) (Heemskerk et al., 2011). A baseline chest radiograph was performed for all patients in all studies, except for those from the Fluoroquinolone trial (Thwaites et al., 2011).

The schedule of procedures for each trial is summarised in Table 2.2. In brief, during hospitalization, patients were monitored daily for clinical events such as death, neurological or drug-related adverse events, and new or recurrent AIDS defining illness (for HIV-infected patients only). Regarding laboratory monitoring, patients normally had weekly haematology and biochemistry assessment, while CSF assessment was often performed monthly. Following discharge, patients were followed-up as part of the Vietnamese National Tuberculosis programme. They were normally checked for haematology and biochemistry monthly until the completion of treatment. For HIV-infected patients, additional immunology and virology tests were performed every three-months. A CSF analysis was performed at 9 months. For HIV-infected patients, another CSF analysis was additionally performed at 6 months in the ART timing trial (Török et al., 2011a) and the TBM HIV cohort study (Torok et al., 2008).

All studies were approved by the Oxford Tropical Research Ethics Committee, and the HTD and Pham Ngoc Thach Hospital Ethics Committees. All RCTs were also approved by the Vietnam Ministry of Health.

Table 2.2: *Schedule of procedures*

Inpatient	**Baseline**	**W1**	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**	**W11**	**W12/M3**
Dexamethasone trial	● ● ●				● ● ●				● ● ●				● ● ●
Fluoroquinolone trial	● ● ●	● ● ●			● ● ●				● ● ●				
TBM HIV cohort	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●	● ● ● ●
ART timing trial	● ● ● ●	● ●	● ●	● ●	● ● ● ●	● ●	● ●	● ●	● ● ● ●	● ●	● ●	● ●	● ● ● ●
Intensified treatment trial	● ● ● ●	● ●	● ●	● ●	● ● ●	● ●	● ●	● ●	● ● ● ●				● ● ● ●
Aspirin trial	● ● ●	● ● ●	● ●	● ●	● ● ●				● ● ●				
Outpatient	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M12**						
Dexamethasone trial						● ● ●							
Fluoroquinolone trial						● ● ●							
TBM HIV cohort			● ● ● ●			● ● ● ●							
ART timing trial	● ●	● ●	● ● ● ●	● ●	● ●	● ● ● ●	● ● ● ●						
Intensified treatment trial	● ●	● ●	● ● ●	● ●	● ●	● ● ● ●							
Aspirin trial					● ●								

Abbreviations: W: Week; M: Month; ● Clinical assessment; ● Blood assessment; ● CSF assessment; ● Immunology assessment

2.1.3 Laboratory investigations

CSF specimens were stained and cultured by standard methods for pyogenic bacteria, mycobacteria, and fungi. In the last two studies (Heemskerk et al., 2016; Mai et al., 2018), CSF was additionally tested with the Xpert MTB/RIF assay (Cepheid, California, USA). Isolates of *M. tuberculosis* were tested for susceptibility to isoniazid, rifampin, ethambutol, and streptomycin by the mycobacterial growth indicator tube method (Ardito et al., 2001). Baseline peripheral blood CD4 cell counts were measured for all HIV-infected adults by flow cytometry.

2.1.4 Creation of the pooled database

A team of experienced clinicians and I defined a list of all potential predictors based on literature review of risk factors for poor outcome in TBM, expert opinion and the availability of the data in the six studies. This included patients demographic information, clinical, laboratory and radiological features assessed at baseline, longitudinal biochemistry and CSF measurements, and outcome (time to death). I then extracted and derived the information from each individual study database and combined them together into a pooled database. I next used the pooled database to reproduce the key results of each study and carefully crosschecked them with the corresponding main publications. All found discrepancies were documented and clarified. I also checked for outliers or implausible values, and compared with the corresponding original Case Report Form (CRF) for correction if needed.

2.2 Common statistical methods in prediction

2.2.1 Model development

2.2.1.1 Treatment of missing data

In Chapter 3 and 5, I imputed missing data using multivariate imputation by chained equations (MICE) which assumes a MAR missing data mechanism (van Buuren, 2007). MICE is a Gibbs sampler, which is a Markov chain Monte Carlo method, based on the property that a multivariate distribution is defined by its conditional distributions (Lesaffre and Lawson, 2012). The algorithm starts with a random draw

from observed data for each variable, and imputes the missing values in a variable-by-variable and iterative fashion (van Buuren and Groothuis-Oudshoorn, 2011). Specifically, variables are imputed one after another in a pre-defined order, and one iteration is completed when all missing values are imputed. The algorithm is repeated for a number of iterations (normally from 5 to 10) to stabilize the result. The whole process is repeated multiple times to obtain multiple imputed data sets. To illustrate how MICE works, lets consider a data set with three variables: age, gender and MRC grade, each containing missing values. The imputation algorithm involves the following steps. First, we temporarily fill in missing values for each variable by randomly sampling with replacement from the observed values. Second, we set imputed values for one variable back to missing, say age. Third, we regress age against gender and MRC grade using all cases with complete information of age. Fourth, missing values of age are then replaced by simulated values drawn from the model's posterior predictive distribution. Steps 2 to 4 are repeated for the second variable, say gender, using the imputed values of the missing age values. Finally, MRC grade is also imputed in the same manner, using the imputed values of age and gender. The imputation finishes a cycle or iteration when all three variables are imputed.

I used logistic regression to impute missing binary variables and predictive mean matching (PMM) to impute missing continuous variables. PMM can be broken down into three steps. First, it calculates the predicted values of a continuous variable based on the specified imputation model. Second, for each missing entry, it identifies q ($q = 3$ is the common choice) observed values that are closest to the corresponding predicted one. Lastly, it randomly chooses one amongst these q values as imputed value for the missing entry. Hence, only observed values are considered for imputation (Little and Rubin, 2002; White et al., 2011).
The imputation models comprised all predictor variables as well as the outcomes, with the main effects and linear terms for the former. Survival outcome was included as a combination of event status and the standard nonparametric Nelson-Aalen estimate of the cumulative hazard function (White et al., 2011).

All analyses in Chapter 3 were performed based on the multiply imputed data sets, except for the univariable analysis which was based solely on the complete cases for each predictor. In Chapter 4, I used a complete case analysis to develop and validate dynamic prediction models.

2.2.1.2 Statistical model

Overall survival is the primary outcome in all studies included in the pooled database. With this type of outcome, both a logistic regression and a Cox proportional hazard regression model could be used. However, the first option requires completely removing patients who were lost to follow-up before the defined time window for mortality, while the second option does not. In addition, a Cox model does not only look at whether an event happened or not, but also when it happened. Hence, Cox proportional hazard regression was the statistical model of choice for building prediction models in Chapter 3 and 4.

In Chapter 5, I used a logistic regression model in the simulated data because of two reasons: i) to be consistent with previous publications on the same topic in the literature, and ii) it is easier than time to event data where one needs to pay attention to the censoring distribution.

2.2.2 Model validation

2.2.2.1 Model performance

The performance of a prediction model is characterized by two main aspects: discrimination and calibration (Collins et al., 2015; Steyerberg, 2008). Discrimination is the ability of a prediction model to discriminate patients who develop the event and those who do not. Patient with an observed event should receive a higher predicted risk by the model. The area under the receiving operating characteristic (ROC) curve (AUC) is the most common measure of discrimination. AUC ranges from 0.5 to 1 with a higher value indicating better discrimination. An AUC of 75% means that there is a 75% chance that the risk prediction is higher for a random patient with an event than for a random patient without an event.

AUC was originally constructed to quantify how well a continuous diagnostic test can discriminate between subjects who truly have the disease according to a gold standard test (cases) versus non-diseased (control). When extending its use to the prediction context for time to event data, several issues should be considered. First, the event status of an individual can vary over time. For example, an individual is event-free when entering the study but can develop the event at later time point. This entails the development of time-dependent AUC. Second, when a subject is right censored, his/her true survival status is unobserved. Ignoring the censoring in estimating AUC may lead to biased estimates (Blanche et al., 2013a). Thus, the

estimation of AUC for time-to-event data requires special estimation methods to accommodate censoring.

Depending on the way cases and controls are defined, one can categorize time-dependent AUC into three different types (Heagerty and Zheng, 2005) cumulative /dynamic, incident /dynamic and incident /static AUC. The cumulative /dynamic AUC quantifies how well a marker can distinguish subjects who develop the event prior to a specific time t (cumulative case) from those who are event-free at t (dynamic control). At any fixed time t, the whole population is divided into two distinct groups based on each individual status at time t: with an event or without an event. Thus, all subjects in the population contribute to the calculation of AUC. Of note in a sample, only subjects who are not censored contribute the calculation of the AUC. An individual can be a control at one time point and then a case at a later time point. The cumulative/dynamic AUC does not distinguish between early and late events; instead, events are accumulated over time (Lambert and Chevret, 2016). The incident/dynamic AUC overcomes that drawback by defining cases as those who fail at an exact time point t (incident case). An individual only contribute as a case at one single time point when it occurs. Finally, the incident/static AUC quantifies how well a marker can distinguish subjects who fail at an exact time t (incident case) from those who are event-free for a certain time window (static control). The static control is also referred to as long-term survivor. Each individual only contributes as either a case or a control for a fixed time window .

Clinicians often focus on a time interval within which the discrimination between case and control is of interest. Therefore, the cumulative/dynamic AUC is the most relevant measure in medical application (Blanche et al., 2013a; Kamarudin et al., 2017; Lambert and Chevret, 2016).

Many estimators for time-dependent AUC have been proposed to account for the presence of censoring (Blanche et al., 2013a, 2012; Kamarudin et al., 2017). One of the standard approaches is based on the Inverse Probability of Censoring Weighting (IPCW)(Hung and Chiang, 2010; Uno et al., 2007). This method accounts for censoring by giving extra weight to individuals who are not censored. The weight is the inverse probability of being uncensored. In both Chapter 3 and 4, I used the cumulative/dynamic AUC to assess the discrimination ability of prognostic models.

Calibration qualifies how closely the predicted risk from the model agrees with the observed risk. It is normally reported graphically using a calibration plot (Collins et al., 2015; Royston and Altman, 2013). In both chapters 3 and 4, I used a calibration plot based on risk groups (Royston and Altman, 2013). Specifically, four risk groups

were created by using cut-off points at the 16th, 50th, and 84th percentiles of the prognostic index from the prediction models, defining "good", "fairly good","fairly poor",and "poor" prognostic subgroups. The prognostic index here is defined as the linear predictor of the Cox model which is monotonically associated with the predicted risk in case there is no stratum variable. Next, I plotted the predicted survival curves and the Kaplan-Meier estimates for the four subgroups in the same figure to examine the agreement between the two estimates. If they align well, it indicates good calibration. The predicted risks were corrected for over-optimism using the validation procedure presented in Section 2.2.2.2. Harrell (2015)(page 506) proposed a similar calibration plot for time-to-event data in which calibration is evaluated at one specific time point (see Figure 3.4). Again, patients are first stratified into risk groups based on their predicted risks at a given time point. Next, the average predicted risks for each stratum are plotted against the respective average observed risk (based on Kaplan-Meier estimates). The difference between the two estimates could be corrected for over-optimism by internal validation (Harrell, 2015). If the points deviate from the diagonal line, it indicates poor calibration. I used the Harrell's calibration plot in Chapter 3.

Another measure to quantify the overall performance of a prediction model is the Brier score. For a binary outcome, the score is the mean squared difference between the observed outcome and the prediction. The Brier score can be mathematically decomposed into two components which subsequently assess the discrimination and calibration of a model (Steyerberg et al., 2010). A lower value of the score indicates a better performance. I used Brier score for binary data to evaluate model performance in Chapter 5. Similar to AUC, calculating the Brier score for a survival outcome requires special techniques to account for censoring (Gerds and Schumacher, 2006; Graf et al., 1999). One could choose between the time-dependent Brier score or the integrated Brier score which measures the overall performance of the model at all times (Gerds et al., 2008). In Chapter 4, I used the time-dependent Brier score to assess the overall performance of the dynamic prediction models.

2.2.2.2 Model validation

As mentioned in Chapter 1, it is very important to validate a prediction model. Model validation is preferably conducted on an independent data set. However, such external data set is often not available and splitting off part of the internal data for

independent validation purposes only is inefficient; therefore internal validation on the same full data set is often performed.

There are several approaches to perform internal validation (Hastie et al., 2009; Wahl et al., 2016). In my thesis, I used the enhanced bootstrap resampling procedure to obtain bias-corrected values of model performance (Harrell (2015), page 114). The method involves several steps. In step 1, I evaluated the prediction model on the original data set to obtain the apparent performance. In step 2, I calculated the bias-corrected performance measure by subtracting the estimated "optimism" from the apparent performance values. The optimism was estimated as follows: I drew 100 bootstrap samples with replacement from the original data set. For each bootstrap data set, I repeated all model building steps. The obtained model was validated on the same bootstrap data set to quantify the bootstrap apparent performance, and on the original data set to quantify the bootstrap test performance. The difference between the bootstrap apparent and the bootstrap test performance was then averaged over 100 bootstrap data sets to obtain an overall estimate of the optimism.
In the presence of multiply imputed data, the apparent performance in the first step is the average of the predictive performances evaluated on each of the m imputed data sets $Imp_i, i = 1, \ldots, m$. In the calculation of the optimism, bootstrap samples were drawn from the original data with missing values. Multiple imputation was performed on each bootstrap sample (denote as $Imp_i^*, i = 1, \ldots, m$). A final model obtained from variable selection was fitted on these data. I validated the obtained model on each of the original imputed data sets (Imp_i) and averaged the results to quantify the bootstrap test performance. I quantified the bootstrap apparent performance in the same manner based on the bootstrap imputed data sets (Imp_i^*). The difference between the bootstrap apparent and the bootstrap test performance was then averaged over 100 bootstrap data sets to obtain an overall estimate of the optimism. The procedure is illustrated in Figure 2.2 (Musoro et al., 2014).

2.2.2.3 Model presentation

Prediction models in Chapter 3 and 4 were implemented in web-based calculators to ease their use in practice (McPherson et al., 2018). In addition, I used a nomogram to graphically display prediction models in Chapter 3. A nomogram is an effective tool in providing an impression of the importance of each predictor to the overall survival probability, and relative to other predictors (Steyerberg, 2008).

2.2.2.4 Statistical software

All analyses were conducted using R software (R Core Team, 2018) with corresponding version and companion packages specified in each chapter.

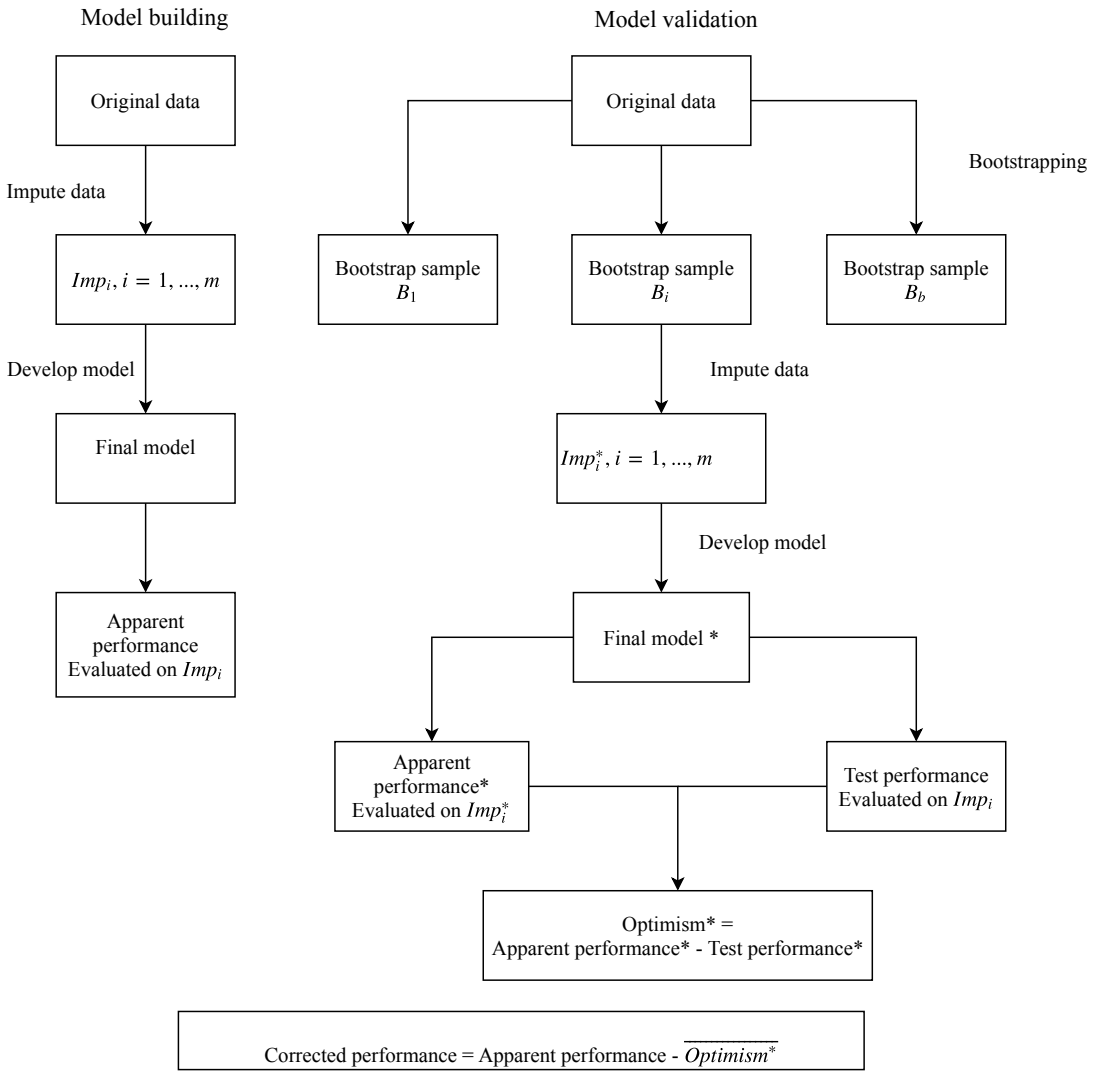


Figure 2.2: *An algorithm to calculate optimism in the presence of multiply imputed data based on the classical enhanced bootstrap resampling procedure (Harrell, 2015).* $*$ *denotes the measurements in bootstrap samples.*

Chapter 3

Prognostic models for 9 month mortality in tuberculous meningitis

Summary: In this chapter, I present the development and validation of prognostic models for 9-month mortality in TBM, using baseline information only. The result of this chapter has been published during my DPhil in the following publications:

- **Le Thi Phuong Thao**, A Dorothee Heemskerk, Ronald B Geskus, Nguyen Thi Hoang Mai, Dang Thi Minh Ha, Tran Thi Hong Chau, Nguyen Hoan Phu, Nguyen Van Vinh Chau, Maxine Caws, Nguyen Huu Lan, Do Dang Anh Thu, Nguyen Thuy Thuong Thuong, Jeremy Day, Jeremy J Farrar, M Estee Torok, Nguyen Duc Bang, Guy E Thwaites, and Marcel Wolbers. Prognostic Models for 9-Month Mortality in Tuberculous Meningitis. *Clinical Infectious Diseases*, 66 (4):523532, feb 2018b. ISSN 1058-4838. doi: 10.1093/cid/cix849. URL `https://academic.oup.com/cid/article/66/4/523/4237651`

- **Le Thi Phuong Thao**, Ronald Geskus, and Guy Thwaites. Reply to Dian et al. *Clinical Infectious Diseases*, 67(12):1955, nov 2018a. ISSN 1058-4838. URL `http://dx.doi.org/10.1093/cid/ciy447`.

3.1 Introduction

The primary objective of this study was to develop and validate novel robust prognostic models for 9-month mortality in adult TBM patients with and without HIV

co-infection. The models were based on a large dataset of 1699 subjects (951 HIV-uninfected, 748 HIV-infected) enrolled in four randomized controlled trials and one prospective cohort study conducted in Vietnam. In addition, we compared the predictive performance of the developed models with the MRC grading system and the GCS, both widely used in the assessment of TBM severity.

3.2 Methods

3.2.1 Study population

The study population comprised subjects enrolled in five out of six TBM studies in the pooled database (Thwaites et al., 2004a, 2011; Torok et al., 2008; Török et al., 2011a; Heemskerk et al., 2016). I did not include data from the last trial (Mai et al., 2018) because it was not available at the analysis time. All study participants were included in the pooled analysis except if they had a confirmed alternative diagnosis, a study drug administration error, or an unknown HIV status. Detailed information regarding study description, inclusion criteria, laboratory investigation, and anti-tuberculosis and adjunctive treatment is presented in Chapter 2.

3.2.2 Primary outcome

The primary endpoint was overall survival during a 9-month follow-up period. Patients without documented death during the follow-up period were censored at 9 months or at the last date they were known to be alive, whichever was earlier. Since anti-tuberculosis treatment usually started immediately following TBM diagnosis, time since TBM diagnosis and time since the initiation of anti-tuberculosis treatment can be seen to refer to the same time point throughout the thesis.

3.2.3 Candidate predictors

Candidate predictors were initially selected based on clinical judgement, their status as established risk factors in previous publications (Erdem et al., 2015; Hosolu et al., 1998; Hsu et al., 2010; Lau et al., 2005; Misra et al., 2000; Thwaites and Tran, 2005; Yasar et al., 2010), and completeness of the data. The number of predictors was further restricted based on a rule of thumb of requiring at least 10 events for each

estimated parameter (Harrell, 2015). The final list of candidate predictors is presented in Table 3.1. For all laboratory parameters and radiology assessments, I used the value recorded closest to enrolment (up to ± 7 days from enrolment). Of note, CSF total white cell count and CSF total lymphocyte count were strongly correlated; therefore only CSF lymphocyte count was included as candidate predictor. Resistance is known to be strongly related to mortality in subjects with TBM (Tho et al., 2012; Thwaites et al., 2005). However, unless earlier isolates are available or rapid molecular tests performed, information about a patients TB drug susceptibility is not available at enrolment. The main models therefore did not include resistance but results from models with resistance are given in the Appendix. For this model, I used resistance information of all patients, not just from those whose results were available at baseline. The cohort variable was included in the full model to represent the change in patient management, treatment, and standard of care over the 15 years' time span of the included studies. I constructed separate prognostic models for the HIV-uninfected and HIV-infected TBM populations because they are clinically distinct populations, and predictors may differ between them.

Table 3.1: *Definition of candidate predictors*

Variable name	**Type**	**Definition**
Cohort	Categorical	5 levels: - Dexamethasone trial - Fluoroquinolone trial - TBM HIV cohort - ART timing trial - Intensified treatment trial
General patient information		
Age	Continuous	Patients age (years)
Sex	Binary	Gender: Male, Female
Weight	Continuous	Patients weight (kg)
Treatment		
Dexamethasone	Binary	Receiving dexamethasone
On ART at enrolment	Binary	On ART at enrolment [in HIV-infected population only]
Baseline clinical variables		

Continued on next page

Table 3.1 – *Continued from previous page*

Variable name	Type	Definition
Baseline clinical variables		
MRC Grade	Categorical	MRC grade: - Grade I: Glasgow coma score (GCS) of 15, without any focal neurological sign - Grade II: GCS of 11 to 14, or a score of 15 with focal neurological signs - Grade III: GSC of 10 or lower
Illness duration at study entry	Continuous	TBM illness duration at study entry (day)
Previous TB treatment	Binary	History of previous TB treatment
Focal neurological signs	Binary	Focal neurological signs: presence of any neurological deficit including hemiplegia, paraplegia, or quadriplegia
Temperature	Continuous	Temperature (degrees Celsius)
Convulsion	Binary	Presence of seizures before enrollment
Baseline laboratory variables		
Plasma sodium	Continuous	Concentration of sodium in blood (mmol/l)
CSF lymphocyte count	Continuous	Number of lymphocytes in CSF (cells/ mm3)
CSF protein	Continuous	Protein level in CSF (g/l)
CSF glucose	Continuous	Glucose level in CSF (mmol/l)
Ratio of blood glucose and CSF glucose	Continuous	Ratio of glucose level in blood and glucose level in CSF
Peripheral blood CD4 count	Continuous	CD4 cell counts in blood (cell/mm3) [in HIV-infected population only]
Baseline radiology variable		
Chest x-ray miliary TB	Binary	Chest x-ray result: miliary tuberculosis

3.2.4 Statistical analysis

3.2.4.1 Transformation of covariates

The duration of illness at enrolment, CSF lymphocyte count, CSF protein, ratio of glucose level in blood and glucose level in CSF, and peripheral blood CD4 cell count had a pronounced right-skewed distribution and were log-2 transformed prior to the analysis. I added 1 unit to the CD4 cell count before applying the transformation to avoid taking the log of zero.

3.2.4.2 Treatment of missing data

Assuming that data are missing at random (MAR), I employed multivariate imputation by chained equations (MICE) to generate imputed data sets (van Buuren and Groothuis-Oudshoorn, 2011). Statistical details regarding the imputation procedure are given in Chapter 2. In the imputation model, the survival outcome was included as a combination of event status and the standard Nelson-Aalen estimate of the cumulative hazard function (White and Royston, 2009). Other potentially relevant variables for the imputation such as hospital and glucose level in blood were included as well. The number of incomplete cases for HIV uninfected and infected patients were respectively 306/951 (32%) and 342/748 (46%) (Figure 3.1). Using the rule of thumb of one imputation per percentage of incomplete cases, I imputed forty-five data sets for each HIV population (White et al., 2011).

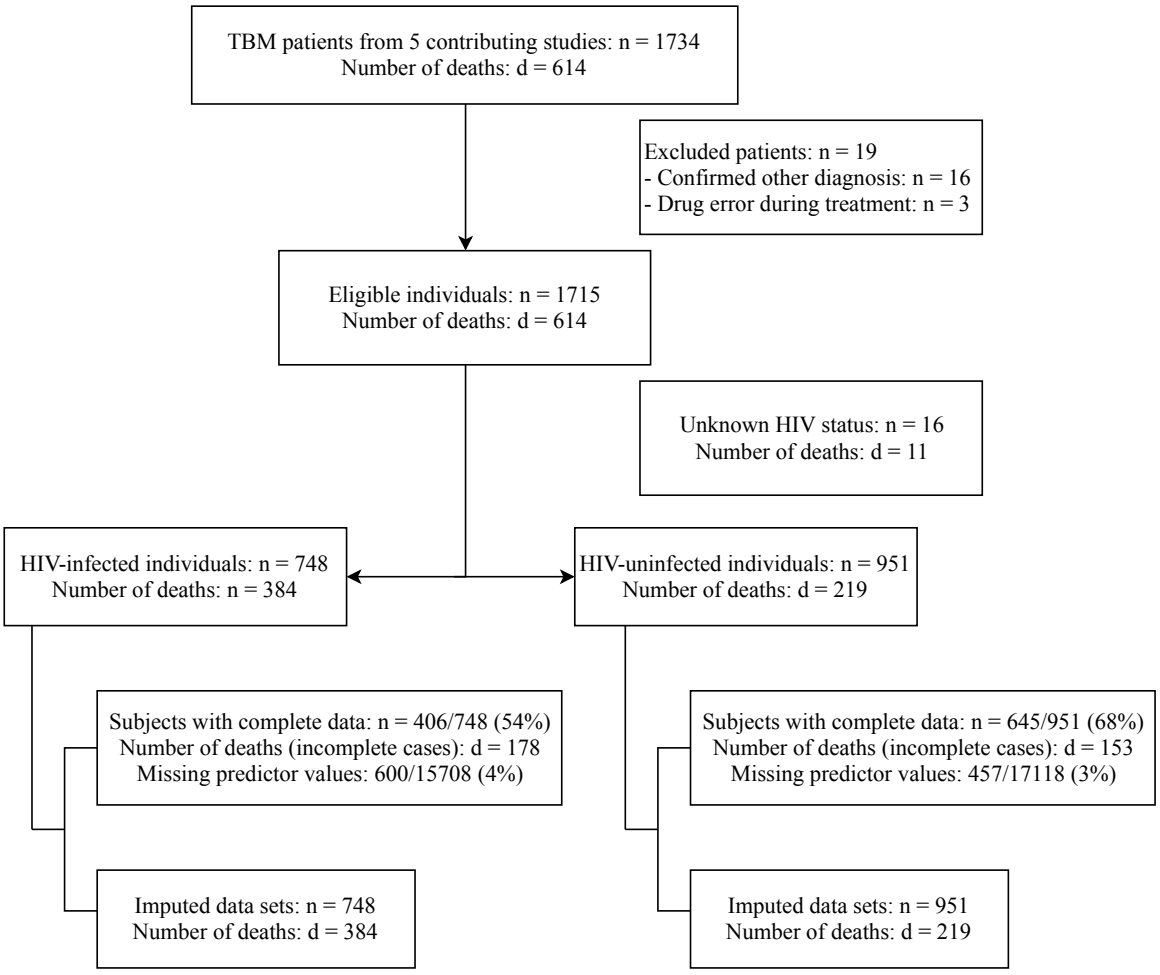


Figure 3.1: *Flow diagram*

3.2.4.3 Model development, validation, and assessment

Statistical models for prediction: I used a Cox proportional hazards regression model. The model building involved 2 steps: i) Evaluating the interaction and non-linear term on the model with all predictors included, and then ii) performing variable selection. Step 1 could be considered as a pre-selection step.

In the first step, I fitted a full model with all pre-specified prognostic variables, with a linear term for all continuous variables and no interactions. Next, I tested for non-linear relationships of the continuous variables by modelling them as restricted cubic spline functions with two degrees of freedom or, equivalently, three knots (Therneau and Grambsch, 2013). Knots were placed at the 10th, 50th, and 90th percentile distribution of each variable. If the spline model was significantly superior (p-value ≤ 0.05 for overall Wald test) to a model with a linear term only, I included the significant non-linear terms. The interaction between age and all other variables was the only biological plausible interaction based on expert opinion. It was pre-defined in the analysis plan. To assess whether age modified the effect of the other predictor variables, I tested for the presence of an interaction between age and all other covariates jointly using an overall Wald test. Again, if the overall test was significant, I included the significant interactions in the full model. The additive assumption was tested after the linearity assumption because the interaction terms could be overstated if the linearity of covariates is falsely assumed (Harrell, 2015).

In the second step, I used two variable selection methods on bootstrap samples of the original data: i) stepwise backwards model selection with a stopping rule based on Akaike's information criterion, and ii) the Least Absolute Shrinkage and Selection Operator (lasso), in which the optimal penalty was determined based on 10-fold cross validation with the partial likelihood as criterion function (Hastie et al., 2009). I chose the largest penalty that was within one standard error from the optimal one (1-se rule). Both methods were applied to 200 bootstrap samples, which were randomly drawn with replacement from each imputed dataset. The proportion of times that each variable appeared in 45 (imputations)*200 (bootstrap samples) = 9000 models was calculated and a variable was selected if this proportion was >60% (Heymans et al., 2007). For the lasso method, a variable represented by more than one parameter (e.g. a factor with > 2 levels, non-linear terms or interaction terms) was retained as a whole if one of its components was selected. For each variable selection method, the final models were then obtained by refitting the selected predictors to each of the 45 imputed datasets and combining results using Rubin's rule (Little and Rubin, 2002; van Buuren, 2012).

Model performance: I used the area under the cumulative/dynamic ROC curve (AUC)(Blanche et al., 2013c) at 9 months to assess the discrimination performance of the models. For calibration, I used calibration plot based on risk groups and Harrell's calibration plot at 9 months (see Chapter 2 (Harrell, 2015)) to visually assess how closely the predicted survival probabilities agreed with the observed probabilities. In each HIV group, model that provided the best predictive performance was chosen as the final model

Model validation: Internal bootstrap validation was performed to correct measures of model performance for overfitting (see Chapter 2). Of note, the pre-selection step in model building process was not repeated in model validation. In addition, I performed temporal validation, wherein data from the Intensified trial (Heemskerk et al., 2016) served as the test data set to validate models developed based on earlier studies (train data set). Here, I performed the imputation step on the test and the train dataset independently. I repeated the two aforementioned variable selection approaches on the train data set with two adaptations. Since none of the HIV-infected patients received ART at enrolment in studies other than the Intensified trial (Heemskerk et al., 2016), this covariate was not included in the temporal validation. The cohort variable was also excluded from the temporal validation.

Model presentation and statistical software: The final prognostic models (with and without resistance as predictor) are graphically depicted by nomograms (Harrell, 2015). Only the final models without resistance have been implemented in a Web-based risk calculator. All analyses were conducted using R version 3.3.1 (R Core Team, 2018) and the following packages: **survival** (Therneau, 2015), **rms** (Harrell Jr, 2016), **mice** (van Buuren and Groothuis-Oudshoorn, 2011), **glmnet** (Friedman et al., 2010), **mgcv** (Wood, 2011), **pec** (Mogensen et al., 2012), and **timeROC** (Blanche et al., 2013b). The web application for risk prediction was programmed using **shiny** (McPherson et al., 2018).

3.3 Results

3.3.1 Baseline characteristics of study participants

The five studies included a total of 1734 subjects, of which 1699 were included in our prognostic model (Figure 3.1). The reasons for excluding 35 subjects were a confirmed

other diagnosis (n = 16), unknown HIV status (n=16), or a study drug administration error (n = 3). Table 3.2 presents the baseline characteristics of all included subjects; 951 were HIV-uninfected (56.0%) and 748 (44.0%) were HIV-infected. The median age was 34 years (the 25% and 75% percentile: 27-45) and 823 (48.4%) had definite, microbiologically confirmed TBM. MRC grade was I in 588 (34.6%) subjects, II in 743 (43.8%) subjects, and III in 367 (21.6%) subjects. Amongst HIV-infected subjects, the median CD4 cell count was 41 cells/mm3, and 124 (16.6%) were on ART at baseline. Compared with HIV-uninfected subjects, those infected with HIV tended to be younger (median 31 vs. 40 years), were more frequently male (85.6% vs. 62.3%) and diagnosed with MRC grade III (25.6% vs. 18.5%). During 9-month follow-up, 219/951 (23.0%) HIV-uninfected and 384/748 (51.3%) HIV-infected subjects with TBM were observed to have died. Amongst the remaining 1096 patients, 998 (91.1%) were followed up for at least 260 days. The Kaplan-Meier estimate of the overall 9-month survival by HIV status is presented in Figure 3.2.

Table 3.2: *Baseline characteristics of patients included in the pooled database, overall and by HIV status*

Characteristic	All TBM patients (N=1699)		HIV-uninfected TBM (N=951)		HIV-infected TBM (N=748)	
	n	Summary statistic	n	Summary statistic	n	Summary statistic
Cohort	1699		951		748	
- Dexamethasone trial		534(31.4%)		436(45.9%)		98(13.1%)
- Fluoroquinolone trial		56(3.3%)		53(5.6%)		3(0.4%)
- TBM HIV cohort		58(3.4%)		0(0%)		58(7.7%)
- ART timing trial		248(14.6%)		0(0%)		248(33.2%)
- Intensified treatment trial		803(47.3%)		462(48.6%)		341(45.6%)
Age [years]	1698	34(27,45)	951	40(27,56)	747	31(26,35)
Sex: Female	1699	514(30.25%)	951	406(42.7%)	748	108(14.4%)
Weight [kg]	1695	46(41,51)	951	46(42,52)	744	45(40,50)
Dexamethasone treatment: Yes	1699	1436(84.5%)	951	742(78.0%)	748	694(92.8 %)
On ART at enrolment: Yes	745	124(16.6%)	-	-	745	124(16.6%)
MRC Grade §	1698		951		747	
- MRC Grade I		588(34.6%)		327(34.4%)		261(34.9%)
- MRC Grade II		743(43.8%)		448(47.1%)		295(39.5%)
- MRC Grade III		367(21.6%)		176(18.5%)		191(25.6%)
Illness duration at study entry [days]	1685	15(10,30)	948	15(10,26)	737	15(9,30)
Previous TB treatment: Yes	1658	236(14.2%)	922	88(9.5%)	736	148(20.1%)
Focal neurological signs: Yes	1683	818(48.6%)	951	517(54.4%)	732	301(41.1%)

Continued on next page

Table 3.2 – *Continued from previous page*

Characteristic	All TBM patients (N=1699)		HIV-uninfected TBM (N=951)		HIV-infected TBM (N=748)	
	n	Summary statistic	n	Summary statistic	n	Summary statistic
Temperature [degrees Celsius]	1697	37.6(37.2,38.5)	950	37.7(37.2,38.5)	747	37.5(37.2,38.5)
Convulsions: Yes	1685	50(3.0%)	950	25(2.6%)	735	25(3.4%)
Plasma sodium [mmol/l]	1537	129(124,133)	843	130(125,134)	694	127(123,132)
CSF lymphocyte count [cells/ mm3]	1614	85(27.6,197.5)	919	94(32.5,200)	695	73 (23.3,188.8)
CSF protein [g/l]	1624	1.3(0.70,1.94)	909	1.2 (0.7,2.0)	715	1.3(0.7,1.9)
CSF glucose [mmol/l]	1636	1.60(1.05,2.30)	920	1.52(1.00,2.30)	716	1.70(1.17,2.33)
Ratio of CSF glucose and blood glucose	1470	0.30(0.21,0.41)	830	0.30(0.20,0.40)	640	0.30(0.21,0.42)
Microbiologically confirmed/definite TBM	1699	823(48.4%)	951	361(38.0%)	748	462(61.8%)
Resistance £	1699		951		748	
- No isoniazid or rifampin resistance		479(28.2%)		203(21.4%)		276(36.9%)
- Isoniazid monoresistance		172(10.1%)		63(6.6%)		109(14.6%)
- Rifampin monoresistance /MDR		35(2.1%)		10(1.1%)		25(3.3%)
- Unknown resistance		1013(59.6%)		675(71.0%)		338(45.2%)
Chest x-ray miliary TB: Yes	1525	279(18.3%)	890	165(18.5%)	635	114(17.9%)

Continued on next page

Table 3.2 – *Continued from previous page*

Characteristic	**All TBM patients (N=1699)**		**HIV-uninfected TBM (N=951)**		**HIV-infected TBM (N=748)**	
	n	**Summary statistic**	**n**	**Summary statistic**	**n**	**Summary statistic**
Peripheral blood CD4 count [cells/mm3]	646	41 (16,108)	-	-	646	41(16,108)

Summary statistics are frequency (%) for categorical variables and mean (25% and 75% percentile) for continuous variables. n refers to the number of subjects with non-missing data for the respective characteristic

£: Isoniazid monoresistance is defined as resistance to isoniazid but not to rifampin. Multidrug resistance (MDR) is defined as resistance to at least isoniazid and rifampin. Unknown resistance is defined as drug-susceptibility test results not available. In all categories, resistance to other drugs may be present.

§: MRC Grade I (GCS 15; no focal neurological signs); MRC Grade II (GCS 11-14, or 15 with focal neurological signs); MRC Grade III (GCS $\leq$10)).

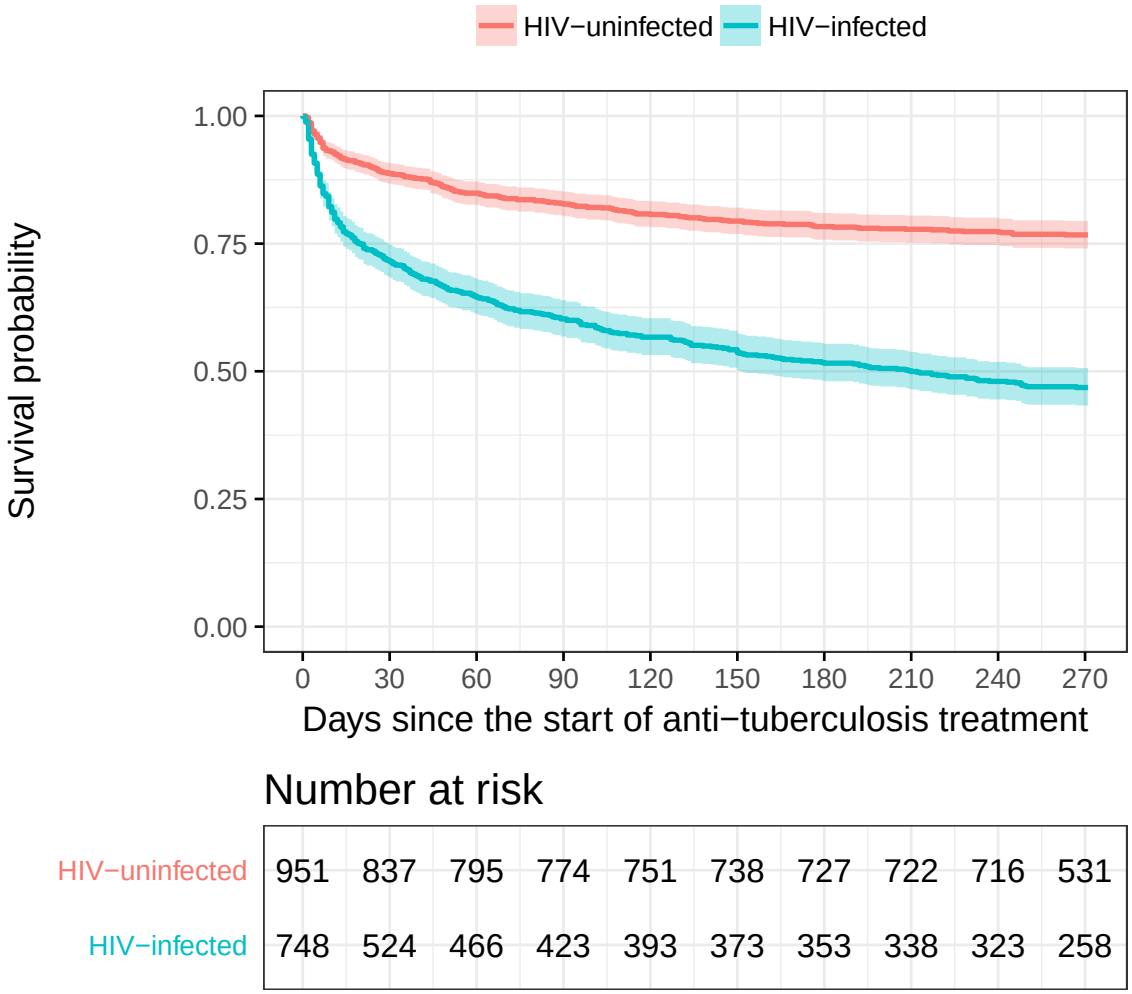


Figure 3.2: *Kaplan Meier survival curves by human immunodeficiency virus (HIV) status*

3.3.2 Prognostic models

Results from the univariable analyses for each candidate predictor are shown by HIV status in Table 3.3 and 3.4. Variables that showed statistically significant associations with death in both HIV groups were age, weight, MRC grade, focal neurological signs, dexamethasone, CSF lymphocyte count, and drug resistant TB. Additional risk factors in HIV-uninfected patients were illness duration and previous TB treatment. In HIV-infected patients, these were temperature, convulsion, plasma sodium, CD4 cell count and ART treatment.

Table 3.3: *Univariable analyses for candidate predictors in the HIV-uninfected TBM population using Cox regression models*

Characteristic	Alive at 9 months or lost to follow-up (N=732)		Dead (N=219)		HR	95% CI	p-value
	n	Summary statistic	n	Summary statistic			
Age [years]∗	732	38 (26,51)	219	48 (35,66)	1.28	1.20 - 1.38	<0.001
Sex: Female	732	313(42.8%)	219	93(42.5%)	1.03	0.78 - 1.34	0.86
Weight [kg] ∗	732	47(42,53)	219	45 (40,50)	0.81	0.69 0.95	0.011
MRC Grade §	732		219				<0.001
- MRC Grade I		288(39.3%)		39(17.8%)	1		
- MRC Grade II		344(47.0%)		104(47.5%)	2.08	1.44 - 3.00	<0.001
- MRC Grade III		100(13.7%)		76(34.7%)	4.72	3.21 - 6.95	<0.001
Illness duration at study entry [days]$	730	15 (10,23)	218	17 (10,30)	1.19	1.05 - 1.34	0.008
Previous TB treatment: Yes	708	51(7.2%)	214	37(17.3%)	2.25	1.58 - 3.21	<0.001
GCS	732	15(13,15)	219	13(8,15)	0.82	0.79 - 0.85	<0.001
Focal neurological signs: Yes	732	365(49.9%)	219	152(69.4%)	2.07	1.55 - 2.75	<0.001
Temperature [Celsius]	731	37.8(37.2,38.5)	219	37.6(37.0,38.5)	0.95	0.82 - 1.11	0.56
Convulsion: Yes	731	19(2.6%)	219	6(2.7%)	1.18	0.53 - 2.66	0.68
Dexamethasone: No	732	142(19.4%)	219	67(30.6%)	1.73	1.30 - 2.31	<0.001
Plasma sodium [mmol/l] ∗	648	130 (125,134)	195	130 (124,134)	1.00	0.83 - 1.21	0.99
CSF lymphocyte count [cells/ mm3] $	711	114(40.65,218.34)	208	47.21(15.92,144.00)	0.82	0.77 - 0.87	<0.001
CSF protein [g/l]	701	1.25(0.71,1.99)	208	1.08(0.60,2.00)	1.01	0.92 - 1.12	0.77
CSF glucose [mmol/l] $	707	1.52(1.00,2.30)	213	1.55(1.00,2.22)	0.93	0.83 - 1.05	0.25

Continued on next page

Table 3.3 – *Continued from previous page*

Characteristic	**Alive at 9 months or lost to follow-up (N=732)**		**Dead (N=219)**		**HR**	**95% CI**	**p-value**
	n	**Summary statistic**	**n**	**Summary statistic**			
Ratio of CSF glucose and blood glucose $	639	0.30(0.20,0.40)	191	0.30(0.24,0.43)	1.17	0.97 - 1.40	0.10
Resistance £	732		219				0.003
- No or other resistance		169(23.1%)		34(15.5%)	1		
- Isoniazid resistant		51(7.0%)		12(5.5%)	1.12	0.58 - 2.16	0.73
- Rifampin resistant/MDR		4(0.6%)		6(2.7%)	5.14	2.16 - 12.27	<0.001
- Unknown resistance		508(69.4%)		167(76.3%)	1.59	1.10 - 2.30	0.014
Chest x-ray miliary TB: Yes	679	119(17.5%)	211	46(21.8%)	1.21	0.87 - 1.68	0.25

Abbreviations: HR=hazard ratio, CI=confidence interval, MDR= multi-drug resistance. Summary statistics are frequency (%) for categorical variables and median (25% and 75% percentile) for continuous variables. Number of censored/lost to follow-up patients before 260 days was 44 (5%). Analysis based on subjects with non-missing data for the respective characteristic.
$: HR per 2-fold increase
$*$: HR per 10-unit increase
£: Isoniazid monoresistance is defined as resistance to isoniazid but not to rifampin. Multidrug resistance (MDR) is defined as resistance to at least isoniazid and rifampin. Unknown resistance is defined as drug-susceptibility test results not available. In all categories, resistance to other drugs may be present.
§: MRC Grade I (GCS 15; no focal neurological signs); MRC Grade II (GCS 11-14, or 15 with focal neurological signs); MRC Grade III (GCS $\leq$10)

Table 3.4: *Univariable analyses for candidate predictors in the HIV-infected TBM population using Cox regression models*

Characteristic	Alive at 9 months or lost to follow-up (N=364)		Dead (N=384)		HR	95% CI	p-value
	n	Summary statistic	n	Summary statistic			
Age [years]∗	364	31(27,36)	383	30(26,35)	0.85	0.73 - 0.98	0.029
Sex: Female	364	53(14.6%)	384	55(14.3%)	1.02	0.77 - 1.36	0.88
Weight [kg] ∗	363	48(44,52)	381	45(40,50)	0.66	0.58 - 0.77	<0.001
MRC Grade §	364		383				
- MRC Grade I		181(49.7%)		80(20.9%)	1		
- MRC Grade II		142(39.0%)		153(40.0%)	2.05	1.56 - 2.68	<0.001
- MRC Grade III		41(11.3%)		150(39.1%)	4.90	3.73 - 6.44	<0.001
Illness duration at study entry [days] $	364	15(9,30)	373	15(9,30)	1.04	0.95 - 1.14	0.42
Previous TB treatment: Yes	362	69(19.1%)	374	79(21.1%)	1.08	0.84 - 1.38	0.56
GCS	364	15(13,15)	383	12(8,15)	0.84	0.82 - 0.86	<0.001
Focal neurological signs: Yes	362	130(35.9%)	370	171(46.2%)	1.39	1.13 - 1.70	0.002
Temperature [Celsius]	364	37.5(37.1,38.3)	383	37.8(37.3,38.5)	1.22	1.09 - 1.36	<0.001
Convulsion: Yes	362	7(1.9%)	373	18(4.8%)	1.81	1.13 - 2.91	0.014
Dexamethasone: No	364	346(95.1%)	384	348(90.6%)	1.43	1.01 - 2.01	0.042
Plasma sodium [mmol/l]∗	339	127 (124,132)	355	126 (121,132)	0.84	0.72 - 0.97	0.015
CSF lymphocyte count [cells/ mm3] $	340	90.90(32.22,214.62)	355	60.90(17.64,160.88)	0.91	0.87 - 0.95	<0.001
CSF protein [g/l]	351	1.30(0.70,1.90)	364	1.25(0.69,1.82)	0.94	0.86 - 1.03	0.20
CSF glucose [mmol/l]	351	1.80(1.20,2.31)	365	1.62(1.12,2.34)	1.06	0.99 - 1.12	0.09

Continued on next page

Table 3.4 – *Continued from previous page*

Characteristic	**Alive at 9 months or lost to follow-up (N=364)**		**Dead (N=384)**		**HR**	**95% CI**	**p-value**
	n	**Summary statistic**	**n**	**Summary statistic**			
Ratio of CSF glucose and blood glucose $	306	0.31(0.22,0.41)	334	0.30(0.21,0.43)	1.06	0.94 - 1.19	0.32
Resistance £	364		384				
- No or other resistance		162(44.5%)		114(29.7%)	1		
- Isoniazid resistant		42(11.5%)		67(17.5%)	1.59	1.18 - 2.16	0.002
- Rifampin resistant/MDR		3(0.8%)		22(5.7%)	4.11	2.59 - 6.50	<0.001
- Unknown resistance		157(43.1%)		181(47.1%)	1.49	1.18 - 1.89	<0.001
Chest x-ray miliary TB: Yes	333	65(19.5%)	302	49(16.2%)	0.84	0.62 - 1.14	0.26
Peripheral blood CD4 count [cells/mm3]$	323	57(23,130)	323	32(12,68)	0.84	0.79 - 0.89	<0.001
On ART at enrolment	364	80(22.0%)	381	44(11.6%)	0.54	0.39 - 0.73	<0.001

Abbreviations: HR=hazard ratio, CI=confidence interval, MDR= multi-drug resistance. Summary statistics are frequency (%) for categorical variables and median (25% and 75% percentile) for continuous variables. Number of censored/lost to follow-up patients before 260 days was 54 (7%). Analysis based on subjects with non-missing data for the respective characteristic.

$: HR per 2-fold increase

∗: HR per 10-unit increase

£: Isoniazid monoresistance is defined as resistance to isoniazid but not to rifampin. Multidrug resistance (MDR) is defined as resistance to at least isoniazid and rifampin. Unknown resistance is defined as drug-susceptibility test results not available. In all categories, resistance to other drugs may be present.

§: MRC Grade I (GCS 15; no focal neurological signs); MRC Grade II (GCS 11-14, or 15 with focal neurological signs); MRC Grade III (GCS ≤10)

Regarding the multivariable analyses, only plasma sodium had a significant non-linear association with survival in HIV-infected patients (p-value < 0.001 for the overall non-linearity test; p-value < 0.001 for the non-linearity of sodium). There was no clear evidence that age modified the association with the outcome of any of the other predictors (p-value for overall interaction test of 0.14 in HIV-uninfected, and of 0.10 in HIV-infected subjects). P-values for the interaction between age and each covariate is reported in Table A3.1. Of note, the p-values were not corrected for multiple testing. The full models and the final models are shown in Table 3.5. Models based on variable selection which were not chosen as final models are shown in Table A3.2. Results of the multivariable analyses where drug resistance was included as a potential predictor are shown in Table A3.3.

Table 3.5: *Final Cox regression models for 9-month survival in each HIV population. Estimates were pooled across multiply imputed datasets.*

Variable	HIV-uninfected TBM population						HIV-infected TBM population					
	Full model			Final model (Model selected by the lasso)			Full model			Final model (Model selected by stepwise backwards model selection)		
	HR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
Age [per +10 years]	1.24	1.15 - 1.34	<0.001	1.24	1.15 - 1.34	<0.001	0.99	0.85 - 1.16	0.89			
Sex: male	1.07	0.79 - 1.44	0.67				0.95	0.70 - 1.30	0.77			
Weight [per +10 kgs]	0.88	0.73 - 1.06	0.17				0.72	0.61 - 0.85	<0.001	0.73	0.63 - 0.85	<0.001
MRC Grade §												
- MRC Grade I	1			1			1			1		
- MRC Grade II	1.36	0.87 - 2.13	0.17	1.41	0.9 - 2.19	0.13	1.71	1.24 - 2.35	0.001	1.88	1.43 - 2.48	<0.001
- MRC Grade III	2.97	1.83 - 4.83	<0.001	3.05	1.92 - 4.86	<0.001	3.76	2.69 - 5.26	<0.001	4.08	3.07 - 5.41	<0.001
Illness duration at study entry [days] ‡	1.03	0.91 - 1.17	0.61				1.02	0.93 - 1.11	0.73			
Previous TB treatment : Yes	1.46	1.00 - 2.13	0.05	1.57	1.09 - 2.26	0.015	1.26	0.97 - 1.65	0.09			
Focal neurological signs: Yes	1.80	1.22 - 2.64	0.003	1.65	1.15 - 2.39	0.007	1.15	0.87 - 1.53	0.33			
Temperature[Celsius]	0.96	0.79 - 1.17	0.71				1.09	0.96 - 1.24	0.18			
Convulsion: Yes	0.96	0.41 - 2.22	0.92				0.95	0.58 - 1.56	0.83			
Dexamethasone: No	1.67	1.15 - 2.40	0.006	1.97	1.46 - 2.67	<0.001	1.12	0.68 - 1.85	0.65			
Plasma sodium [per +10 mmol/l] £	1.01	0.85 - 1.21	0.89						<0.001			<0.001
- 135 vs. 125	1.01	0.85 - 1.21					1.09	0.91 - 1.31		1.07	0.90 - 1.28	
- 115 vs. 125	0.99	0.83 - 1.17					1.63	1.31 - 2.04		1.06	1.29 - 2.00	
CSF lymphocyte count [cells/ mm3] ‡	0.88	0.82 - 0.94	<0.001	0.86	0.81 - 0.92	<0.001	0.93	0.87 - 0.99	0.017	0.93	0.88 - 0.98	0.004
CSF protein [g/l] ‡	0.95	0.84 - 1.07	0.38				0.95	0.85 - 1.06	0.37			
CSF glucose [mmol/l]	1.03	0.87 - 1.23	0.70				1.08	0.98 - 1.19	0.10			

Continued on next page

Table 3.5 – *Continued from previous page*

Variable	HIV-uninfected TBM population						HIV-infected TBM population					
	Full model			Final model (Model selected by the lasso)			Full model			Final model (Model selected by stepwise backwards model selection)		
	HR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
Ratio of CSF glucose and blood glucose ‡	1.02	0.81 - 1.28	0.88				0.91	0.77 - 1.08	0.30			
Chest x-ray miliary TB: Yes	0.82	0.56 - 1.18	0.29				0.95	0.70 - 1.29	0.73			
Peripheral blood CD4 [cells/mm3] ‡	-	-	-				0.91	0.85 - 0.98	0.012	0.9	0.84 - 0.96	0.002
On ART at enrolment [Yes]	-	-	-				0.87	0.60 -1.26	0.45			
Cohort												
- Intensified trial	1						1			1		
- Dexamethasone trial	1.46	0.94 - 2.27	0.09				1.96	1.23 - 3.14	0.005	1.72	1.25 - 2.37	<0.001
- Fluoroquinolone trial	1.12	0.51 - 2.46	0.77									
- TBM HIV cohort							2.58	1.63 - 4.06	<0.001	2.64	1.81 - 3.84	<0.001
- ART timing trial							1.78	1.28 - 2.47	<0.001	1.71	1.34 - 2.19	<0.001

Abbreviations: HR=hazard ratio, CI=confidence interval. 95% CI and p-value for final models do not take into account the uncertainty of model selection.
‡ HR per 2-fold increase.
£ In HIV-infected subjects, the effect of sodium on mortality was significantly non-linear and modelled with a restricted cubic spline function with 2 degrees of freedom. To simplify interpretation of the corresponding regression coefficients, only HRs for two derived sodium contrasts from that model are given. P-values for sodium in the HIV-infected population are based on overall Wald-tests of the restricted cubic spline function
§ MRC Grade I (GCS 15; no focal neurological signs); MRC Grade II (GCS 11-14, or 15 with focal neurological signs); MRC Grade III (GCS≤10)).

The final models after variable selection identified a higher Medical Research Council (MRC) disease severity grade and lower cerebrospinal fluid lymphocyte cell counts as risk factors for mortality in both HIV-uninfected and HIV-infected subjects (Table 3.5). Importantly, the mortality rate in MRC grade III (GCS $\leq$ 10) was three-fold higher in HIV-uninfected patients and four-fold higher in HIV-infected patients compared to MRC grade I (GCS 15; no focal neurological signs). Mortality for MRC grade II (GCS 11-14, or 15 with focal neurological signs) was intermediate in both populations. In HIV-uninfected subjects, older age, previous tuberculosis, not receiving adjunctive dexamethasone, and focal neurological signs were additional predictors of higher mortality. Of note, the mortality rate was increased by more than 50% in patients with previous tuberculosis and those with focal neurological signs, and almost doubled for patients who did not receive dexamethasone. In the HIV-infected population, lower weight, lower peripheral blood CD4 cell count, and lower or higher plasma sodium were additional predictors of higher mortality. Figure 3.3 illustrates the association of sodium with sodium 127 mmol/l as the reference (i.e. the hazard ratio is 1 for sodium 127 mmol/l). Both decreased and elevated sodium values were associated with a higher predicted mortality compared to intermediate values. Ongoing ART therapy at enrolment was associated with a more favourable outcome in univariable analysis (Table 3.4), but was no longer significantly associated with survival after adjusting for the other factors. In the final models with drug resistance information (Table A3.3), this predictor was identified as a risk factor for mortality in both HIV populations. The mortality rate in patients with rifampin mono-resistance or MDR was more than five times higher than in those who were not resistant to isoniazid or rifampin.

3.3.3 Model validation

The final models showed good discrimination between TBM survivors and deaths at 9 months; the corresponding AUCs in internal validation (corrected for over-fitting) were 0.77 in the HIV-uninfected population and 0.78 in the HIV-infected population, respectively (Table 3.6). The addition of drug resistance as a covariate resulted in only a slightly improved model discrimination with AUC values ranging between 0.77-0.79 (Table 3.6).

Table 3.6: *Discrimination of candidate models for TBM mortality by HIV population (with and without drug resistance information) measured by the area under the cumulative/dynamic ROC curve (AUC) at nine months.*

Model	Internal validation		Temporal validation
	Apparent AUC (95% CI)∗	**Optimism Corrected AUC** ‡	**AUC (95% CI)**
Model without drug resistance information			
HIV-uninfected population			
Full model	0.79 (0.76-0.83)	0.76	0.77 (0.72-0.83)
Model selected by stepwise backwards model selection	0.78 (0.74-0.81)	0.76	0.77 (0.71-0.82)
Model selected by the lasso method **	0.78 (0.75-0.82)	0.77	0.82 (0.77-0.87)
MRC Grade §	0.66 (0.62-0.70)	0.66	0.70 (0.64-0.75)
GCS	0.68 (0.64-0.72)	0.68	0.68 (0.62-0.75)
HIV-infected population			
Full model	0.79 (0.75-0.83)	0.77	0.76 (0.70-0.82)
Model selected by stepwise backwards model selection **	0.79 (0.75-0.82)	0.78	0.73 (0.67-0.79)
Model selected by the lasso method	0.78 (0.74-0.82)	0.77	0.75 (0.69-0.81)
MRC Grade §	0.70 (0.66-0.74)	0.70	0.69 (0.63-0.75)
GCS	0.71 (0.67-0.75)	0.71	0.68 (0.62-0.74)
Model with drug resistance information			
HIV-uninfected population			
Full model	0.80 (0.76-0.83)	0.77	0.78 (0.72-0.83)
Model selected by stepwise backwards model selection	0.79 (0.75-0.82)	0.77	0.76 (0.70-0.82)
Model selected by the lasso method **	0.79 (0.76-0.83)	0.77	0.80 (0.75-0.86)
HIV-infected population			
Full model	0.81 (0.78-0.85)	0.79	0.78 (0.73-0.84)
Model selected by stepwise backwards model selection **	0.80 (0.77-0.84)	0.79	0.78 (0.73-0.84)
Model selected by the lasso method	0.81 (0.77-0.84)	0.79	0.78 (0.73-0.84)

Abbreviations: CI=confidence interval, GCS = Glasgow coma score.
∗ Refers to performance estimated directly from the original 45 imputed data sets that were used to develop the prediction models
** Final chosen model
‡ Adjusted performance corrected for over-optimism through internal bootstrap validation.
§ MRC Grade I (GCS 15; no focal neurological signs); MRC Grade II (GCS 11-14, or 15 with focal neurological signs); MRC Grade III (GCS ≤10)).

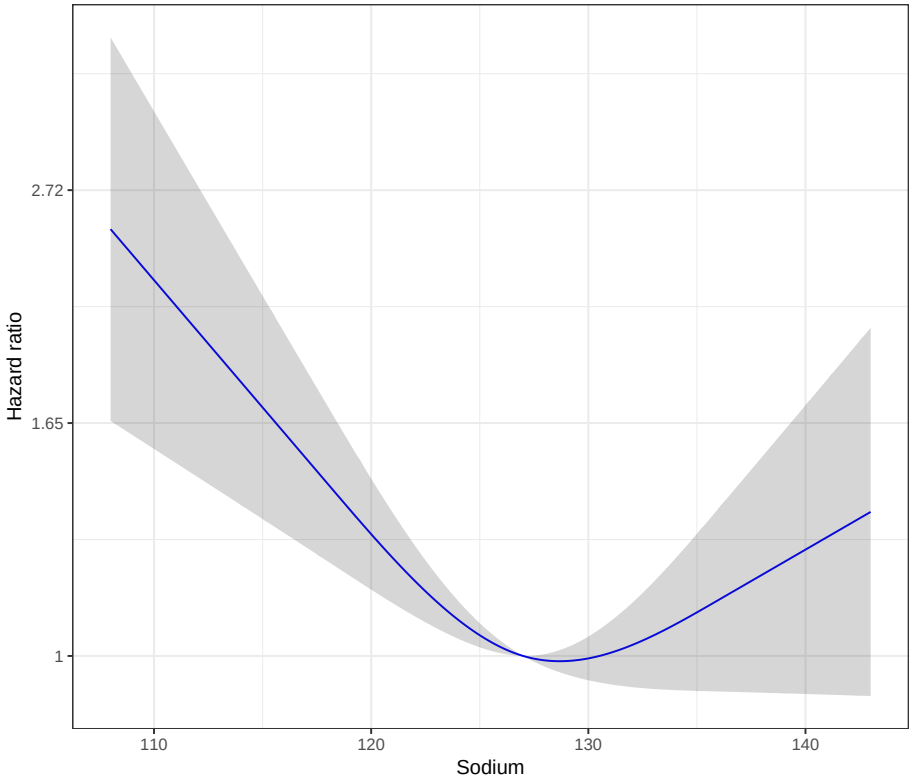
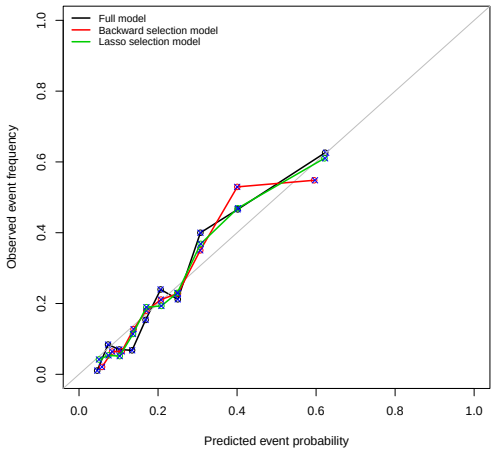


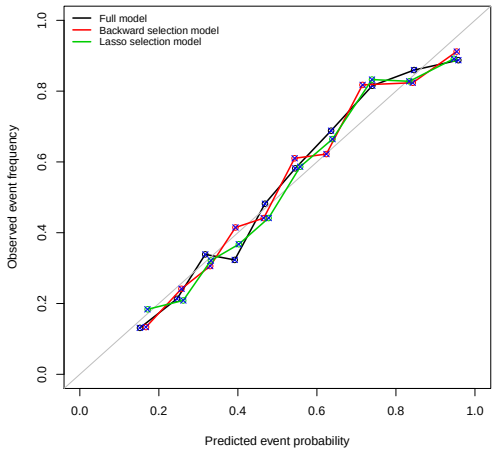
Figure 3.3: *Non-linear effect of plasma sodium (mmol/L) in HIV-infected patients. The plot is generated based on one imputed data set.*

In terms of calibration, the models showed good agreement between predicted and observed mortality in internal validation (Figures 3.4a and 3.4b). In temporal validation, calibration in the HIV-uninfected groups remained satisfactory whereas in the HIV infected group, models developed on earlier studies systematically overestimated the actual mortality in the most recent trial (Figures 3.4c and 3.4d).

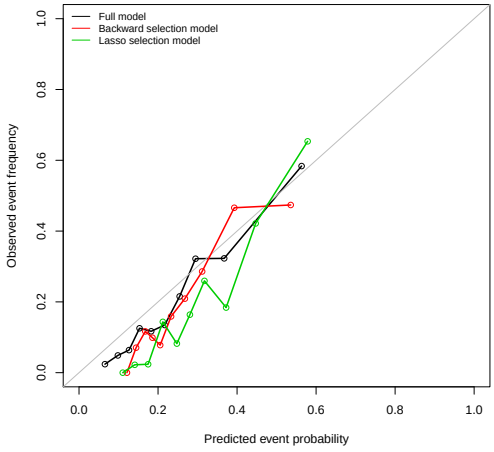
Figure 3.5 shows predicted survival curves of the final simplified models in four risk groups, defined using cut-off points at the 16th, 50th, 84th percentiles of the prognostic index generated by these models (Royston and Altman, 2013). They show strong prognostic separation across risk strata, good agreement between predicted survival curves and Kaplan-Meier estimates using the full dataset, and mild to substantial risk overestimation in temporal validation in HIV-uninfected and HIV-infected subjects.



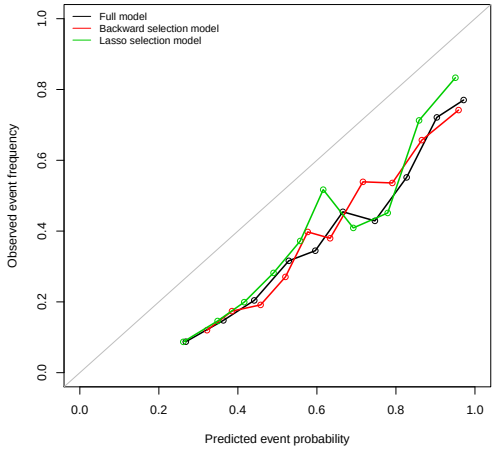
(a) *HIV-uninfected, internal validation*



(b) *HIV-infected, internal validation*



(c) *HIV-uninfected, temporal validation*



(d) *HIV-infected, temporal validation*

Figure 3.4: *Calibration plots of candidate models by HIV status in internal and temporal validation. The circles and lines display the predicted 9 month mortality (from the prognostic model) versus the observed 9-month month mortality (based on Kaplan-Meier estimation) in ten patient groups defined by quantiles of the predicted event probabilities. Each circle represents one quantile group. For internal validation, blue crosses display bootstrap bias-corrected Kaplan-Meier estimates.*

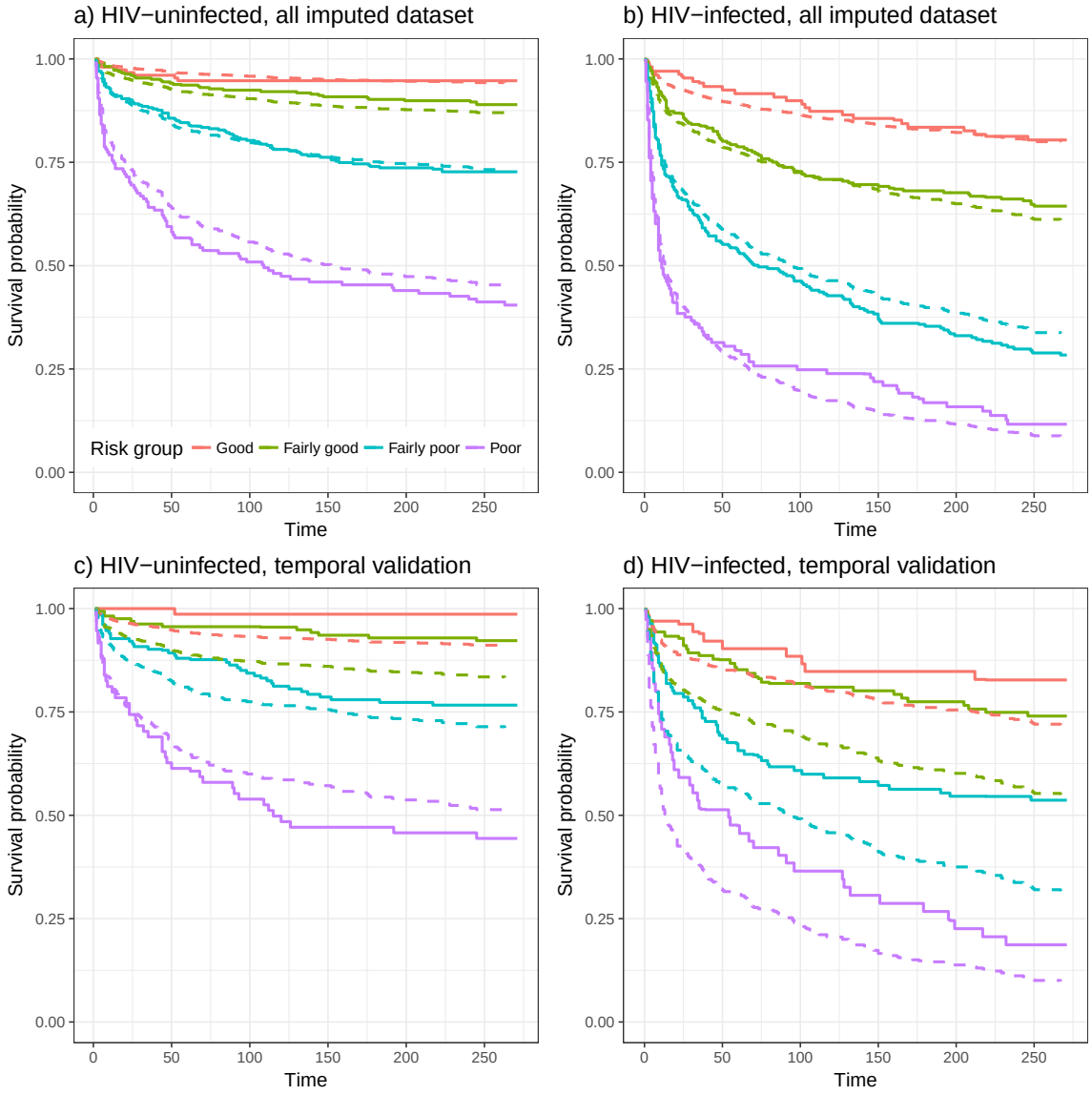
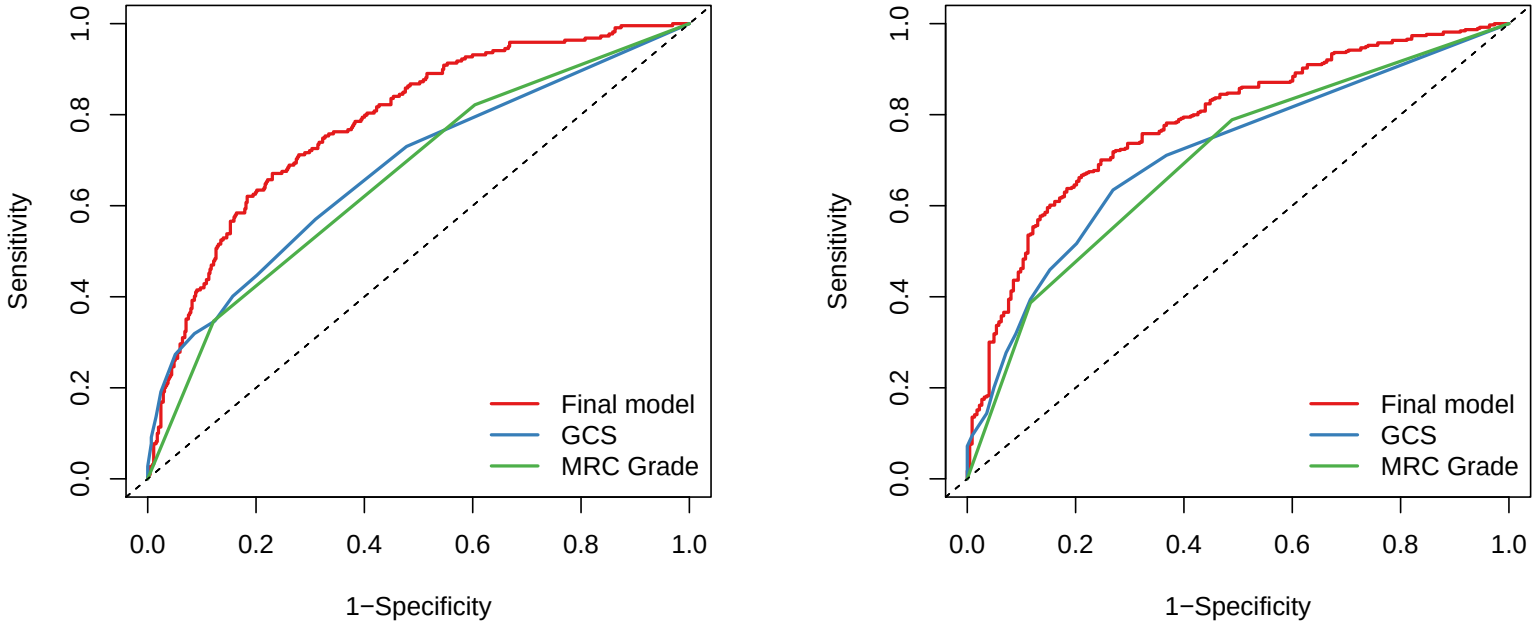


Figure 3.5: *Comparison between predicted survival of the final models (dashed lines) and observed Kaplan-Meier estimates (solid lines). Risk groups are defined using cut-points at the 16th, 50th, 84th percentiles of the prognostic index as generated by the final models (defining good, fairly good, fairly poor, and poor prognostic subgroups).*

3.3.4 Comparison of prognostic models with MRC grade and GCS alone

MRC grade or GCS alone predicted mortality similarly well (AUC ranging from 0.66-0.71) (Table 3.6). The corresponding ROC curves for the MRC grade, GCS and the final models are shown in Figure 3.6. MRC grade and GCS were both clearly

inferior to the developed multivariable models in terms of discrimination (p-values for tests of equality between AUC values < 0.001 for all four comparisons) whereas discrimination did not differ significantly between MRC grade and GCS (p-value $= 0.23$ in HIV uninfected and p-value $= 0.19$ in HIV infected).



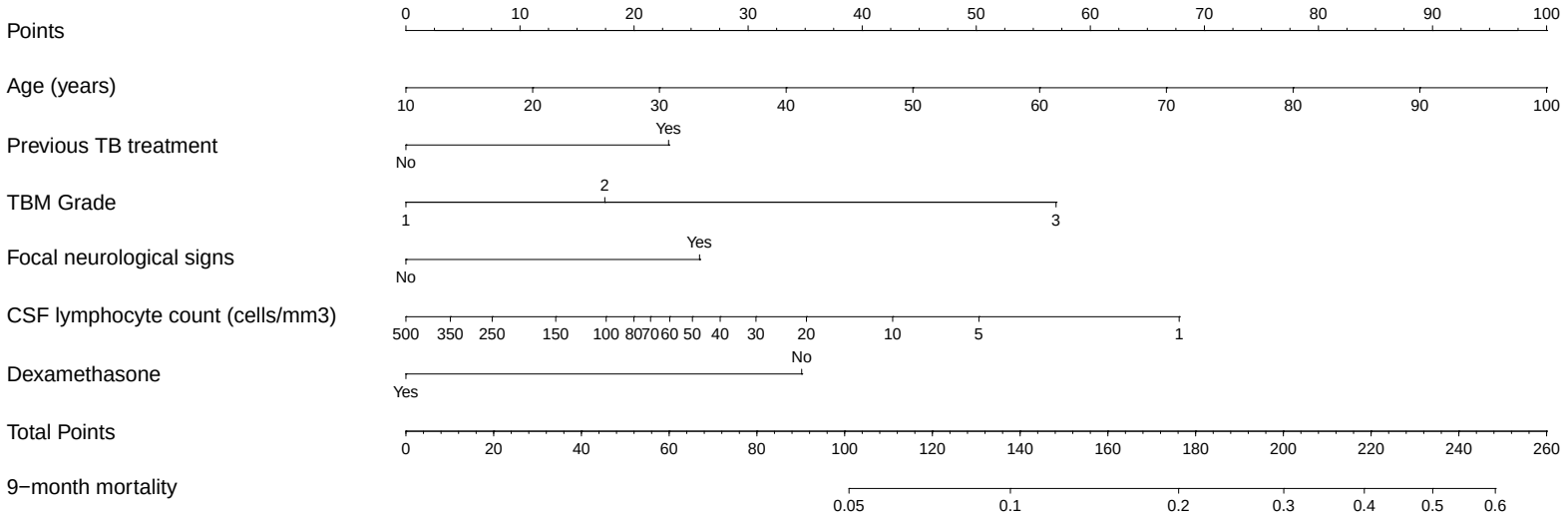
(a) *HIV-uninfected TBM population*

(b) *HIV-infected TBM population*

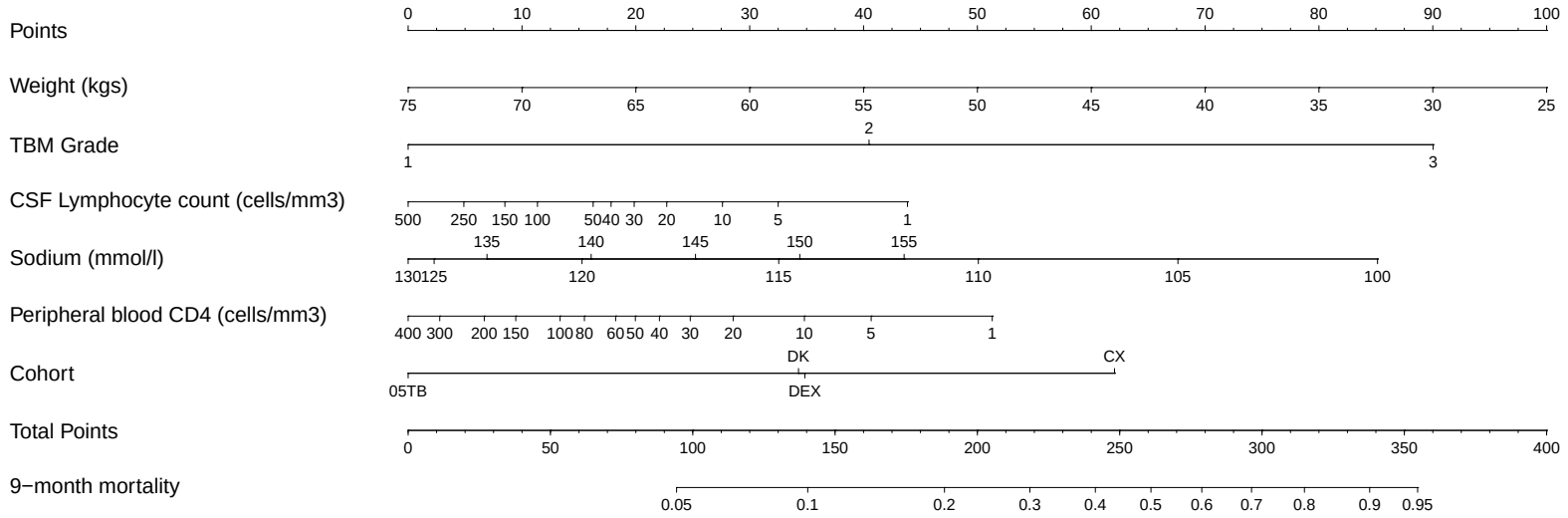
Figure 3.6: *Cumulative/dynamic ROC curves (apparent estimate) for mortality evaluated at 9 months for the final prognostic model, GCS, and MRC Grade in the HIV-uninfected and HIV-infected TBM populations.*

3.3.5 Nomograms and web app for mortality prediction

Figure 3.7 visualizes the final models as nomograms. For clinical use, this model has also been implemented as a user-friendly web app available at `https://thaole.shinyapps.io/tbmapp/` . I also generated nomograms for use when resistance information is available (Figure 3.8). Of note, for creation of the nomogram and the web-app in the HIV-infected population, I chose our most recent trial as value for the cohort variable (Heemskerk et al., 2016) since this is most relevant for future prediction.

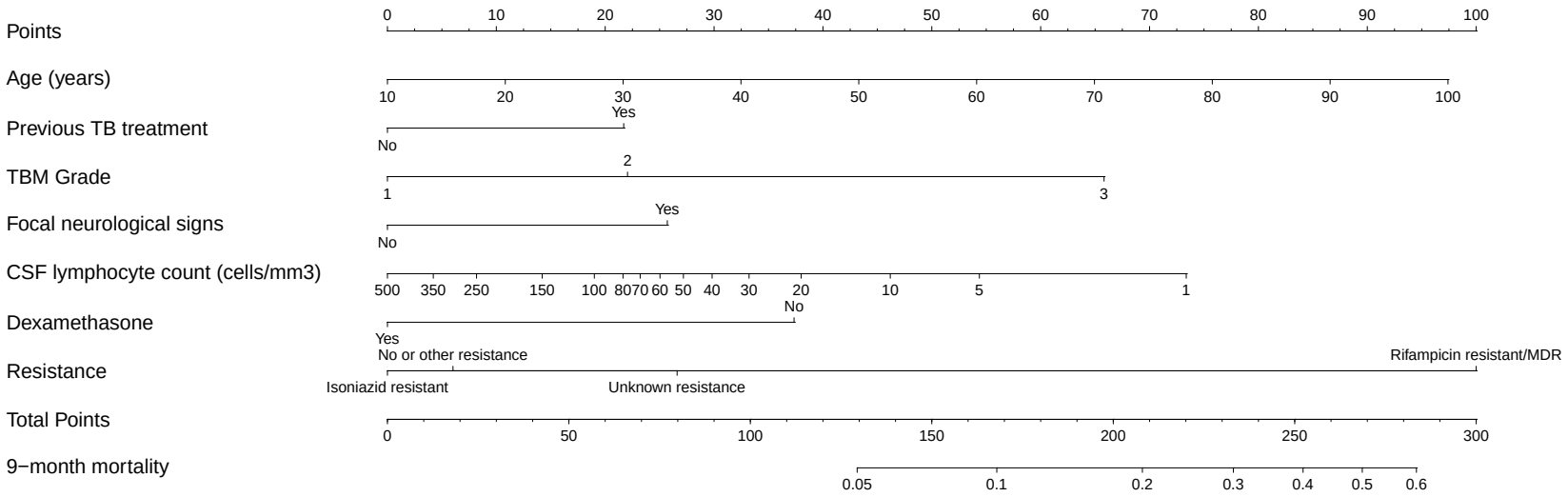


(a) *HIV-uninfected TBM population*

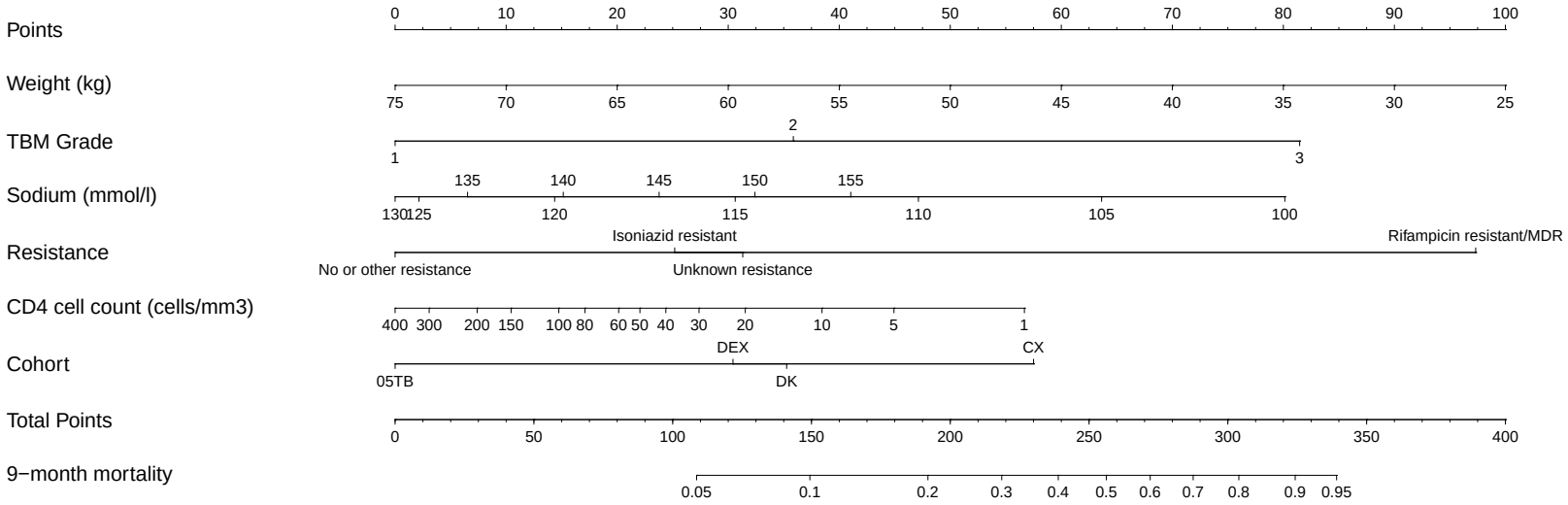


(b) *HIV-infected TBM population*

Figure 3.7: *Nomograms for the prediction of 9-month mortality based on the final prognostic models for the HIV-uninfected and HIV-infected TBM populations, respectively. To derive a prediction, locate the value of each predictor on the corresponding variable line, read the corresponding points assigned on the 0-100 scale and sum all of these points to a total point score. Then read the result on the Total Points scale and its corresponding prediction below. For HIV-infected nomogram, the cohort variable was chosen as the most recent trial (Heemskerk et al., 2016) since this is most relevant for future prediction*



(a) *HIV-uninfected TBM population*



(b) *HIV-infected TBM population*

Figure 3.8: *Nomograms for the prediction of 9-month mortality based on the final prognostic models for the HIV-uninfected and HIV-infected TBM populations respectively including drug resistance information. To derive a prediction, locate the value of each predictor on the corresponding variable line, read the corresponding points assigned on the 0-100 scale and sum all of these points to a total point score. Then read the result on the Total Points scale and its corresponding prediction below. For HIV-infected nomogram, the cohort variable was chosen as the most recent trial (Heemskerk et al., 2016) since this is most relevant for future prediction*



(a) *HIV-uninfected TBM population*



(b) *HIV-infected TBM population*

Figure 3.9: *Screenshot for the web application*

3.4 Discussion

I developed and validated prognostic models for 9-month mortality after TBM diagnosis in a large dataset of 951 HIV-uninfected and 748 HIV-infected TBM patients. I found that a higher MRC grade and lower CSF lymphocyte cell counts were predictors for mortality in both groups. MRC grade assesses the severity of neurological impairment and numerous studies have shown its strong association with poor outcome (Kalita and Misra, 1999; Misra et al., 1996). The predictive value of CSF lymphocyte counts is intriguing. Others have recently found higher CSF neutrophils predicted death from TBM (van Laarhoven et al., 2017); both findings reflect the poorly understood importance of neuro-inflammation with respect to outcome from TBM.

Additional predictors for mortality in HIV-uninfected patients were older age, previous TB treatment, not receiving dexamethasone, and having focal neurological signs. In HIV-infected subjects, additional risk factors for mortality were lower weight, lower CD4 cell count, and abnormal plasma sodium level. These finding are consistent with results from previous published studied conducted in India (Yasar et al., 2010), Hong Kong (Lau et al., 2005), Taiwan (Hsu et al., 2010), Turkey (Hosoglu et al., 2002), South Africa (Marais et al., 2011), and Brazil (Croda et al., 2010). The inclusion of different predictors in the two HIV populations may suggest important differences in disease pathogenesis and outcomes in these two subgroups.

The main analysis did not include drug-resistance as a covariate because this information is often not available at enrolment. Alternative models including drug resistance showed only relatively little improvement in model performance, especially in HIV-uninfected subjects. However, I note the low prevalence of multidrug resistant TB (MDR-TB), an important predictor of poor outcome in TBM (Tho et al., 2012; Thwaites et al., 2005), in the TBM cohort which may have contributed to this finding.

The final prediction models were carefully developed and validated with proper statistical methods (Collins et al., 2015). They showed good discrimination between TBM-survivors and deaths in both internal and temporal validation, and substantially improved over the MRC grade or the GCS alone. Although the models had excellent calibration in internal validation, they over-estimated mortality in temporal validation where the model was developed on earlier studies and tested on subjects included in the most recent trial. This overestimation was mild in HIV-uninfected subjects but substantial in HIV-infected subjects. It may be explained by improvements in treatment, especially the availability of ART, and patients' supportive care over time.

Importantly, because ART only became widely available in Vietnam in 2005 (Nguyen et al., 2013) and being on ART was an exclusion criterion for the ART timing trial, the intensified treatment trial (Heemskerk et al., 2016) was the only study that included subjects on ART at TBM diagnosis. Therefore, the effect of ART was neglected in the temporal validation, and this may have contributed to the observed overestimation. Temporal heterogeneity was addressed by including the cohort as a covariate in the statistical regression models. Survival for future patients is anticipated to be close to predictions adjusted to the most recent study (Heemskerk et al., 2016).

The study has several limitations, mostly owing to the characteristics of the TBM database. Firstly, CSF lactate, which may be an important risk factor (Thwaites et al., 2003), was not taken into account in the final model because it was missing in $> 40\%$ of subjects. Secondly, previous studies suggested that leukotriene A4 hydrolase (LTA4H) genotype is a determinant of the inflammatory response in HIV-uninfected adults with TBM, and consequently might predict who benefits from adjunctive corticosteroid treatment (Thuong et al., 2017). This factor, therefore, might have an influence on survival of HIV-uninfected TBM subjects. As the information of LTA4H genotype was not available for all included studies, I did not consider this covariate in these analyses. Finally, the models were developed and validated based on data from two large hospitals in Southern Vietnam only. Thus, it would be desirable to validate and possibly improve the models in an independent external database.

After publishing in the Clinical Infectious Disease journal, the model was validated by an independent research group. The validation data set consisted of 524 Indonesian HIV-uninfected patients with TBM. The validation showed that the model did not perform well, with an AUC value of 0.6 (Dian et al., 2018). The discrepancy, however, may be explained by the biological difference between the two populations (Thuong et al., 2017; van Laarhoven et al., 2017). For example, LTA4H genotype influenced CSF inflammation and survival in the Vietnamese cohort but not in the Indonesian cohort (Thuong et al., 2017; van Laarhoven et al., 2017). In addition, it has been shown that high CSF neutrophil numbers are associated with the culture of *Mycobacterium tuberculosis* from CSF in Vietnam (Thwaites et al., 2004c), but unlike the Indonesian cohort, lower neutrophil numbers were associated with death (Thwaites et al., 2003). Studies of TBM in HIV co-infected patients in South Africa have also reported that high CSF neutrophil numbers predicted positive CSF mycobacterial cultures and, intriguingly, the development of CNS immune reconstitution inflammatory syndrome which is an unfavourable outcome (Marais et al., 2013, 2014). These studies highlight shared and discrepant pathophysiological mechanisms

between populations with TBM, which may be driven by genomic variation in the bacteria or their hosts and may undermine attempts to define universally applicable clinical prognostic models. This validation study again highlights the importance of validating prognostic models in different settings before they can be widely applied in clinical practice.

In conclusion, the presented models are based on one of the largest sources of prospectively collected clinical data for TBM disease worldwide. The final prognostic models included variables which are usually available for patients with TBM. Overall, they have good discrimination and calibration performance after correcting for optimism by internal validation, and clearly outperformed the MRC grading score. However, they only included risk factors available at TBM diagnosis. Changes in biomarkers that are repeatedly measured during the disease course may carry important information which could allow updating and improving an initial prognosis during TB treatment. The value of such longitudinally measured biomarkers is examined in the next chapter.

3.5 Appendix

Table A3.1: *P-values for the interaction between age and other covariates. Of note, the p-values were not corrected for multiple testing.*

Variable	HIV-uninfected population	HIV-infected population
Weight	0.050	0.445
Cohort	0.004	0.356
Convulsion	0.742	0.578
Chest X-ray miliary TB	0.038	0.692
Dexamethasone treatment	0.002	0.866
Focal neurological signs	0.937	0.300
CSF glucose	0.593	0.465
Illness duration	0.256	0.145
CSF Lymphocyte	0.542	0.600
Previous TB treatment	0.766	0.269
CSF Protein	0.078	0.529
Ratio of CSF glucose and blood glucose	0.414	0.748
Sex	0.508	0.211
Plasma sodium	0.478	0.097
MRC Grade	0.051	0.141
Temperature	0.117	0.763
Peripheral blood CD4		0.097
On ART at enrolment		0.140

Table A3.2: *Models based on variable selection which were not chosen as final models in HIV-uninfected and HIV-infected TBM populations.*

Variable	HIV-uninfected TBM			HIV-infected TBM		
	Backward selection model			Lasso selection model		
	HR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
Age [years]∗	1.24	1.15 - 1.33	<0.001			
Weight [kg] ∗				0.93	0.90 - 0.96	<0.001
MRC Grade §						
- MRC Grade I	1			1		
- MRC Grade II	1.93	1.33 - 2.79	<0.001	1.89	1.44 - 2.49	<0.001
- MRC Grade III	4.43	3.01 - 6.52	<0.001	4.37	3.30 - 5.79	<0001
Dexamethasone: No	1.72	1.29 - 2.29	<0.001			
CSF lymphocyte count [cells/ mm3] ‡	0.85	0.80 - 0.91	<0.001			
Peripheral blood CD4 count [cells/mm3]				0.90	0.84 - 0.96	0.002
On ART at enrolment [Yes]				0.92	0.64 - 1.32	0.64
Cohort						
- Intensified treatment trial				1		
- Dexamethasone trial				1.59	1.13 - 2.23	0.008
- TBM HIV cohort				2.35	1.60 - 3.45	<0.001
- ART timing trial				1.55	1.19 - 2.02	0.001

95% confidence intervals and p-value for final models do not take into account the uncertainty of model selection.
Abbreviations: HR=hazard ratio, CI=confidence interval.
∗: HR per 10-unit increase.
‡: HR per 2-fold increase.
§: MRC Grade I (GCS 15; no focal neurological signs); MRC Grade II (GCS 11-14, or 15 with focal neurological signs); MRC Grade III (GCS $\leq$ 10)).

Table A3.3: *Final Cox regression models for 9-month survival in each HIV population including drug resistance information. Estimates were pooled across multiply imputed datasets.*

Variable	HIV-uninfected TBM population						HIV-infected TBM population					
	Full model			Final model (Model selected by the lasso)			Full model			Final model (Model selected by backwards stepwise selection method)		
	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value
Age [per +10 years]	1.23	1.14 - 1.33	<0.001	1.23	1.14 - 1.32	<0.001	0.96	0.82 - 1.12	0.64			
Sex: male	1.04	0.77 - 1.41	0.78				0.98	0.72 - 1.35	0.92			
Weight [per +10 kgs]	0.88	0.73 - 1.06	0.20				0.71	0.60 - 0.84	<0.001	0.70	0.60 - 0.82	<0.001
MRC Grade §												
- MRC Grade I	1			1			1			1		
- MRC Grade II	1.45	0.93 - 2.27	0.10	1.51	0.97 - 2.34	0.07	1.73	1.26 - 2.39	<0.001	1.89	1.43 - 2.48	<0.001
- MRC Grade III	3.32	2.03 - 5.43	<0.001	3.43	2.14 - 5.48	<0.001	3.95	2.82 - 5.53	<0.001	4.22	3.18 - 5.60	<0.001
Illness duration at study entry [day] ‡	1.05	0.92 - 1.19	0.50				1	0.91 - 1.10	0.97			
Previous TB treatment: Yes	1.38	0.93 - 2.05	0.11	1.49	1.02 - 2.16	0.037	1.16	0.88 - 1.52	0.28			
Focal neurological signs: Yes	1.77	1.20 - 2.60	0.004	1.62	1.12 - 2.33	0.010	1.14	0.86 - 1.51	0.37			
Temperature[Celsius]	0.95	0.78 - 1.16	0.64				1.09	0.96 - 1.24	0.20			
Convulsion: Yes	0.89	0.37 - 2.13	0.79				1.01	0.61 - 1.66	0.98			
Dexamethasone: No	1.71	1.19 - 2.5	0.004	2.0	1.45 - 2.70	<0.001	1.04	0.63 - 1.72	0.88			
Plasma sodium [per +10 mmol/l] £	0.99	0.83 - 1.19	0.93				-	-	-	-	-	-
- 135 vs. 125	0.99	0.83 - 1.19					1.08	0.89 - 1.30		1.08	0.90 - 1.30	
- 115 vs. 125	1.01	0.85 - 1.20					1.67	1.33 - 2.09		1.66	1.33 - 2.06	
CSF lymphocyte count [cells/ mm3] ‡	0.88	0.81 - 0.94	<0.001	0.86	0.81 - 0.92	<0.001	0.95	0.89 - 1.01	0.09			
CSF protein [g/l] ‡	0.97	0.85 - 1.10	0.60				0.96	0.86 - 1.07	0.47			
CSF glucose [mmol/l]	0.99	0.83 - 1.19	0.95				1.05	0.95 - 1.15	0.32			
Ratio of CSF glucose and blood glucose ‡	1.07	0.84 - 1.36	0.57				0.93	0.79 - 1.10	0.41			
Resistance ♯												
- No isoniazid or rifampin resistance	1			1			1			1		
- Isoniazid monoresistance	0.87	0.45 - 1.71	0.69	0.9	0.46 - 1.73	0.74	1.56	1.14 - 2.14	0.005	1.57	1.15 - 2.13	0.004
- Rifampin monoresistance /MDR	5.78	2.32 - 14.42	<0.001	5.8	2.39 - 14.04	<0.001	5.38	3.31 - 8.75	<0.001	5.56	3.48 - 8.89	<0.001
- Unknown resistance	1.37	0.93 - 2.03	0.12	1.47	1.01 - 2.14	0.046	1.66	1.28 - 2.16	<0.001	1.74	1.37 - 2.22	<0.001
Chest x-ray miliary TB: Yes	0.85	0.58 - 1.24	0.39				1.01	0.74 - 1.37	0.97			
Peripheral blood CD4 [cells/mm3] ‡							0.91	0.84 - 0.98	0.012	0.88	0.82 - 0.94	<0.001
On ART at enrolment [Yes]							0.77	0.53 - 1.12	0.17			
Cohort												
- Intensified trial	1						1			1		
- Dexamethasone trial	1.37	0.88 - 2.15	0.17				1.56	0.97 - 2.52	0.07	1.54	1.12 - 2.11	0.008
- Fluoroquinolone trial	1.02	0.46 - 2.28	0.95				-	-	-	-	-	-

Continued on next page

Table A3.3 – *Continued from previous page*

Variable	HIV-uninfected TBM population						HIV-infected TBM population					
	Full model			Final model (Model selected by the lasso)			Full model			Final model (Model selected by backwards stepwise selection method)		
	HR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
- TBM HIV cohort	-						2.69	1.70 - 4.23	<0.001	2.84	1.96 - 4.13	<0.001
- ART timing trial	-						1.79	1.28 - 2.49	<0.001	1.84	1.44 - 2.36	<0.001

HR=hazard ratio, CI=confidence interval. 95% CI and p-value for final models do not take into account the uncertainty of model selection.
‡: HR per 2-fold increase.
£: In HIV-infected subjects, the effect of sodium on mortality was significantly non-linear and modelled with a restricted cubic spline function with 2 degrees of freedom. To simplify interpretation of the corresponding regression coefficients, only HRs for two derived sodium contrasts from that model are given. P-values for sodium In the HIV-infected population are based on overall Wald-tests of the restricted cubic spline function
♯: Isoniazid monoresistance is defined as resistance to isoniazid but not to rifampin. Multidrug resistance (MDR) is defined as resistance to at least isoniazid and rifampin. Unknown resistance is defined as drug-susceptibility test results not available. In all categories, resistance to other drugs may be present.
§ MRC Grade I (GCS 15; no focal neurological signs); MRC Grade II (GCS 11-14, or 15 with focal neurological signs); MRC Grade III (GCS $\leq$ 10)).

Chapter 4

Dynamic prediction of death in patients with tuberculous meningitis using time-updated Glasgow coma score and plasma sodium measurements

Summary: I present the development and validation of dynamic prediction models for TBM using longitudinal information of two biomarkers: Glasgow coma score and plasma sodium. The major part of this chapter has been submitted to the Clinical Infectious Diseases Journal and is currently under review.

4.1 Introduction

The pre-treatment predictors of poor outcome from TBM are well-studied. In the previous chapter, I have presented the prognostic models for 9 month mortality. The models were based on several pre-treatment clinical variables and better predicted subsequent death than the MRC grade alone.

TBM requires 9-12 months of anti-tuberculosis drug treatment, during which complications from both the disease and the drugs can occur (Wilkinson et al., 2017). These complications change the likelihood of a poor outcome from TBM, but are not captured by current prognostic models based on pre-treatment variables alone. Currently, there are no published models that take into account changing clinical variables after the start of TBM treatment. The aim of this study was to develop and validate dynamic prediction models that incorporate baseline patient characteristics, as well as follow-up measurements of GCS and plasma sodium. The models were constructed using a large sample of 1048 Vietnamese adults with TBM. During the first 30 days of anti-tuberculosis treatment, I used updated information to predict mortality within the first 120 days. This time range of 30 days was selected based on the clinical experience that the first 30 days are critical to outcome. Most patients are hospitalised during this period and repeated measures of GCS and plasma sodium are usually available. I compare the predictive performance of these dynamic models to the baseline models presented in Chapter 3 (Thao et al., 2018b).

4.2 Methods

4.2.1 Study population

4.2.1.1 Study participants

I used data from three RCTs (Heemskerk et al., 2016; Török et al., 2011b; Mai et al., 2018) and one observational study (Torok et al., 2008) on TBM which provided longitudinal information of GCS and plasma sodium during follow-up. The studies were conducted between 2004 and 2016. A summary of the studies and their inclusion criteria, the laboratory investigations and anti-tuberculosis and adjunctive treatment are described in detail in Chapter 2. Study participants were excluded if no information on GCS or plasma sodium was available, if any of the baseline risk factors were missing, if an alternative diagnosis to TBM was confirmed, or if standard first-line anti-tuberculosis drug treatment was not given.

4.2.1.2 Primary outcome

The primary outcome for this study was defined as overall mortality during 120 days from the start of anti-tuberculosis treatment. Follow-up was restricted to 120 days

because deaths occurring within this period are much more likely to be attributable to TBM rather than an alternative cause.

4.2.1.3 Predictors

As in the previous chapter, I constructed separate prediction models for HIV-uninfected and HIV-infected patients. The new models included both baseline patient characteristics and the longitudinal measurements of GCS and plasma sodium. I chose GCS and plasma sodium because both have been shown to be good pre-treatment predictors of death from TBM and both are monitored regularly in almost all settings (see section 2.1.2). In the first 30 days of follow-up, most patients had daily GCS measurements in one trial (Heemskerk et al., 2016) and weekly in the other studies. Plasma sodium was mostly measured daily for patients in one trial (Mai et al., 2018) and weekly in the other studies. I did not include other biomarkers such as CSF protein or ratio of CSF to blood glucose because on average there were only two measurements per patient in the first 30 days, precluding the development of a robust model.

I included the baseline characteristics that were identified as predictors of death within 9 months in the models presented in Chapter 3 . Specifically, in the HIV-uninfected population, these were age, history of previous TB treatment, MRC grade, focal neurological signs, and CSF lymphocyte count. In the HIV-infected population, these were weight, MRC grade, CSF lymphocyte count, study cohort, and peripheral blood CD4 cell count. Baseline laboratory values were defined as values recorded closest to enrolment (up to $\pm$7 days from enrolment) (Thao et al., 2018b). Compared to the previous analysis, I excluded patients enrolled in two studies (Thwaites et al., 2004a, 2011) because their GCS and plasma sodium after the start of anti-tuberculosis treatment were not available. However, I retained the study cohort variable in the model for HIV-infected patients to account for improvements in treatment and patient supportive care over the last 15 years in Vietnam. The most significant of these was the nation-wide scale-up of free HIV antiretroviral therapy since 2005 (Nguyen et al., 2013). I included sodium in the model for both populations, and used it as a time-updated instead of as a baseline variable.

4.2.2 Statistical analysis

4.2.2.1 Transformation of covariates

Prior to the analysis, I log2-transformed CSF Lymphocyte counts, and CD4 cell count to make their distribution less skewed. I added 1 unit to CD4 cell count before applying the transformation to avoid log transforming a value of zero. In addition, I subtracted 100 units and log2-transformed plasma sodium in the HIV-uninfected subgroup to make the variable approximately normal (a required assumption for the mixed effects model). For sodium in HIV-infected patients, I only centered the value by subtracting 100 units, as the distribution of sodium was relatively normal without the log transformation.

4.2.2.2 Statistical models for prediction

I used a mixed model landmarking approach (Rizopoulos et al., 2017), also called a two-stage partly conditional method (Maziarz et al., 2017). Specifically, for a given landmark time t_{LM}, I selected subjects who were still alive and observed at t_{LM}. In this data set, denoted as $R(t_{LM})$, events that occurred after $t_{LM} + 120$ days were censored. An alternative approach is to censor events after 120 days from baseline. We chose a fixed time window and include events beyond day 120 (up today 150 maximum) from baseline, which are also likely to be due to TBM.

I treated the two biomarkers differently when incorporating their values in the model. With plasma sodium, I first fitted a mixed-effects model for the longitudinal profile of the patients in $R(t_{LM})$:

$$y_i(t) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t + \epsilon(t) \tag{4.1}$$

$$\boldsymbol{b_i} \sim N(0, \boldsymbol{D}), \epsilon(t) \sim N(0, \sigma^2)$$

where $y_i(t)$ denotes the observed value of plasma sodium for patient i at time t , $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is the vector of fixed effects, $\boldsymbol{b_i} = (b_{i0}, b_{i1})^T$ is the vector of random effects for patient i , and $\epsilon(t)$ is the error term which is assumed to be white noise. The mixed model for each landmark time takes all of the longitudinal measurements from baseline to the corresponding landmark time. For example, if we want to predict

mortality at day 120 using data up to day 30, then all of the longitudinal measurement of sodium up to day 30 will be used in the mixed model.

I derived the fitted value for sodium at t_{LM} for each patient in $R(t_{LM})$ using the best linear unbiased predictor (BLUP) estimator (Maziarz et al., 2017), denoted as $\hat{y}(t_{LM})$. This approach uses all earlier recorded information and eliminates the error term which is assumed to represent measurement error and short-term fluctuation that is irrelevant for prediction (Rizopoulos et al., 2017). It is expected to provide a good approximation to the true unobserved sodium value. With GCS, I did not use a longitudinal model because of two reasons. First, unlike plasma sodium, GCS is often measured with high precision. Second, GCS can have a highly nonlinear pattern, which is difficult to specify in a model. Instead, I used the last available measured value up till 10 days prior to the landmark time. An individual was removed from $R(t_{LM})$ if his last GCS was more than 10 days earlier.

Finally, I combined the 31 data sets from day 0 to day 30 to construct a super-model for prediction of death (van Houwelingen and Putter, 2011). The stacked data was analysed using a Cox proportional hazards model, in which the landmark time was used as a stratification factor. I did not check for proportional (PH) assumption because recent simulation (Ferrer et al., 2017) showed that the violation of PH assumption in landmark and joint model has limited impact on predictive performance. I included the baseline predictors, the most recent GCS value, and the fitted value of plasma sodium at t_{LM}. Based on the results from Chapter 3, the effect of plasma sodium on survival may be nonlinear. Thus, I modelled it using a restricted cubic spline function with two degrees of freedom. I included a restricted interaction (Harrell, 2015) between sodium and landmark time to allow the association between sodium and survival to change over landmark time. The restriction was imposed to make sure that the formed interaction is not "doubly nonlinear". In other words, the restricted interaction between time and sodium is nonlinear in each covariate but not in both at the same time. The final model can be written as:

$$\begin{aligned} h(u) = \quad & h_{0t_{LM}}(u) \exp\{\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{X} + \alpha_1 GCS(t_{LM}) + \alpha_2^1 \hat{y}(t_{LM}) + \alpha_2^2 \hat{y}\prime(t_{LM}) + \\ & \alpha_3^1 \hat{y}(t_{LM}) t_{LM} + \alpha_3^2 \hat{y}\prime(t_{LM}) t_{LM} + \alpha_3^3 \hat{y}(t_{LM}) t\prime_{LM}\} \end{aligned} \tag{4.2}$$

where $\boldsymbol{X}$ are the baseline predictors with the corresponding vector of coefficients $\boldsymbol{\gamma}$; $\hat{y}_{LM}$ and $\hat{y}\prime_{LM}$ denote the first (linear) and second (non-linear) spline term for sodium.

t_{LM} and $t\prime_{LM}$ are the first (linear) and second (non-linear) spline term for landmark time.

4.2.2.3 Extra analyses and alternative modelling approach

I explored a number of extra analyses as specified below. They are all extensions of the more parsimonious model described in the previous section.

- In the mixed-effects model that was used to predict the sodium concentration at the landmark time points (equation 4.1), I fitted a more flexible model by using a spline function for time.
- I allowed the risk of dying to depend on the slope of the longitudinal profiles for both GCS and sodium. In this model, two patients who have the same marker values at a specific landmark time can have a different predicted mortality if one of them has an increasing profile and the other one has a decreasing profile. The slope for GCS was defined based on the last two adjacent measurements, or received a value of zero if there was only one available measurement. The slope for sodium was derived from the mixed-effects model.
- I modelled GCS using a restricted cubic spline function with two degrees of freedom, and added a restricted interaction between GCS and landmark time to allow the effect of GCS to change over time.

I also used a joint modelling approach that combines a multivariate mixed-effects model for the two longitudinal variables (GCS and sodium) and a Cox model for mortality (Rizopoulos, 2011; Rizopoulos and Ghosh, 2011). The general introduction about joint models was presented in Chapter 1. I considered two association structures: i) The "current value" association structure: the hazard of death depends linearly on the "true" values of the markers at the same time point, and ii) the current value and the slope: the hazard of death depends on the current value and the slope of the longitudinal markers. The considered joint models are formulated as follows:

$$h_i(t|X_i) = h_0(t)\exp\{\boldsymbol{X_i^T}\boldsymbol{\gamma} + \sum_{k=1}^{2}\alpha_{0k}f_{ik}(t)\}$$

$$h_i(t|X_i) = h_0(t)\exp\{\boldsymbol{X_i^T}\boldsymbol{\gamma} + \sum_{k=1}^{2}\alpha_{0k}f_{ik}(t) + \alpha_{1k}\frac{df_{ik}(t)}{dt}\}$$

where:

- $f_{ik}(.)$ describes the true and unobserved longitudinal trajectory for each of the markers with corresponding fixed effects β_k and random effects b_{ik},
- $\boldsymbol{X_i}$ denotes the vector of baseline covariates and $\boldsymbol{\gamma}$ is the vector of the corresponding coefficients,
- α_{0k}, α_{1k} denote the association parameter that measures the effect of the k^{th} longitudinal marker to the survival outcome.

The longitudinal models ($f_{ik}(.)$) assumed a random intercept and random slope with time (linear form) included as the only covariate as the landmark model for both sodium and GCS. The Cox model included all baseline predictors (X_i) as in the landmark model (equation 4.2). Estimation was performed under a Bayesian framework as implemented in the package **JMbayes** (Rizopoulos, 2016).

4.2.2.4 Model performance and validation

AUC and Brier score

I assessed the discrimination and calibration of the dynamic models at days 10, 20 and 30 using the cumulative/dynamic AUC, and the Brier score. The event was death until day 60, 90 and 120 after diagnosis. In other words, I used data of GCS and Sodium up till day 10, 20 or 30 from baseline to predict mortality at day 60, 90, and 120 from baseline. I also calculated the AUC and Brier score of the baseline model (Thao et al., 2018b) at the same prediction times using the corresponding landmark data sets at that day. Both measures are based on the conditional survival probability and accounted for right censored event times via IPW as mentioned in Chapter 1 (Rizopoulos et al., 2017). The conditional survival probability is defined as the probability of being alive up to time $u > t$ conditioning on being at risk at landmark time t. These estimates have been implemented in the `aucJM()` and `prederrJM()` functions in the **R** package **JMbayes**(Rizopoulos, 2016).

For the model performances of the landmarking approach, I corrected for overfitting via bootstrap internal validation (Harrell (2015), page 114) which was described in Chapter 2. Of note, in each bootstrap sample, I selected by patient, i.e. if a patient was selected, his whole longitudinal profile was included in the data set. For each bootstrap data set, I repeated the model building step by first creating a stacked

landmark data set, in which I used the BLUP estimates for sodium from a mixed effects model and then fitting the Cox model as in 4.1.

Calibration risk groups plot
I generated calibration plots (Royston and Altman, 2013) at landmark times 10, 20 and 30. The predicted risks were corrected for over-optimism using the same validation procedure as with AUC and Brier score.

4.2.2.5 Statistical software

All analyses were conducted using **R** version 3.5.0 (R Core Team, 2018), and the following packages: **survival** (Therneau, 2015), **JMbayes** (Rizopoulos, 2016)), **rms** (Harrell Jr, 2016), **tidyverse** (Wickham, 2017) , **survminer** (Alboukadel Kassambara and Marcin Kosinski, 2018).

4.3 Results

4.3.1 Baseline characteristics of study participants

In total, there were 1248 patients in the four studies, of which 1068 were included in the analysis. The reasons for excluding the other 180 patients were: confirmed other diagnosis (n = 16), first-line anti-tuberculosis drugs not received (n = 3), missing information on either of the two markers (n = 7), missing values for at least one of the baseline patient characteristics (n = 154). Baseline characteristics of the included patients are given in Table 4.1. Amongst 1068 patients, 550 (51%) were HIV-uninfected, and 518 (49%) were HIV-infected. The median age was 34 years (25% and 75% percentile: 27-44). Out of 1068 patients, 395 (37%) were MRC grade I, 459 (43%) were MRC grade II, and 214 (20%) were MRC grade III. The number of definite TBM (microbiologically confirmed TBM) cases was 636/1068 (60%). In HIV-infected subjects, the median peripheral blood CD4 count was 40 cells/μL (the 25% and 75% percentile: 15 - 101). Compared to the HIV-uninfected patients, HIV-infected patients were younger (median age 31 years vs. 41 years), were admitted to the hospital in more severe condition (MRC Grade III 24.9% vs. 15.5%), and more likely to be classified as having definite TBM (70% vs. 50%).

Figure 4.1 displays the Kaplan-Meier estimate of the overall survival by HIV status. During the 120 days of follow-up, the number of deaths was 76/550 (14%) in the HIV-uninfected group, and 202/518 (39%) in the HIV-infected group.

Table 4.1: *Baseline characteristics of patients with tuberculous meningitis included in the pooled database, overall and by human immunodeficiency virus status*

Characteristic	All TBM patients (N=1068)		HIV-uninfected TBM (N=550)		HIV-infected TBM (N=518)	
	n[a]	Summary statistic	n[a]	Summary statistic	n[a]	Summary statistic
Cohort ∗	1068		550		518	
- Intensified treatment trial		689 (64.5%)		431 (78.4%)		258 (49.8%)
- Aspirin trial		119 (11.1%)		119 (21.6%)		0(0%)
- TBM HIV cohort		44 (4.1%)		0 (0%)		44 (8.5%)
- ART timing trial		216 (20.2%)		0 (0%)		216 (41.7%)
Age (yrs), median (IQR)	1067	34(27,44)	550	41(29,53)	517	31(27,36)
Weight (kg), median (IQR)	1068	48(43,54)	550	50(45,55)	518	46(41,51)
Previous tuberculosis treatment	1062	162 (15.3%)	550	50 (9.1%)	512	112 (21.9%)
MRC grade [b]	1068		550		518	
- Grade I		395 (37%)		208 (37.8%)		187 (36.1%)
- Grade II		459 (43%)		257 (46.7%)		202 (39%)
- Grade III		214 (20%)		85 (15.5%)		129 (24.9%)
Focal neurological signs present	1056	528 (50%)	550	321 (58.4%)	506	207 (40.9%)
CSF lymphocyte count (cells/μL), median (IQR)	1068	94.5(35,208.7)	550	104.8(42,213.1)	518	80.7(25.2,204.8)
Microbiologically confirmed/definite TBM	1068	636 (59.6%)	550	274 (49.8%)	518	362 (69.9%)

Continued on next page

Table 4.1 – *Continued from previous page*

Characteristic	**All TBM patients (N=1068)**		**HIV-uninfected TBM (N=550)**		**HIV-infected TBM (N=518)**	
	n[a]	**Summary statistic**	**n**[a]	**Summary statistic**	**n**[a]	**Summary statistic**
Peripheral blood CD4 count (cells/μL), median (IQR)	-	-	-	-	518	40(15,101)

Abbreviations: CSF = cerebrospinal fluid; HIV = human immunodeficiency virus; IQR = interquartile range (25% and 75% percentile); MRC = Medical Research Council; TBM = tuberculous meningitis.

[a]Number of subjects with non-missing data for the respective characteristic. Note: the number of individuals with measured values can be lower than the maximum (550 resp. 518), but these risk factors were not part of the model for that HIV subpopulation.

[b]Grade I: Glasgow Coma score (GCS) 15 with no focal neurological signs; Grade II: GCS 1114, or 15 with focal neurological signs; Grade III: GCS $\leq$ 10.

$*$ Included studies are described in Chapter 2.

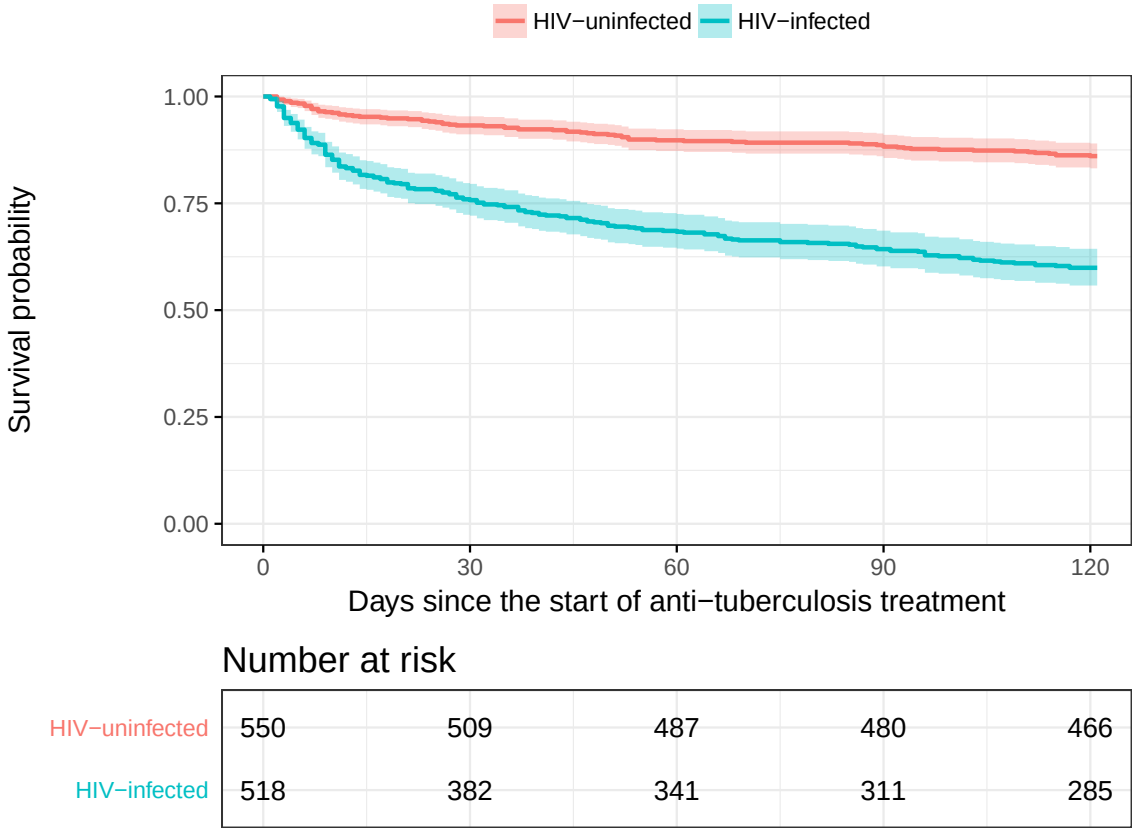


Figure 4.1: *Kaplan Meier survival curves by human immunodeficiency virus (HIV) status*

4.3.2 Longitudinal profile of GCS and plasma sodium

During the first 30 days of treatment, HIV-uninfected patients had a mean of 21 (median (25% and 75% percentile): 27 (7-31)) GCS assessments and HIV-infected patients had a mean of 14 (median (25% and 75% percentile): 5 (5-29)) assessments. The mean numbers of plasma sodium measurements was 6 (median (25% and 75% percentile): 5 (4-7)) and 4 (median (25% and 75% percentile): 5(3-5)) in HIV-uninfected and HIV-infected patients, respectively. Figure 4.2 depicts the individual trajectories of the two markers as well as their median value and quartiles during the first 30 days. All survivors, independent of HIV status, tended to have higher GCS than those who died. Plasma sodium values were also higher in HIV-infected survivors, with a less clear relationship between sodium and survival in HIV-uninfected patients.

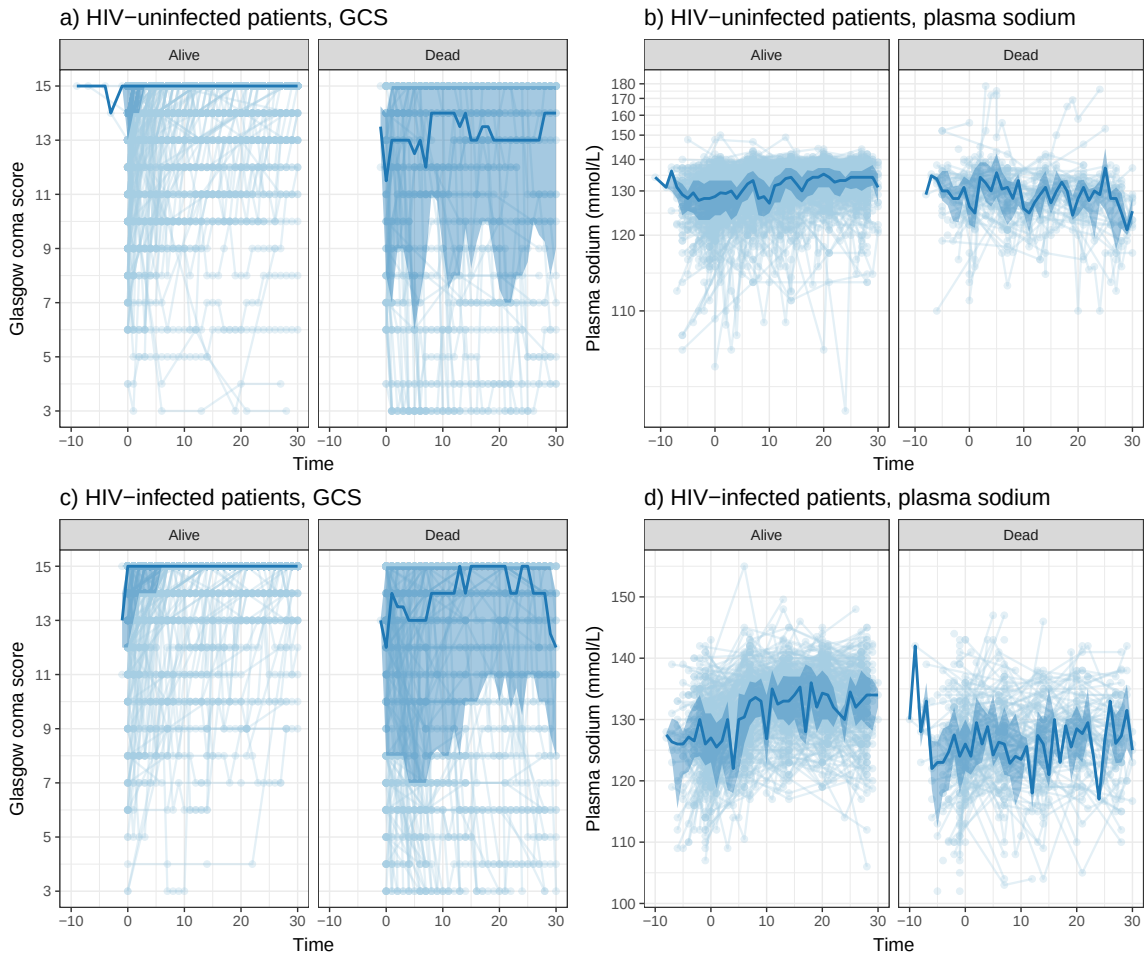


Figure 4.2: *Subject specific profiles for Glasgow coma score (GCS) and plasma sodium over the first 30 days, stratified by observed survival status within 120 days. The dark blue line represents the median value per day, the blue area represents the interquartile range. Plasma sodium of HIV-uninfected patients is plotted as log2(plasma sodium - 100) with the original value as label.*

4.3.3 Parameter estimates

Table 4.2 gives parameter estimates of the dynamic landmark supermodels. Regarding GCS, a higher value indicated lower mortality with a hazard ratio of 0.76 (95% confidence interval (CI): 0.71-0.81) and 0.85 (95% CI: 0.81- 0.91) in HIV-uninfected and HIV-infected patients respectively. Figure 4.3 displays the hazard ratio for sodium at day 0, 10 and 30. Sodium concentrations of 135 mmol/L were taken as the reference, i.e. hazard ratio is 1 for sodium 135 mmol/L. As can be seen, the association between plasma sodium and mortality differed by HIV status. In HIV-uninfected patients, it was heterogeneous over time. Higher sodium concentrations were associated with worse prognosis from baseline until day 10 of treatment. However, by day 30, higher

sodium concentrations were associated with increased survival. In HIV-infected patients, higher sodium concentrations uniformly associated with better prognosis across all prediction time points. Sensitivity analysis showed that the association is sensitive to the number of knots in the spline function for sodium, mostly in HIV-uninfected patients (Figure A4.1, A4.2). As the number of knots increases, high sodium values showed an erratic pattern with (again) a very large confidence interval for the hazard ratio.

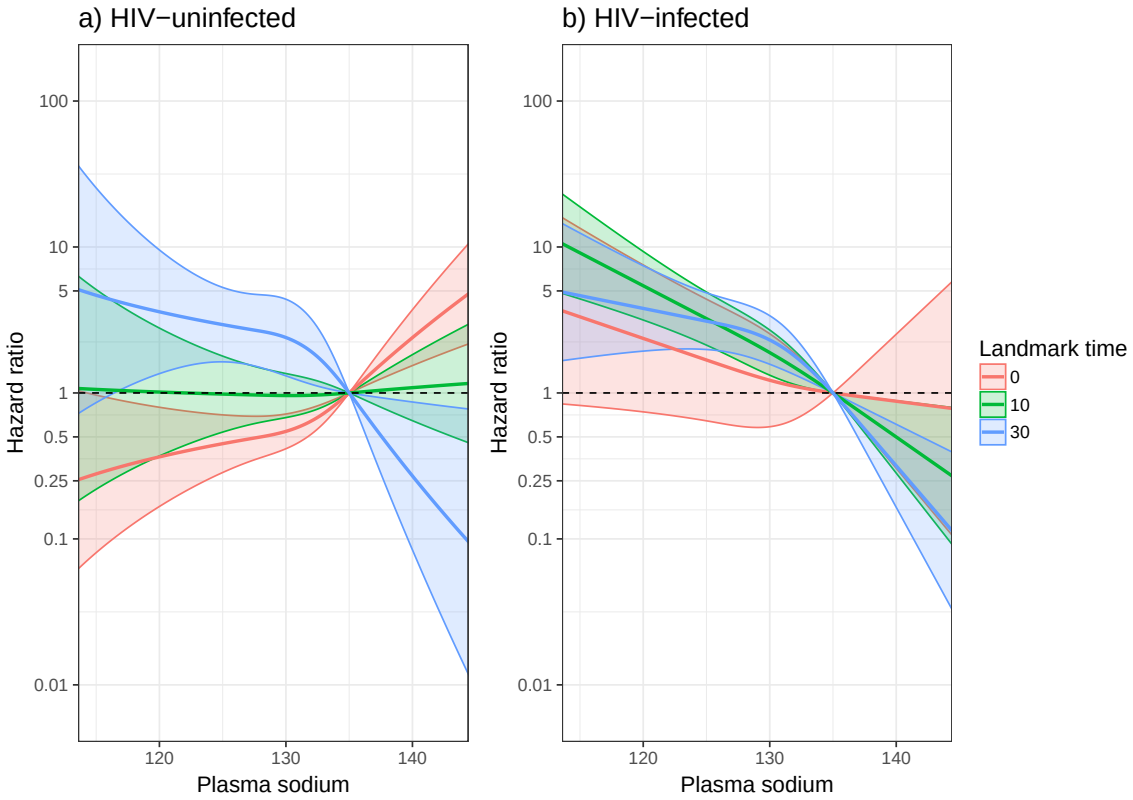


Figure 4.3: *Hazard ratio for plasma sodium at three landmark times: day 0, day 10 and day 30 after the start of TB treatment. A sodium value of 135 mmol/L was taken as the reference. The shaded areas represent the 95% confidence intervals.*

Compared with the baseline model (Thao et al., 2018b), the strength of the association between MRC grade and mortality was attenuated with the inclusion of time-updated GCS. This is not surprising as GCS is strongly correlated with MRC grade (GCS is the main component of MRC grade). The strength of the association for the other baseline covariates was similar between the dynamic models and the baseline models, except for history of previous TB treatment in the model for HIV-uninfected patients. The hazard ratio for this predictor was 0.57 indicating a beneficial effect which contradicted with that in the baseline model (HR = 1.57, CI

1.09 -2.26). However, it was not statistically significant at 0.05 level (p-value = 0.27, CI 0.21-1.54).

Table 4.2: *Parameter estimates (with 95% confidence intervals) from landmark supermodel by HIV status*

Variable	**HIV-uninfected TBM**		**HIV-infected TBM**	
	HR	**95% CI**	**HR**	**95% CI**
Age (per +10y)	1.51	(1.34; 1.69)	-	-
Weight (per +10kg)	-	-	0.59	(0.47; 0.75)
MRC grade [a]				
- Grade I	1	-	1	-
- Grade II	1.02	(0.37; 2.83)	1.48	(0.99; 2.21)
- Grade III	1.48	(0.43; 5.05)	1.37	(0.77; 2.45)
Previous tuberculosis treatment	0.57	(0.21; 1.54)	-	-
Focal neurological signs present	1.45	(0.60; 3.54)	-	-
CSF lymphocyte count [b], cells/μL	0.78	(0.68; 0.89)	0.99	(0.91; 1.08)
Cohort	-	-		
- Intensified trial	-	-	1	
- TBM HIV cohort	-	-	2.55	(1.62; 4.02)
- ART timing trial	-	-	1.55	(1.10; 2.18)
Peripheral blood CD4 count [b], cells/μL	-	-	0.92	(0.83; 1.01)
Time-updated Glasgow coma score	0.76	(0.71; 0.81)	0.85	(0.81; 0.91)
Time-updated plasma sodium, mmol/L [c]	-	-	-	-
- Plasma sodium 125 vs 135 mmol/L	-	-	-	-
- LM time = day 0 (baseline)	0.45	(0.29; 0.71)	1.67	(0.64; 4.36)
- LM time = day 10	0.98	(0.54; 1.77)	3.21	(2.11; 4.87)
- LM time = day 30	2.93	(1.63; 5.25)	3.10	(1.98; 4.83)
- Plasma sodium 140 vs 135 mmol/L	-	-	-	-
- LM time = day 0 (baseline)	2.37	(1.53; 3.67)	0.88	(0.31; 2.52)
- LM time = day 10	1.09	(0.65; 1.82)	0.50	(0.28; 0.88)
- LM time = day 30	0.27	(0.09; 0.87)	0.32	(0.16; 0.61)

Abbreviations: CI = confidence interval; CSF = cerebrospinal fluid; HIV = human immunodeficiency virus; HR = hazard ratio; MRC = Medical Research Council; TBM = tuberculous meningitis, LM = landmark.

[a] Grade I: Glasgow Coma score (GCS) 15 with no focal neurological signs; Grade II: GCS 1114, or 15 with focal neurological signs; Grade III: GCS $\leq$ 10

Continued on next page

Table 4.2 – *Continued from previous page*

[b] HR per 2-fold increase.
[c] The association of sodium levels and mortality was allowed to vary in a flexible manner. The p-value for the interaction between sodium and time was < 0.001 and 0.18 respectively for the HIV-uninfected and HIV-infected subgroups. To simplify interpretation, I give HRs for two derived sodium contrasts at three different landmark time points.

4.3.4 Predictive performance

The dynamic prognostic model showed good discrimination between survivors and deaths, with corrected AUC ranging from 0.82-0.92 in HIV-uninfected, and from 0.81-0.85 in HIV-uninfected patients across the chosen landmark and prediction times (Table 4.3). The dynamic models clearly improved over the published baseline models (Thao et al., 2018b), which had AUCs from 0.75-0.81 in HIV-uninfected and from 0.71-0.76 in HIV-infected subjects (Table 4.3). In terms of calibration, Figure 4.4 shows good agreement between the predicted risks and the observed risks for all three chosen landmark time points. Overall performance, as measured by the Brier score, was also improved in the dynamic models compared to the baseline models (Table 4.3).

All the considered alternative analyses did not improve the predictive performance of the models. AUC values of the alternative models and the presented models were analogous to the second decimal. The presented model was chosen as the final model because of its parsimony. Regarding the joint model, the two association structures gave similar performance so only the result of the first one is reported (Figure 4.5). The joint models gave an apparent predictive performance that was worse than that of the final models. As the joint model required a substantially large amount of time to fit, I did not perform the internal validation for this model.

Table 4.3: *Optimism-corrected AUC and Brier score of the landmark model and the baseline model by human immunodeficiency virus (HIV) status. The prediction was carried out at day 10, 20 and 30 for mortality until day 60, 90 and 120 after diagnosis.*

Prediction time	Prediction horizon∗	HIV-uninfected TBM population				HIV-infected TBM population			
		AUC		Brier score		AUC		Brier score	
		LM	Baseline	LM	Baseline	LM	Baseline	LM	Baseline
10	60	0.819	0.769	0.051	0.061	0.829	0.762	0.109	0.146
	90	0.834	0.787	0.061	0.071	0.825	0.753	0.130	0.167
	120	0.839	0.810	0.073	0.083	0.814	0.745	0.150	0.184
20	60	0.866	0.775	0.042	0.053	0.832	0.737	0.090	0.136
	90	0.869	0.795	0.051	0.064	0.842	0.728	0.110	0.162
	120	0.866	0.818	0.062	0.076	0.825	0.721	0.136	0.182
30	60	0.919	0.748	0.027	0.043	0.843	0.725	0.066	0.126
	90	0.890	0.780	0.036	0.055	0.845	0.716	0.091	0.155
	120	0.890	0.811	0.046	0.068	0.816	0.709	0.122	0.179

Abbreviations: Baseline = Baseline model; LM = Landmark model

∗ Prediction horizon refers to days after the initiation of anti-tuberculous treatment

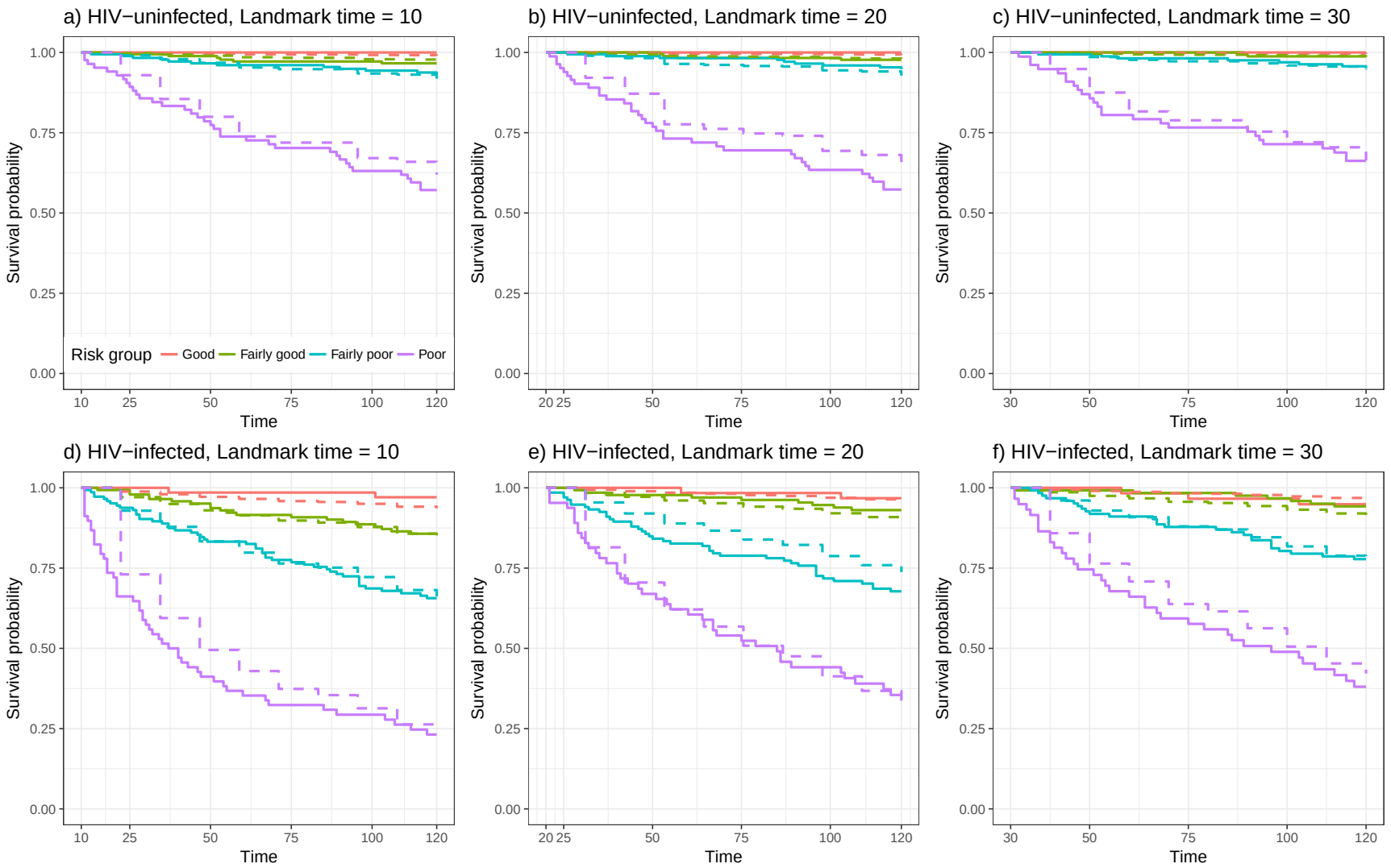


Figure 4.4: *Comparison between average predicted survival (dashed line) and Kaplan-Meier estimates (solid line) at day 0, day 10 and day 30. Groups are defined using cut-off points at the 16th, 50th, and 84th percentiles of the prognostic index as generated by the final models (defining "good", "fairly good", "fairly poor", and "poor" prognostic subgroups).*

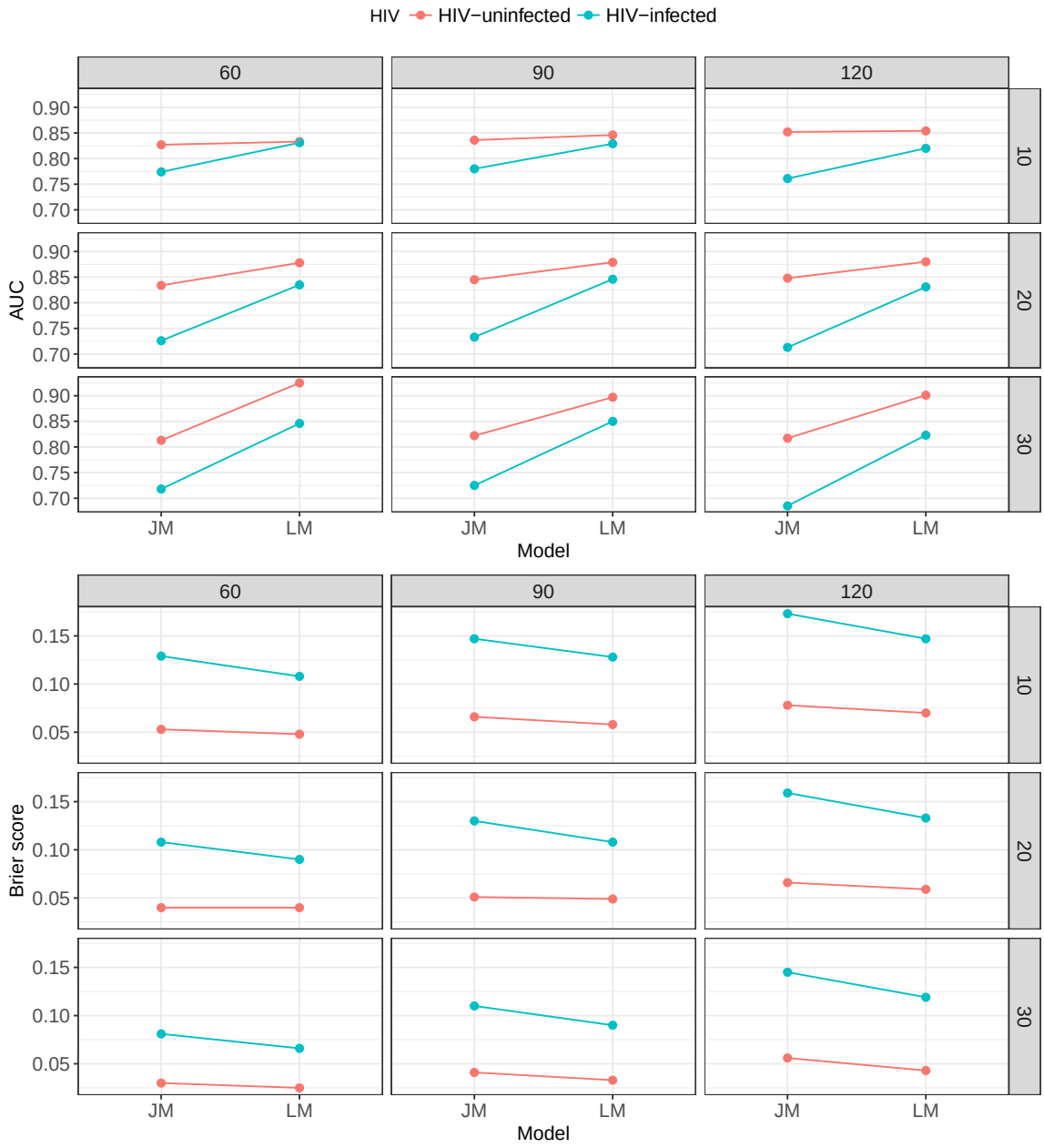


Figure 4.5: *Apparent AUC and Brier score of the landmark model (LM) and the joint model (JM) by human immunodeficiency virus (HIV) status. The prediction was carried out at day 10, 20 and 30 for mortality until day 60, 90 and 120 after diagnosis*

4.3.5 Model display and illustration

I implemented the models in a user-friendly web-based calculator, which is available at: `https://thaole.shinyapps.io/DynamicTBMApp/`. The screen shot of the app is presented in Figure A4.3. The web calculator allows the user to make predictions of mortality within 120 days after the initiation of TB treatment from any of the first 30 days. As in the baseline models web calculator, I chose the intensified trial (Heemskerk et al., 2016) as cohort variable for HIV-infected patients because it is the most recent cohort.

In Figure 4.6, I illustrate the use of the model to dynamically predict mortality until day 120 for a 39 year old HIV-uninfected patient without previous TB, having MRC grade III, no focal neurological signs at enrolment and with CSF lymphocyte count at baseline of 73.5 cells/μL (patient A). A similar illustration for an HIV-infected patient (patient B) is in Figure 4.7. Baseline characteristics of the two patients are documented in Table 4.4.

Table 4.4: *Baseline characteristic of patients A and B*

Variable	**Patient A**	**Patient B**
HIV infection	No	Yes
Age (years)	39	32
Weight (kg)	53	45
History of previous tuberculosis treatment	No	Yes
MRC grade [a]	III	III
Focal neurological signs present	No	Yes
Plasma sodium (mmol/L)	124	127
CSF lymphocyte count (cells/μL)	73.5	134.2
Plasma CD4 count, (cells/μL)	-	42
Definite TBM	No	Yes

Abbreviations: CSF = cerebrospinal fluid; HIV = human immunodeficiency virus; MRC = Medical Research Council; TBM = tuberculous meningitis.
[a] MRC grade I is defined as Glasgow Coma Score (GCS) 15 with no focal neurological signs; grade II, GCS score 1114, or 15 with focal neurological signs; grade III, GCS score $\leq$ 10.

On the left hand side of each panel in Figure 4.6 (and Figure 4.7), I display the longitudinal trajectory of plasma sodium (upper panel), and GCS (lower panel). The blue line in the upper panel represents the BLUP fitted value for sodium. On the

right hand side, I display the prediction of survival until day 120. Of note, at each evaluation time, the predicted survival probability always starts at 100%. As can be seen from Figure 4.6, at baseline, patient A had a sodium value of 124 mmol/L and a GCS value of 8. The predicted 120-day survival was about 75%. At day 9, I observed an increasing trend in both plasma sodium and GCS value. The predicted 120-day survival was updated to about 90%, indicating an improvement in patient conditions. However, at day 20, as the two markers decreased, the predicted survival dropped to 50%. At day 30, the patient showed a steeper decrease in plasma sodium and a further drop in GCS value from 8 to 7. The updated predicted survival was about 25%, indicating a poor prognosis. In fact, patient A died at day 37. Of note, the risk predictions can vary substantially over the landmark days, indicating the importance of updating prediction using the most recent information.

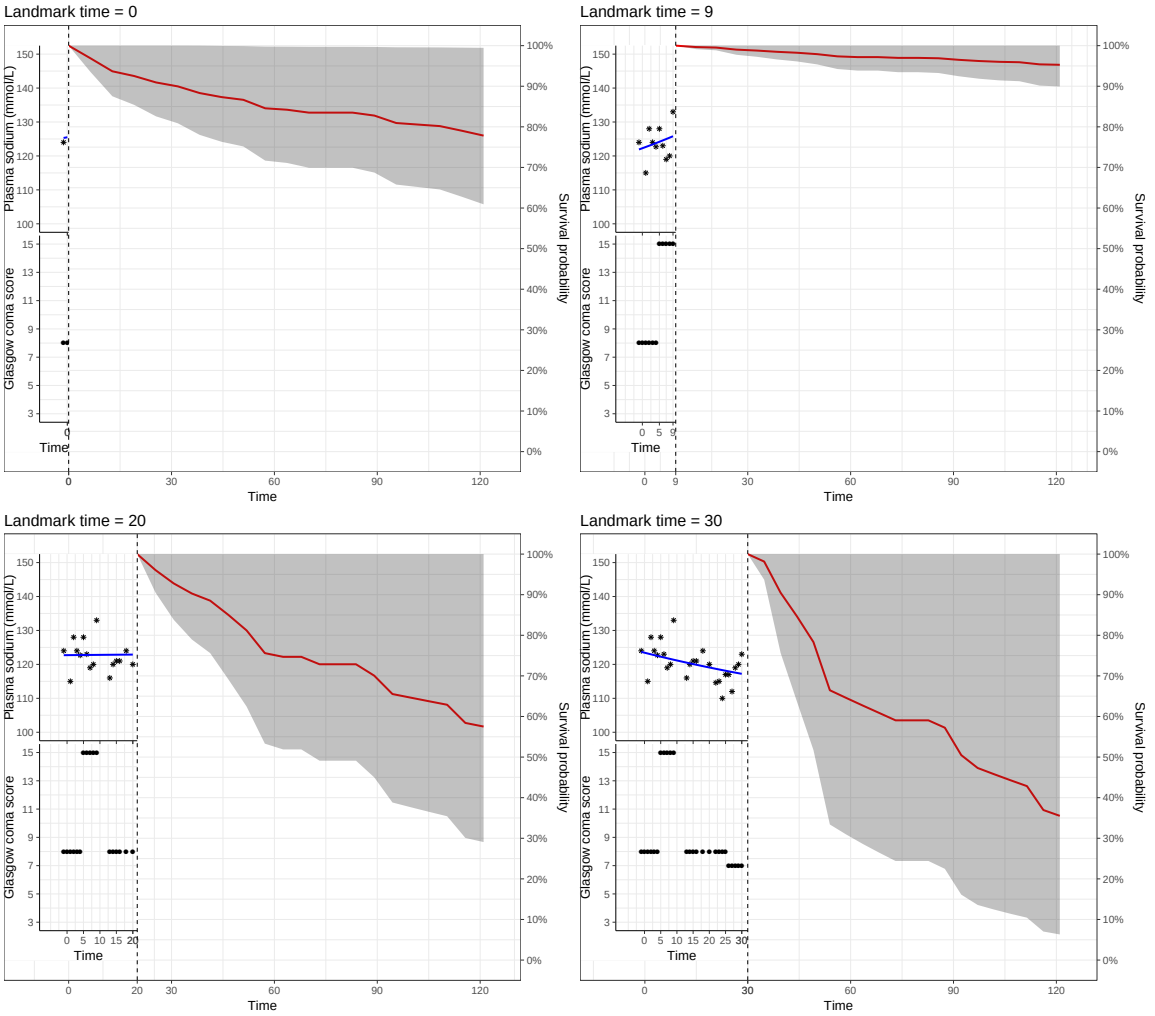


Figure 4.6: *Dynamic predictions of survival for patient A at four different days after the start of TB treatment. For each day, the left panels display the observed values of GCS and plasma sodium over time, the blue line represents the fitted value for sodium. The right panels represent the estimated survival probability in red with 95% confidence intervals in grey*

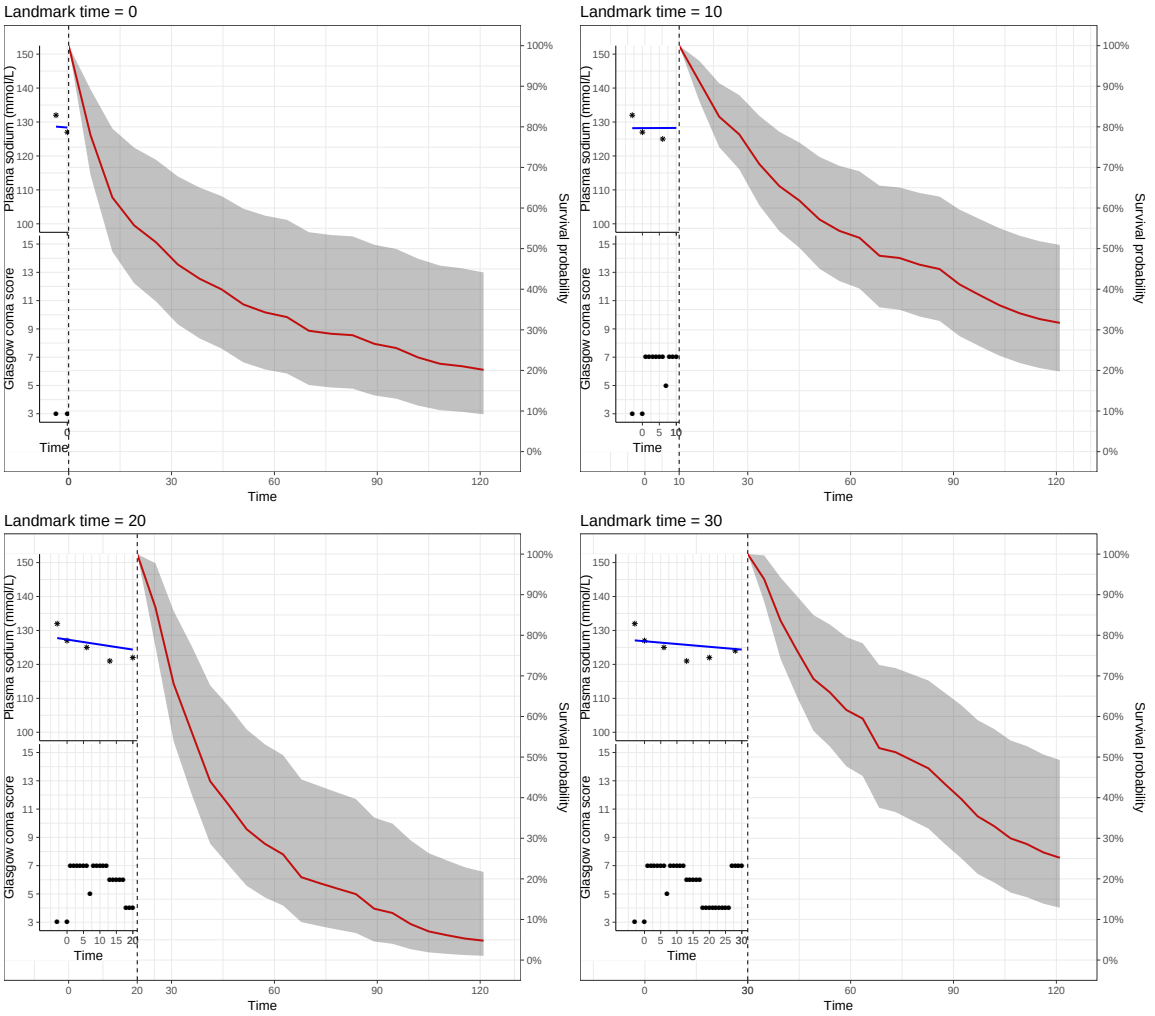


Figure 4.7: *Dynamic predictions of survival for patient B at four different days after the start of TB treatment. For each day, the left panels display the observed values of GCS and plasma sodium over time, the blue line represents the fitted value for sodium. The right panels represent the estimated survival probability in red with 95% confidence intervals in grey.*

4.4 Discussion

Using most recent information on disease state is important for monitoring patients, tailoring treatment and improving disease outcomes. I developed and validated a dynamic prediction model for death in TBM patients with and without HIV coinfection. The models used time-updated information on Glasgow coma score and plasma sodium as well as baseline patient characteristics. Indeed, the dynamic model provided better predictive performance than the model that only used baseline characteristics (Thao et al., 2018b).

The dynamic model confirmed that GCS is a strong predictor of outcome in both populations. Regarding plasma sodium, hyponatraemia is a common complication of TBM, usually caused by cerebral salt wasting or the syndrome of inappropriate anti-diuretic hormone (Misra et al., 2016). Its impact upon survival has not been well-studied, although low sodium concentrations can cause reduced consciousness and exacerbate raised intracranial pressure. Mostly, higher sodium concentrations during treatment indicated better outcome.

This association was uniformly observed over the first 30 days of treatment in HIV-infected patients. The trend was more or less in line with the respective association in the baseline model (Figure 3.3) for sodium lower than 127 mmol/L. Of note, the normal range for plasma sodium in healthy people is between 135 and 145 mmol/L. These numbers could be different in patients with TBM. A wider confidence interval was observed in the plot for the dynamic model, especially at landmark 0. The higher uncertainty could be explained by the smaller number of events considered in this analysis (deaths within 120 days vs deaths within 270 days).

In HIV-uninfected patients higher concentration of sodium predicted worse outcomes in the first 10 days of treatment, and switched to predicting better outcomes thereafter (Figure 4.3). This is an intriguing finding which is not easily explained. A possible explanation may be the contribution of neurogenic diabetes insipidus, which occurs in severe disease affecting the hypothalamic-pituitary access (Anderson et al., 2010), and can cause harmful hypernatraemia. It is unclear why HIV-infection may drive these differences and the findings increase the importance of better understanding and managing disorders of sodium homeostasis associated with TBM.

In the database, there are seven HIV-uninfected patients and eight HIV-infected patients with at least one measurement of plasma sodium higher than 145 mmol/L (hypernatraemia) in the first 30 days. Of those, seven patients died (four in HIV-uninfected and three in HIV-infected patients). All of these deaths happened very

early, within the first 25 days since the start of anti-tuberculosis treatment. Due to the low number of hypernatraemia cases and the timing of the events, the impact of high sodium values on mortality was not captured thoroughly in the dynamic model, especially at later landmark times. The low number of high sodium values may also contribute to the sensitive pattern of sodium to the number of knots in the sensitivity analysis.

Regarding the extra analysis, using a spline function for time in the longitudinal model did not improve the performance. On average, there were only 4 to 6 measurements for sodium in the first 30 days. This low number might prevent a more flexible functional form from showing its advantage in the predictive performance. Regarding the joint model approach, it is expected to be more efficient than the landmarking approach as it models the whole longitudinal process of both markers jointly with the survival process. Indeed, the superior performance of the joint model over the landmarking approach has been demonstrated in several publications (Maziarz et al., 2017; Rizopoulos et al., 2017; Suresh et al., 2017) in case of a correctly specified model. However, in case the longitudinal process is misspecified, the landmarking approach tends to provide better performance (Ferrer et al., 2017; Rizopoulos et al., 2017; Sweeting et al., 2017). As mentioned before, it is difficult to specify the highly nonlinear pattern of GCS in a model. Hence, the misspecification of the longitudinal model for the marker could contribute to the inferior performance of the joint models. In addition, the joint model is very computationally intensive, especially when multiple markers are under consideration.

The study has several limitations. First, I only considered GCS and plasma sodium as time-updated markers. Including other markers could potentially improve the model predictive performance. In order to include them in future studies, they should be collected more frequently over time. Second, the models were developed and validated using data solely from two recruiting centres in the South of Vietnam. This restriction may affect the generalisability of the models as discussed in Chapter 3. Thus, the models require validation in different patient populations.

In conclusion, the dynamic prediction models allowed us to predict 120-day mortality from TBM using time-updated data from the first 30 days of treatment. The models were carefully constructed and validated internally based on a large data set of 1048 patients. To the best of my knowledge, these are the first published dynamic prediction models for TBM. They could be used to provide clinicians and patients better information and facilitate decision making regarding treatment and patient

management. I hope that they will be validated in other patient populations and prove useful in practice.

4.5 Appendix

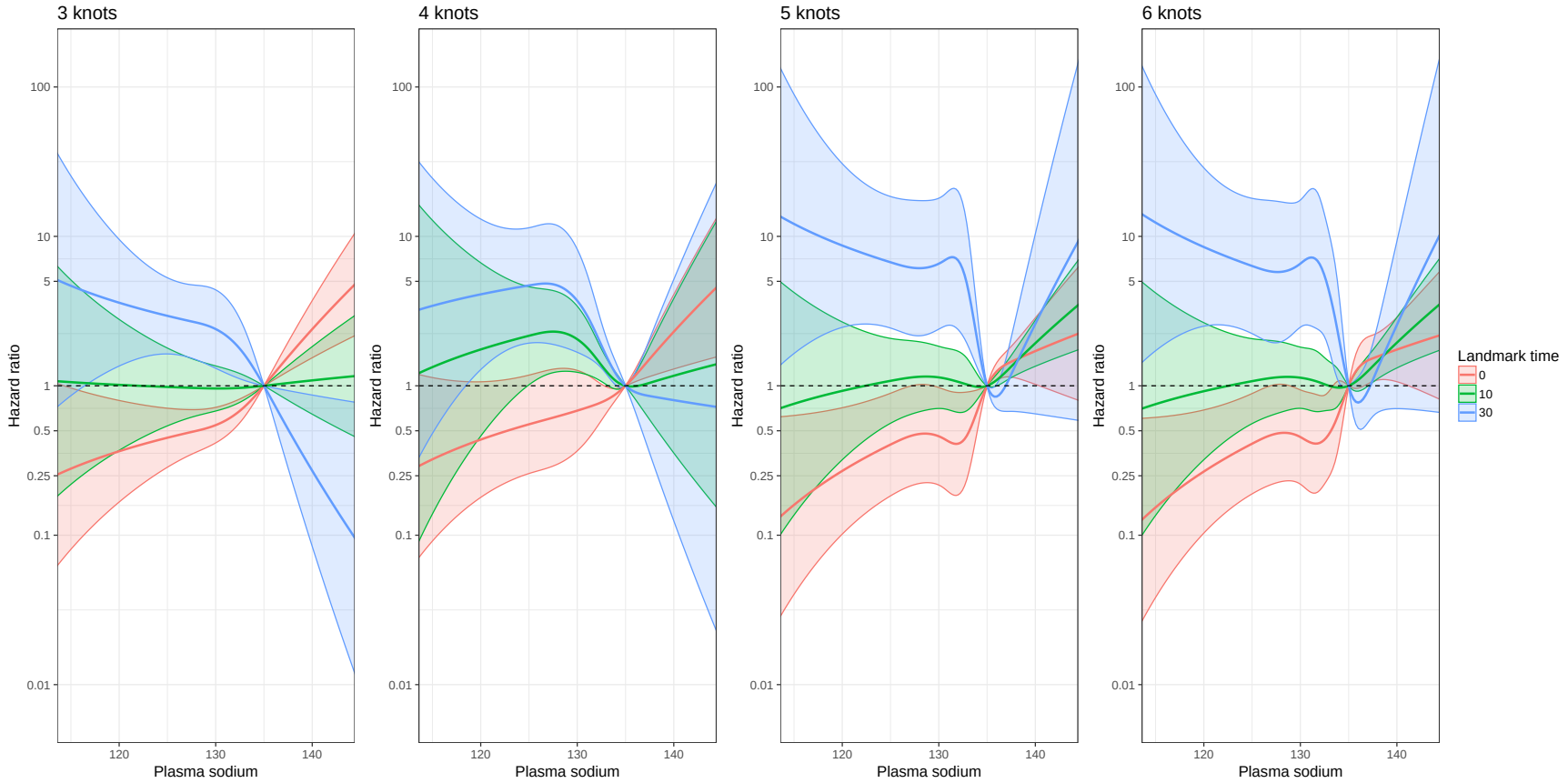


Figure A4.1: *Sensitivity analysis for sodium (HIV-uninfected): Hazard ratio for plasma sodium at three landmark times: day 0, day 10 and day 30 after the start of TB treatment. A sodium value of 135 mmol/L was taken as the reference. The shaded areas represent the 95% confidence intervals*

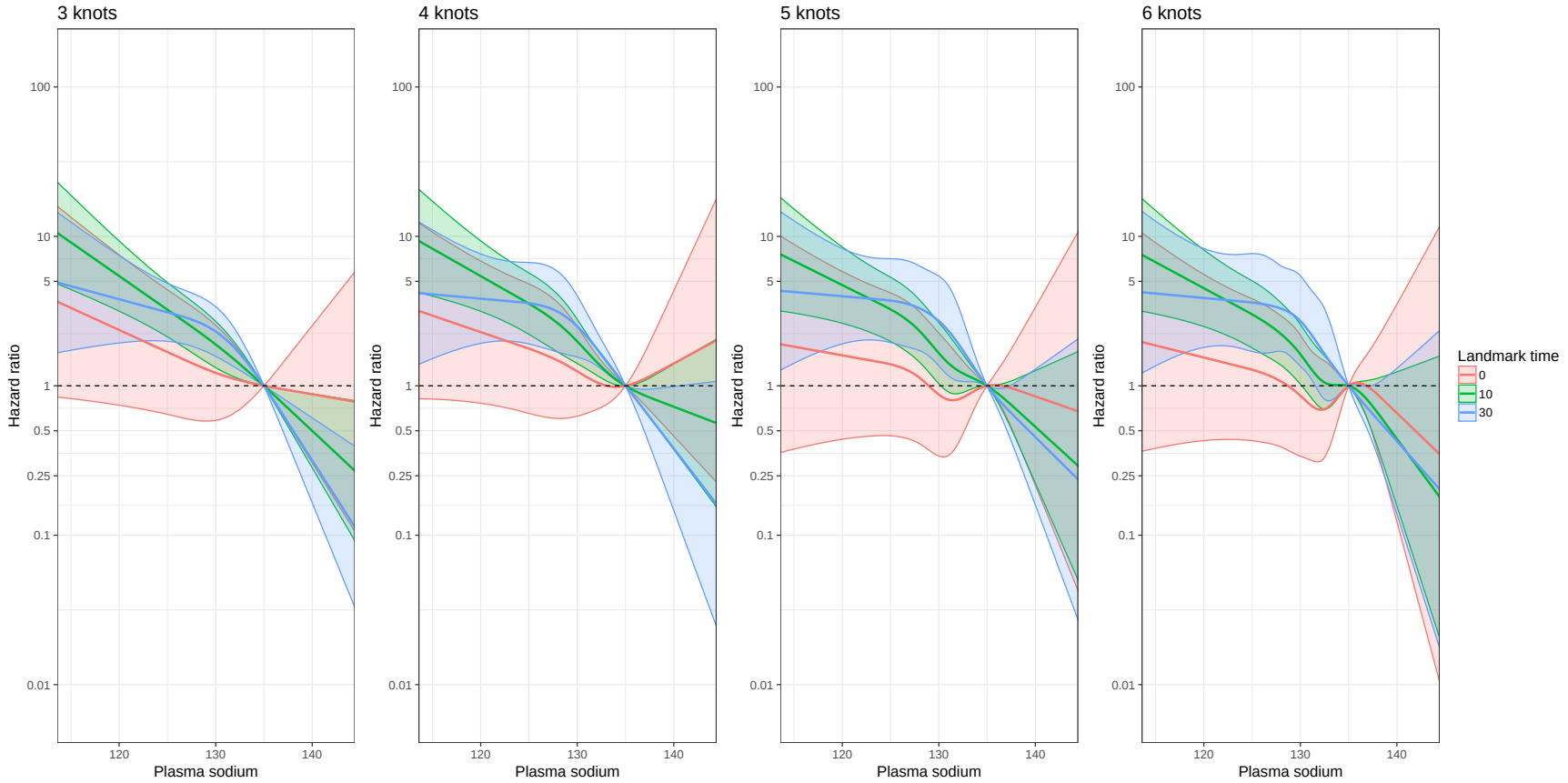
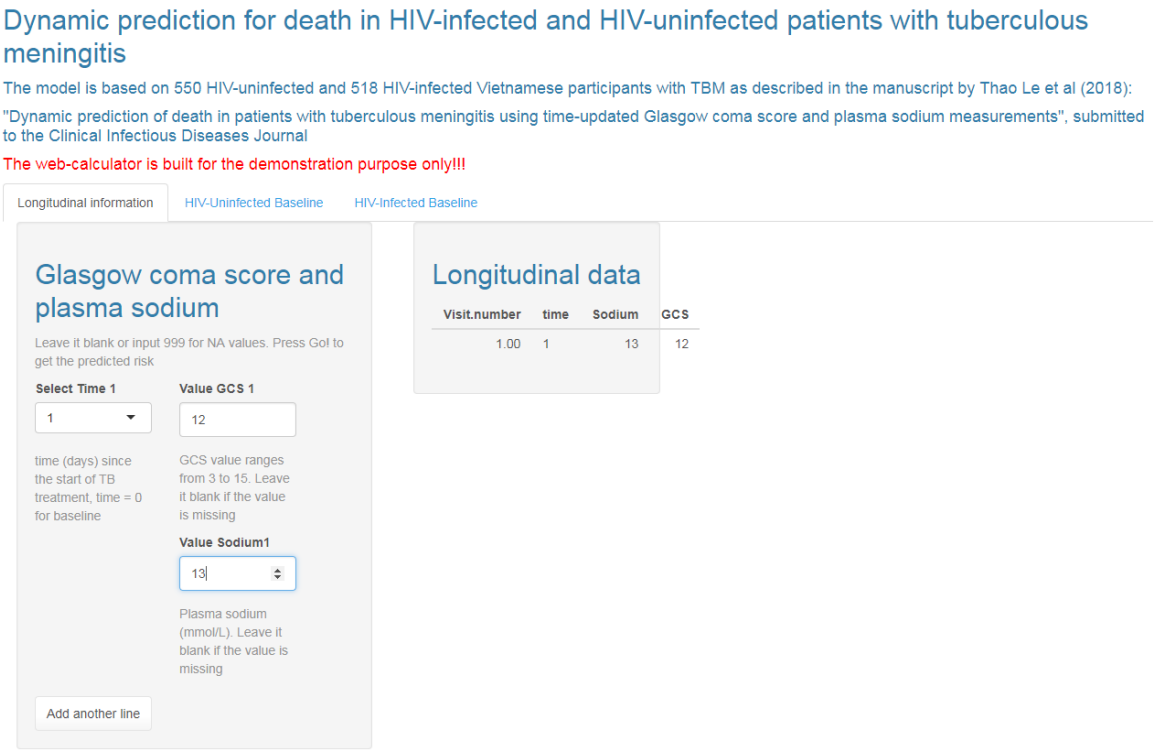
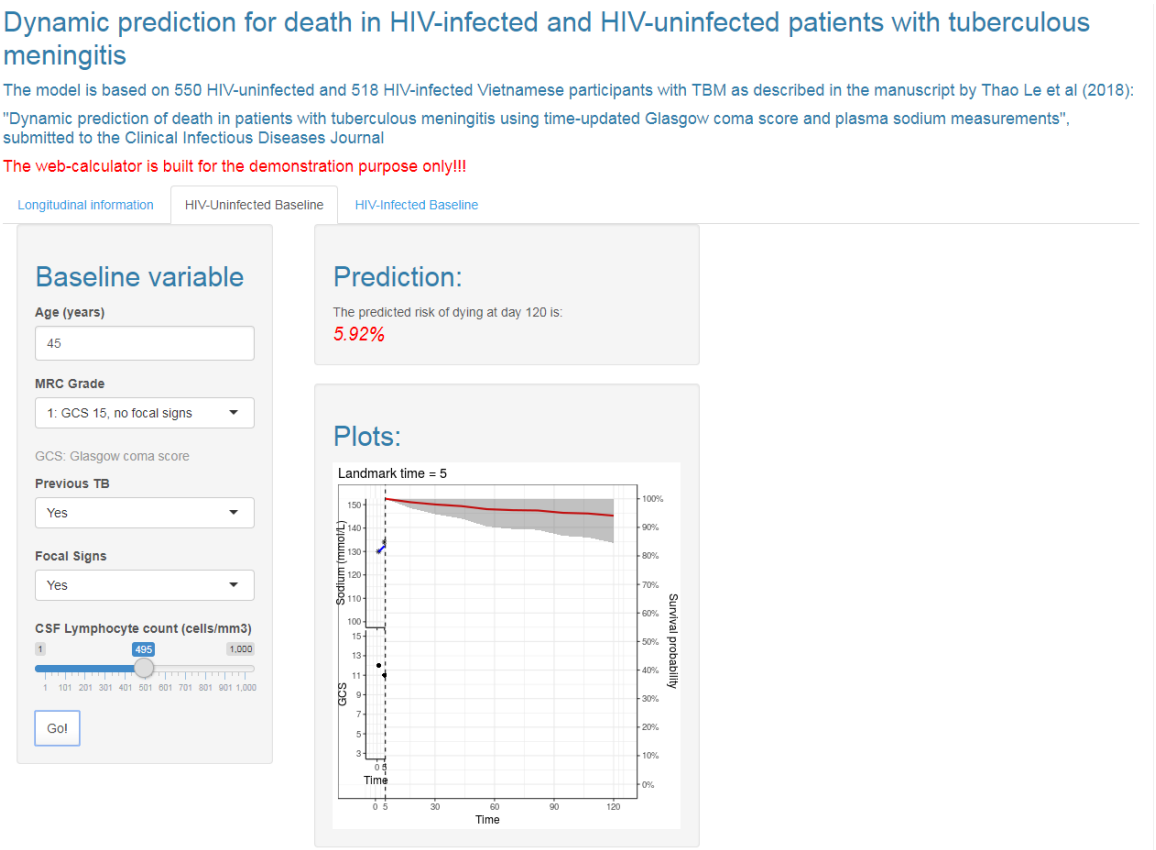


Figure A4.2: *Sensitivity analysis for sodium (HIV-infected): Hazard ratio for plasma sodium at three landmark times: day 0, day 10 and day 30 after the start of TB treatment. A sodium value of 135 mmol/L was taken as the reference. The shaded areas represent the 95% confidence intervals*



(a) *Longitudinal information tab*



(b) *Baseline information tab*

Figure A4.3: *Screenshot for the web-app*

Chapter 5

A comparison of model selection methods for prediction in the presence of multiply imputed data

Summary: In this chapter, I compare two classes of model selection methods in the presence of multiply imputed data. I evaluated the performance of the considered methods on simulated data and on a real dataset of 951 adult patients with tuberculous meningitis to predict mortality within nine months. The contents of this chapter have been published in the Biometrical journal with the title:

Le Thi Phuong Thao and Ronald Geskus. A comparison of model selection methods for prediction in the presence of multiply imputed data. *Biometrical Journal*, oct 2018. ISSN 03233847. doi: 10.1002/bimj.201700232. URL `http://doi.wiley.com/10.1002/bimj.201700232`.

5.1 Introduction

In Chapter 1, I gave an overview of current strategies for incorporating variable selection with multiply imputed (MI) data in prognostic modelling. For the construction of the prognostic model in Chapter 3, I used a bootstrap approach in which I implemented lasso and backward elimination on resampled data sets. After that study had been finished, I reconsidered the approach and wanted to find out whether model

performance could be improved upon if a different model selection approach was chosen. In this chapter, I will describe and compare two classes of approaches: i) lasso or backward elimination in combination with bootstrap sampling, as I used in Chapter 3; and ii) lasso selection on the original data. The chapter is organized as follows. In section 5.2, I describe the model selection methods and the performance criteria. In section 5.3, I compare the methods using a simulation study with a logistic regression model to predict a binary outcome. In section 5.4, I demonstrate the performance of the methods on the data set of patients with tuberculous meningitis (TBM) to predict mortality within nine months. I discuss the findings in the final section.

5.2 Model selection strategies

I consider a logistic regression model given by :

$$\text{logit P}(Y_i = 1|X_{1i}, \ldots, X_{pi}) = \beta_0 + \boldsymbol{X_i^T \beta}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ denotes the binary outcome vector of n subjects, $\boldsymbol{X_i} = (X_{1i}, X_{2i}, \ldots, X_{pi})^T$ is a set of p predictors for subject i, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of p regression coefficients.

5.2.1 Lasso models

If we have a fully observed data set, the lasso coefficient estimates are the solution to the $L1$ optimization problem:

$$\boldsymbol{\beta_{lasso}} = \text{argmin}\Big[- \Big\{ \frac{1}{n} \sum_{i=1}^{n} Y_i(\beta_0 + \boldsymbol{X_i^T \beta}) - \log(1 + \exp(\beta_0 + \boldsymbol{X_i^T \beta})) \Big\} + \lambda \|\boldsymbol{\beta}\|_1 \Big]$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ denotes the $L1$ norm, and $\lambda \geq 0$ is a chosen regularization parameter that controls the amount of shrinkage. The larger the λ, the larger the amount of shrinkage. Large enough λ will set some coefficients exactly equal to zero, thus lasso can perform variable selection.
The λ is chosen based on some loss function that measures the predictive performance. The loss function for logistic regression could be the deviance (minus two times the

maximum log likelihood; the smaller the deviance, the better the fit), the Brier score, or the AUC. I used a single 10-fold cross-validation to select λ. More specifically, I first specified a sequence of values for λ that covers a broad enough range. Next, for each λ, I computed the cross-validated value of the loss function as follows:

- Randomly split the data into 10 equal parts
- For each $k = 1, 2, \ldots, 10$,
  - Leave the k^{th} portion of the data out, and fit the lasso model with the corresponding λ using the other $k - 1$ parts.
  - Evaluate the loss function on the left-out k^{th} part.
- Take the mean of the 10 values of the loss function

This procedure has been implemented in the `cv.glmnet` function in the **R** package `glmnet` (Friedman et al., 2010) with deviance as the default loss function. On fully observed data, I also considered the AUC and the Brier score as alternative loss functions.

I considered both the optimal λ and a suboptimal λ based on the 1-se rule. The former gives the model with the optimal value of the cross-validated loss function. The latter gives the most regularized model (e.g. the simplest model with the largest amount of shrinkage in coefficient estimates) such that the value of the cross-validated loss function remains within one standard error from the optimal value (Hastie et al., 2009). By choosing this suboptimal penalty, we obtain a more parsimonious model. Due to the stronger penalty, the coefficient estimates are shrunk more towards zero, which may impair predictive performance. To improve predictive performance and correct for the over-shrinkage, I considered recalibrating the parameter estimates. Specifically, I refitted the model via unpenalized maximum likelihood based on the selected predictors or the linear predictor (score) from the lasso procedure. For the recalibration by score method, we have a fitted logistic regression model

$$\text{logit}(Y) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k \tag{5.1}$$

with k predictors obtained by performing lasso selection. We fit a regression model with the score $Z = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$ as the only covariate, i.e. $\text{logit}(Y) = \gamma_0 + \gamma_1 Z$. The recalibrated coefficients become $\hat{\beta}_{i,cal} = \hat{\gamma}_1 * \hat{\beta}_i, i = 1, \ldots, k$. Hence, the recalibrated model by score retains the relative effects of the lasso regression coefficients in 5.1. In the more flexible recalibration, a new model is refitted based on

the k selected variables: $\text{logit}(Y) = \beta_{0new} + \beta_{1new}X_1 + \beta_{2new}X_2 + \ldots + \beta_{knew}X_k$. The extension to MI data is explained in the next subsection. Of note, the recalibration was performed on the same data set that was used to derive the model. Hence, it is different from the classical recalibration methods in model validation, which ideally use independent observations, or at least use a resampling technique like cross-validation or bootstrap. For lasso models with the optimal penalty, I did not perform any type of recalibration as the amount of shrinkage is considered to be optimal.

5.2.2 General model selection approaches

I describe the two classes of model selection approaches that I considered in more detail. A summary of all methods is given in Table 5.1.

i. "**Model selection on bootstrap data**": From each of the m MI data sets, I drew B samples with replacement. On each of these mB bootstrap samples, I performed lasso selection (method BLaF) or backward elimination with AIC as selection criterion (method BBeF). The penalty in the lasso model was chosen per bootstrap sample. For each method, the proportion of times each variable appears in the mB models was calculated, and a variable was selected if this proportion exceeds a certain percentage x. I chose $x = 50\%$. To obtain the final model, I refitted the selected predictors to each of the m imputed data sets via unpenalized maximum likelihood and combined results by taking the mean of the coefficient estimates. These approaches are similar to the ones that have been used in Heymans et al. (2007) and Long and Johnson (2014).

ii. "**Lasso on original data**": I performed lasso on the original m MI data sets. I considered two different approaches to combine the m data sets: separate and stacked. In the first one, I applied lasso on each imputed data set separately. The final model was obtained by taking the mean of the coefficients of predictors that appear in a) any of the m models (method SepAv), or b) at least $x\%$ of the models (method SepAvF). I chose $x = 50$. SepAv is similar to method S1 in Wood et al. (2008), while SepAvF is similar to the method investigated by Lachenbruch (2010) , and method S2 in Wood et al. (2008) except that the latter used backward elimination instead of lasso. For lasso models with the 1-se penalty, I first recalibrated the parameter estimates as described above on each MI data set and then took the mean of the estimates over the imputed data set.

Table 5.1: *Summary of considered model selection methods*

Method	Description
FULL	**Full** model with all covariates included
TrueC	Model with all **true** covariates only (i.e. covariates with non-zero coefficients)
Model selection on bootstrap data	
BBeF	**B**ootstrap resampling. **B**ackward **e**limination (with AIC as the stopping rule) on bootstrap data and fit the final model based on variables with inclusion **f**requency $\geq 50\%$
BLaF†	**B**ootstrap resampling. **La**sso selection on bootstrap data and fit the final model based on variables with inclusion **f**requency $\geq 50\%$
Lasso on original data	
SepAv*†	Lasso selection on each original MI data set **sep**arately, and **av**erage coefficient estimates
SepAvF*†	Lasso selection on each original MI data set **sep**arately, and **av**erage coefficient estimates of the variables with inclusion **f**requency $\geq 50\%$
Stack*†	Lasso selection on the **stack**ed data
StackW*†	Lasso selection on the **stack**ed data, with a **w**eighting scheme for each observation

† I add letter "o" after the method abbreviation for model obtained by the optimal λ
$*$ I additionally consider recalibration by score and by selected variables

In the stacked approach, I applied lasso on a single stacked data set (method Stack). I also investigated the value of adding a weighting scheme for each observation to account for the amount of missingness (method StackW). I employed the same weighting scheme as implemented in method MI-WENet by Wan et al. (2015). Specifically, each individual i received a weight w_i, which was defined as :

$$w_i = \frac{1}{m}\Big(\frac{\text{Number of non-missing variables of subject } i}{p}\Big)$$

where m is the number of imputations, and p is the number of covariates. Hence, subjects with more missing values received a smaller weight. In the stacked methods, the cross validation procedure for the selection of λ was implemented "by individual", i.e. in each fold, a subject is either included m times or excluded altogether. By doing so, we account for the variability incurred by the imputation. For each method that uses the 1-se penalty, the recalibrated models were obtained by fitting the score or the selected variables on the stacked MI data.

I compared the methods with a model that uses unpenalized maximum likelihood on either the full set of variables (method FULL) or all variables for which the coefficient is not equal to zero (method TrueC). These two models were fitted on the stacked MI data.

5.3 Simulation study

5.3.1 Simulation design

I considered two data generating mechanisms, differing in the number of covariates p, the percentage of noise variables, and the correlation structure of the covariates.

a) Data generating mechanism one: 15 variables (8 noise) with specific correlation structure

I adapted the simulation design from van Houwelingen and Sauerbrei (2013). Fifteen variables were generated from a multivariable normal distribution with mean 0 and unit variance. I postulated a low overall correlation structure, as visualized in Figure A5.1. Specifically, all correlations were 0 except for $\rho_{1,10} = \rho_{2,6} = \rho_{7,14} = \rho_{9,13} = 0.5, \rho_{4,8} = -0.7, \rho_{7,8} = 0.3, \rho_{11,12} = 0.7$. I dichotomized variables $X_2, X_4, X_5, X_9, X_{10}, X_{11}, X_{12}$ at 0 to create binary variables. I set 7 covariates, X_4 to X_{10}, to be true predictors with the following regression coefficients: $\beta_4 = -0.5, \beta_5 = \beta_6 = \beta_7 = 0.5, \beta_8 = \beta_9 = 1, \beta_{10} = 1.5$. The other 8 covariates were noise variables with $\beta_j = 0$. The binary outcome Y_i for subject i was generated according to: logit $\mathrm{P}(Y_i = 1|X_{1i}, \ldots, X_{15i}) = -0.25 + \boldsymbol{X_i^T \beta}$. The empirical prevalence of the outcome was about 60%. The outcome vector Y was fully observed, and X had missing values. Missing data was generated for variables $X_2, X_5, X_7, X_8, X_{12}$ and X_{14} as specified below.

b) Data generating mechanism two: 25 variables (18 noise) with first-order autoregressive (AR(1)) correlation

I generated 25 variables (X_1 to X_{25}) from a multivariate normal distribution with mean 0 and unit variance, and an AR(1) correlation with $\rho = 0.5$ (the ordering was by variable number) as displayed visually in the Figure A5.1. The coefficients and variable types of the first 15 covariates were the same as in the previous data generating mechanism. The extra ten covariates (X_{16} to X_{25}) were continuous variables with coefficients of zero. The prevalence of the outcome was again about 60%. Miss-

ing data was additionally generated for variables X_{16}, X_{17}, X_{19} and X_{23} as specified below.

Denote R_{ji} as the missing indicator for variable X_j of subject i ; $R_{ji} = 1$ if X_{ji} is missing and $R_{ji} = 0$ if not. I only considered scenarios in which the data is MAR (Little and Rubin, 2002). Let J denote the set of covariates that I chose to contain missing values. The missing data indicator R_{ji} for covariate X_j, $j \in J$ of subject i was generated following the logistic regression model:

$$\text{logit P}(R_{ji} = 1 | \{X_{ki}\}_{k \notin J}, Y_i) = \alpha_0 + 0.5 \sum_{k \notin J} X_{ki} + 0.5 Y_i$$

where α_0 for each simulation run was estimated by numerically solving the equation:

$$\frac{1}{n} \sum_{i=1}^{n} P(R_{ji} = 1) = miss \tag{5.2}$$

Here, I specified value for $miss$. As the missingness probability only depends on the fully observed data, the left hand side of equation 5.2 is the same for all variables with incomplete data. Thus, on average, the proportion of missing values for any variable $X_j, j \in J$ is equal to $miss$, and the total proportion of missing values in the data is equal to: (The number of variables with missing data*$miss$)/p.

Table 5.2: *Simulation scenarios*

	Data generating mechanism one	**Data generating mechanism two**
Number of covariates (p)	15	25
Proportion of missing values per variable ($miss$)	0.1, 0.2, 0.3, 0.4, 0.5	0.1, 0.2, 0.3, 0.4, 0.5
The corresponding:		
- Percentage (%) of missing values in the data *	4, 8, 12, 16, 20	4, 8, 12, 16, 20
- Percentage (%) of complete cases *	65, 44, 30, 20, 12	65, 47, 34, 24, 16
Sample size n (Events per variable EPV*)	200 (5), 400 (10), 600 (15)	200 (3), 400 (7), 600 (10)
Number of imputed data sets (m)	10, 20, 30	10, 20, 30
Number of true/noise covariates	7/8	7/18

* Mean value

I imputed missing data using multivariate imputation by chained equations (mice) (van Buuren, 2012) as implemented in the **R** package `mice` (van Buuren and Groothuis-Oudshoorn, 2011). Details of the procedure can be found in Chapter 2.

I investigated the performance of the methods described in Section 5.2 for different values of the proportion of missing values (*miss*), the sample size, and the number of imputations. Table 5.2 gives a summary of all the scenarios that I considered in this study. I defined the number of events per variable (EPV) as the number of "events" divided by the number of parameters in the model. The number of "events" is the smaller value when comparing the number of subjects who have the outcome versus the number of subjects who do not (Austin and Steyerberg, 2017). As the proportion of events in each data set was around 60%, in each data generating scheme a higher EPV corresponded to a larger sample size. The EPV, the percentage of missing values in the data, and the percentage of complete cases were empirically estimated. For each scenario, 500 data sets were generated. Regarding the loss function in lasso models, I only considered the fully observed data with different EPV values (Table 5.2). Again, 500 data sets were generated for each scenario.

Performance criteria:
I assessed the predictive performance of the models using Brier score and AUC (Steyerberg et al., 2000). The evaluation was performed via external validation. A validation data set was generated from the same data generating mechanism with a sample size of 5000 fully observed observations. I also compared selection performance via the number of selected true and the number of selected noise variables. Better performance is observed with higher value of AUC, larger number of selected true predictors, lower value of Brier score and smaller number of selected noise predictors.

5.3.2 Results

5.3.2.1 Fully observed data

Predictive performance
I compare the performance of lasso models obtained by using different loss functions for the selection of the penalty: deviance, brier score and AUC. Figure 5.1 represents the mean AUC and Brier score over 500 data sets generated from data generating mechanism one. A similar figure for data generating mechanism two is displayed in Figure A5.2. Of note, recalibrating a model by score does not change its AUC value because the ordering of the predictions is preserved.

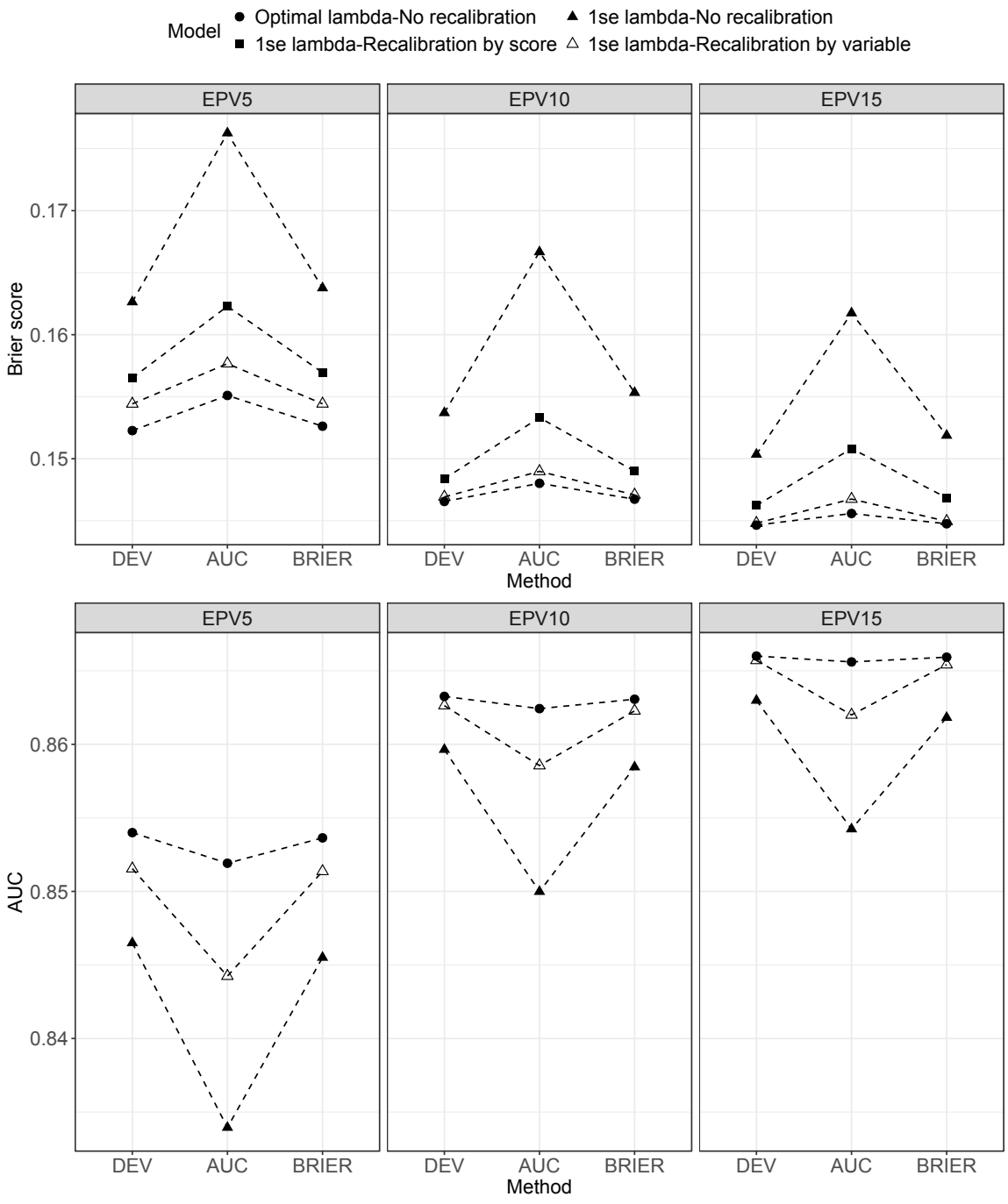


Figure 5.1: *Fully observed data - first data generating mechanism (section 5.3.1a)): Mean AUC and Brier score over 500 fully observed generated data sets for lasso models with three loss functions for the selection of the penalty. DEV: deviance; BRIER: brier score; AUC: AUC as loss function.*

The results from the two data generating mechanisms were largely similar. Specifically, the deviance and the Brier loss function resulted in models with analogous per-

formance. They outperformed lasso models with AUC as a loss function, even when AUC was the performance measure. Regarding the choice of λ, for each loss function, the optimal λ provided better performance than the 1-se. However, this comes at a price. On average, the optimal λ included about three more noise variables in the model (Figure 5.2).

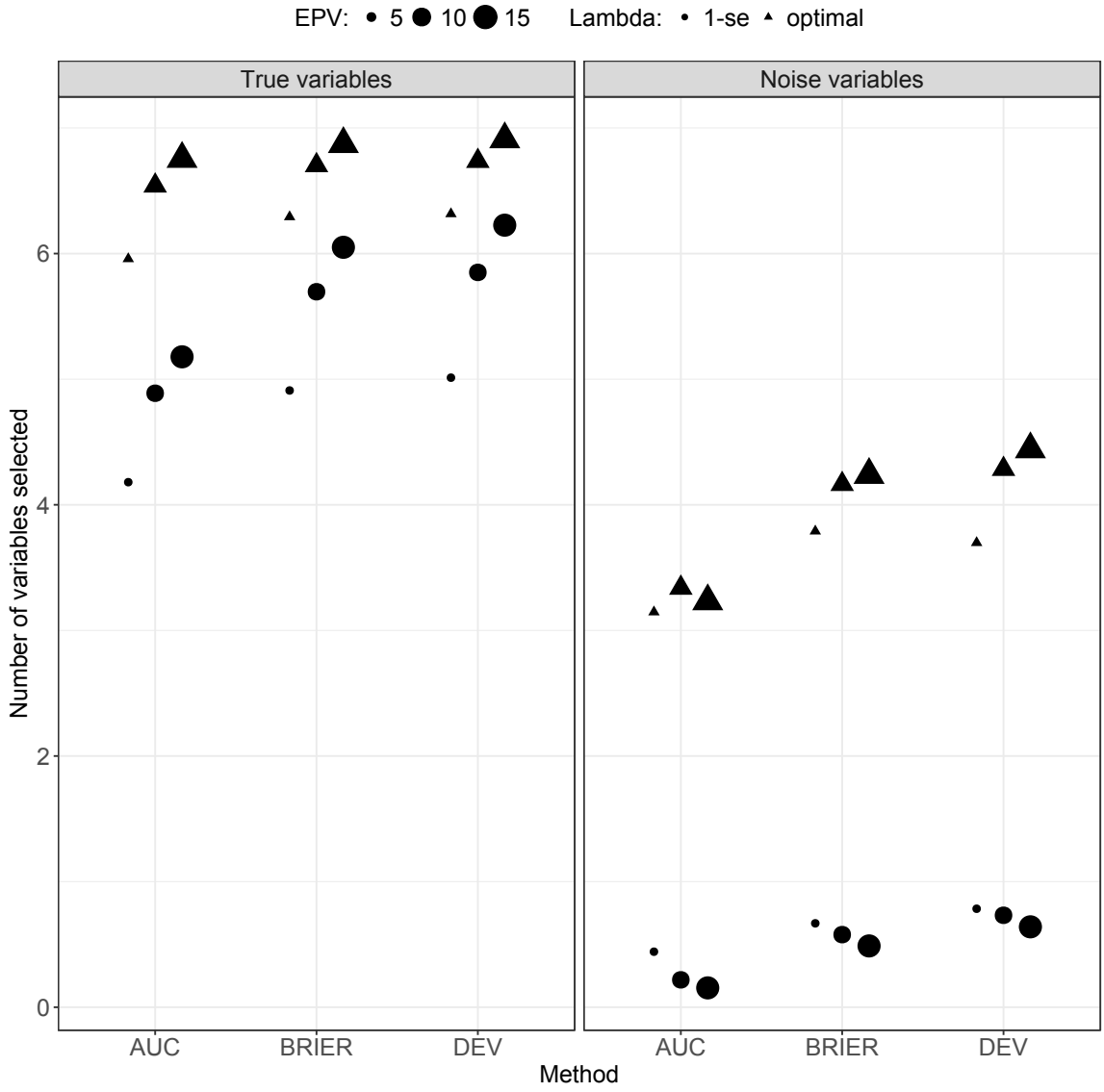


Figure 5.2: *Fully observed data - first data generating mechanism (section 5.3.1.a): Mean number of selected variables for the lasso models with three loss functions for the selection of the penalty. DEV: deviance; BRIER: brier score; AUC: AUC as loss function*

Interestingly, recalibration improved performance in models with the 1-se λ. In

fact, after recalibrating by selected variables, the obtained performance was approximated that of the optimal λ. The difference in performance between the recalibration by variables models and the optimal penalty models decreased with increasing EPV. Because the other loss functions did not improve the model performance, I used the deviance as loss function for all lasso models in the subsequent analyses.

5.3.2.2 Multiply imputed data

Data generating mechanism one:

Predictive performance

The AUC and Brier score are presented in Figure 5.3 for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. The results for all scenarios together with the $\pm$ 1 empirical standard deviations are summarised in the Figure A5.3.

Amongst "model selection on bootstrap data" methods, the lasso-based method BLaF with the 1-se penalty resulted in a model with higher AUC and lower Brier score. The difference between the three bootstrap methods (backward elimination BBeF, and the two BLaF methods with optimal and 1-se penalty) diminished with increasing EPV.

In the "lasso on original data" approach, models obtained by the optimal penalty outperformed those obtained by the 1-se penalty without recalibration, except for the AUC value with the smallest number of EPV (EPV 5) and the largest percentage of missing (20%). The differences between the two penalty selection strategies were more pronounced in scenarios with 4% missing than with 20% missing data. Hence, the difference in performance seems to increase with a larger amount of information in the data.

In data with a low to moderate amount of missing values (4-8%), recalibrating by selected variables improved model performance. In the scenario with the largest amount of information (4% of missing values and EPV 15), it even yielded models with comparable AUC and Brier score to those obtained with the optimal penalty. On the other hand, when the information in the data was low (EPV5 or 10, and 20% missing), recalibrating the models by score gave the better Brier score, and with EPV 5 performance was comparable to the optimal penalty. Interestingly, in this scenario, the recalibration by score model of SepAv (i.e. the mean of the lasso parameter estimates over all imputed data sets) even gave a lower Brier score than when the optimal penalty was chosen. However, the difference was very small (0.158 vs. 0.159).

The most noticeable discrepancy between the two recalibrated models was observed in the method SepAv .

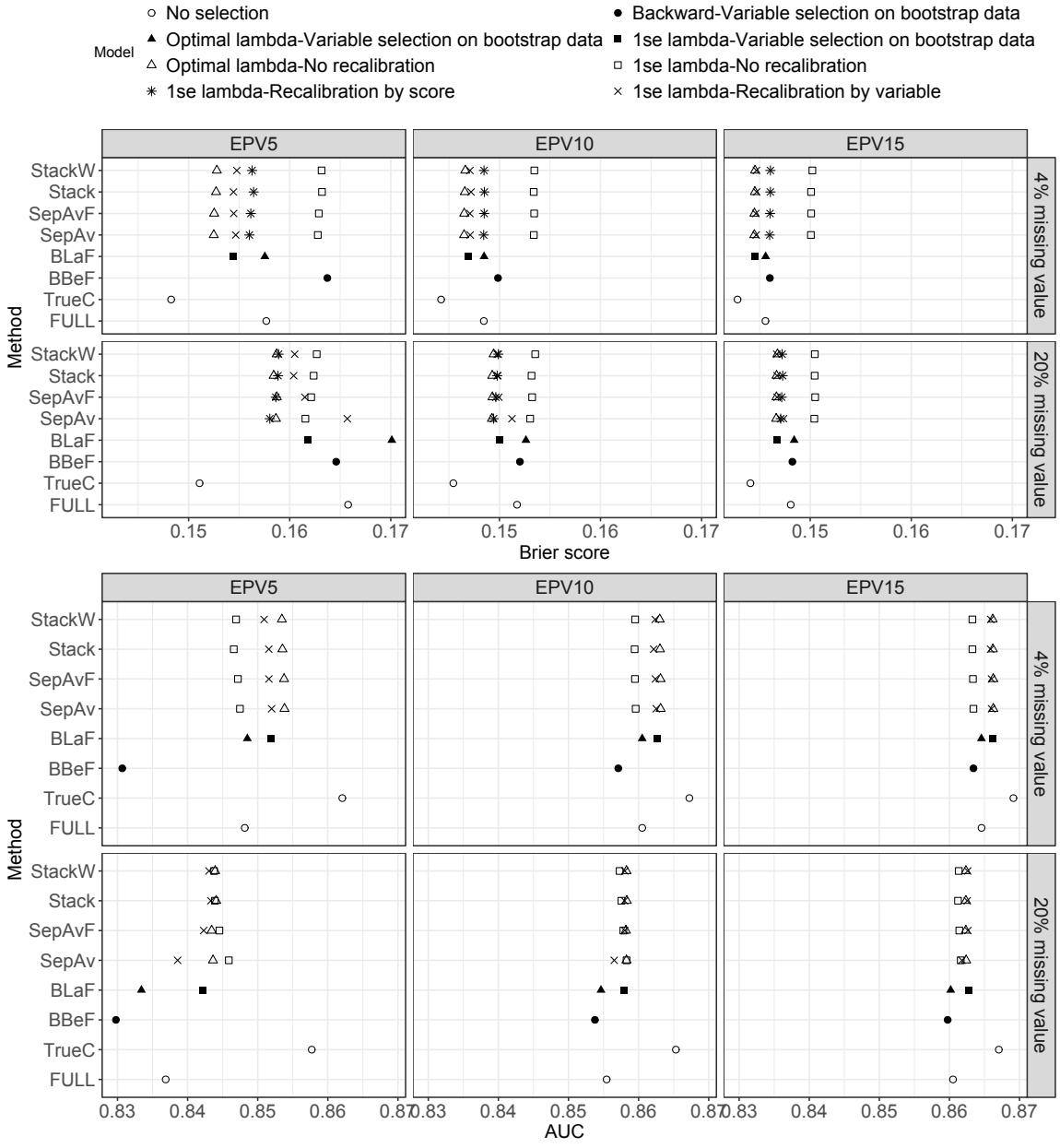
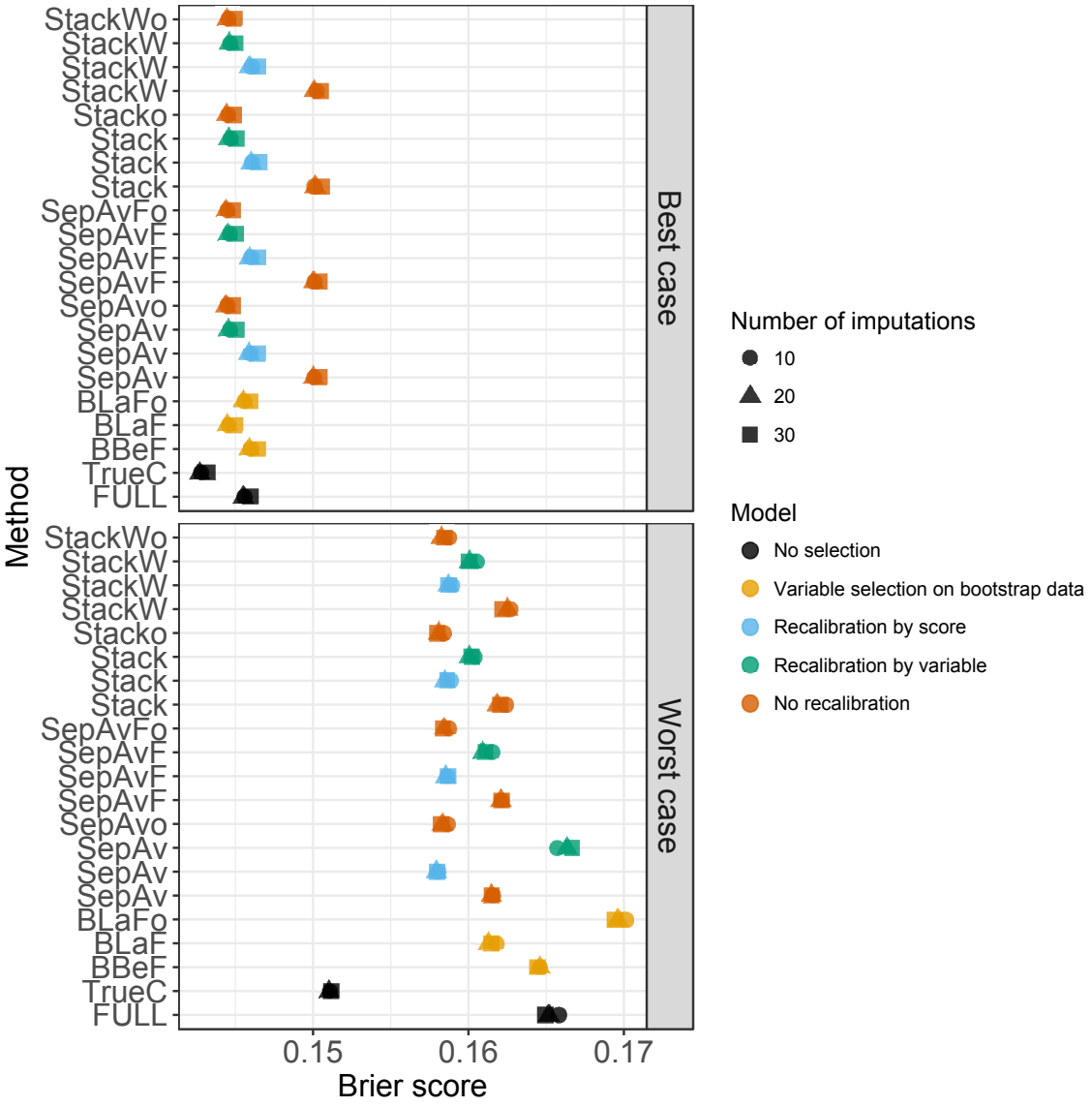


Figure 5.3: *First data generating mechanism (section 5.3.1.a): Mean Brier score (top figure) and AUC (bottom figure) over 500 generated data sets for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets.*

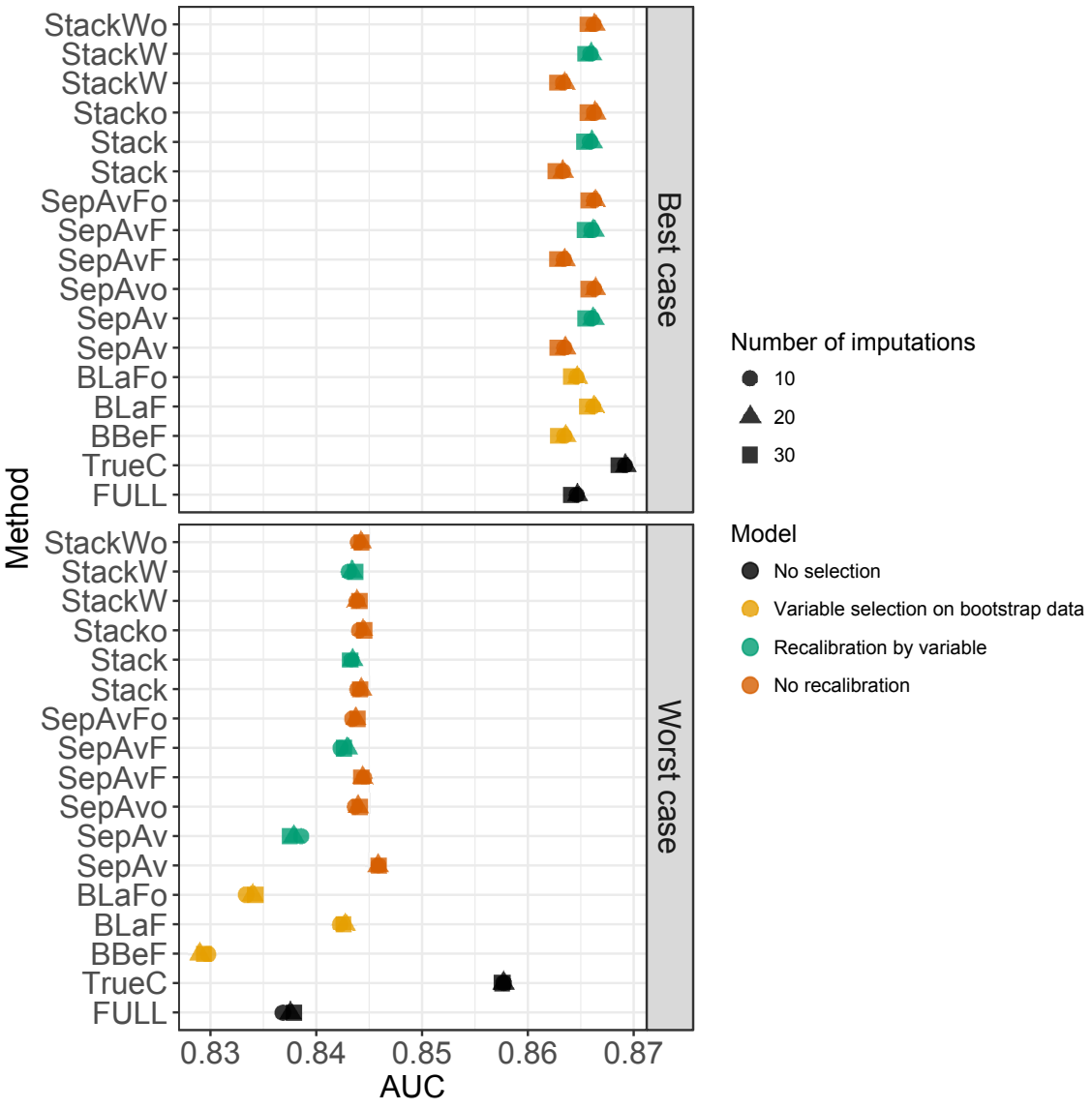
The Stack and StackW methods that select the variables based on the stacked data yielded very similar predictive performance overall. BLaF with the 1-se penalty

provided comparable performance with SepAvF, Stack and StackW if recalibration was done by the selected variables.

The number of imputations had little impact on model performance (see Figure 5.4). Hence, ten imputed data sets seems adequate in this setting.



(a) *Brier score*



(b) *AUC*

Figure 5.4: *First data generating mechanism: The effect of the number of imputations on model performance in the two most extreme data scenarios. Brier score (top figure), AUC (bottom figure). Best case: 4% missing and 15 EPV; Worst case: 20% missing and 5 EPV*

Regarding dispersion (e.g. the variability in the performance estimates across simulated data sets) (see Figure A5.3), all methods gave smaller empirical standard deviation with increasing EPV. The amount of dispersion was similar amongst methods across all scenarios. The only exception was in data with 5 EPV and 20% missingness. In this setting, method BLaF with the optimal penalty and the recalibration-by-selected-variables version of SepAv showed a larger dispersion in their predictive

performance results. Regarding confidence intervals, the widths of the confidence intervals were small in all scenarios, ranging from 0.001 to 0.004 for AUC, and 0.001 to 0.003 for Brier score.

Selection performance

The numbers of selected true and noise variables are depicted in Figure 5.5, again for the two most extreme cases of missing values with 10 imputed data sets. As

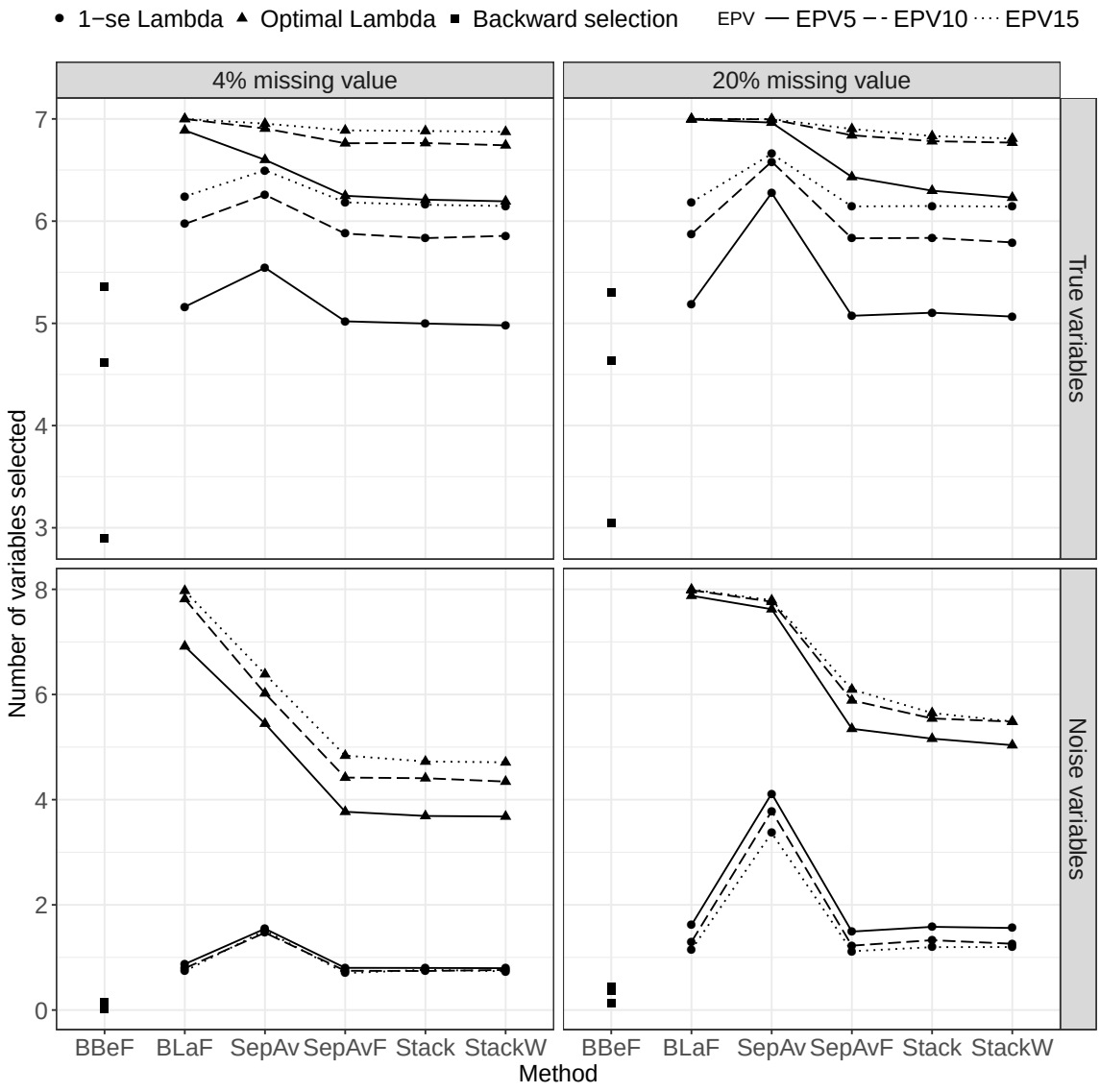


Figure 5.5: *First data generating mechanism (section 5.3.1.a) : Mean number of selected variables for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets.*

expected, lasso models with the optimal penalty included more variables than those with the 1-se penalty. It included most of the true variables, but also a fair amount of noise variables. If parsimony is the primary concern, this suggests that choosing the optimal penalty is not recommended as it results in a model with very low selection specificity. BLaF with the optimal penalty was the worst method in terms of selection specificity. This method selected nearly all noise variables on average. After BLaF with the optimal penalty, SepAv selected the highest number of both noise and true variables. This is not surprising, as it included all variables that were selected based on at least one of the imputed data sets. It also selected more noise variables as the number of imputations increased from 10 to 30 (see Figure 5.6).

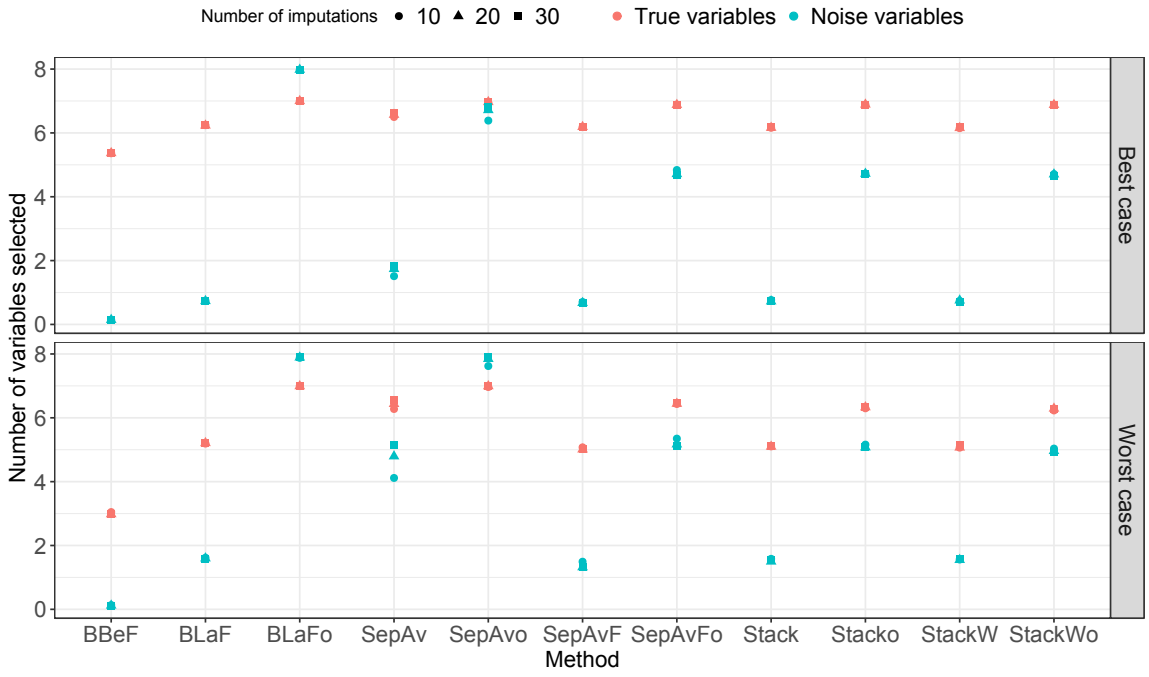


Figure 5.6: *First data generating mechanism: The effect of the number of imputations on the number of selected variables in the two most extreme data scenarios. Best case: 4% missing and 15 EPV; Worst case: 20% missing and 5 EPV*

Of note, BBeF gave the most parsimonious model in all considered scenarios and selected very few noise variables. However, in data with low EPV, this method on average included less than half of the number of the true predictors.

Interestingly, models with the optimal penalty included more noise variables as the EPV increased. On the other hand, with increasing EPV, model selection approaches with the 1-se penalty tended to include nearly the same amount of noise variables in data with 4% missing values, and even fewer in data with 20% missing. The

increase in EPV clearly improved the 1-se penalty models with respect to the number of selected true predictors. On average, less than one true variable was missed if EPV is 15.
As the missingness increased from 4% to 20%, all methods selected more noise variables. The difference was more pronounced in method SepAv and models with the optimal penalty except for method BLaF as it already selected almost all 8 noise variables.

Data generating mechanism two:
The predictive and selection performance for the two most extreme cases of missing values with 10 imputed data sets are presented in Figure 5.7 and Figure A5.4. With 4% missing, conclusions were very similar to those from the first data generating mechanism. In data with 20% missing, lasso on the original data using the 1-se penalty and recalibrating by score showed the best performance with respect to Brier score in all three EPV scenarios.

I observed interesting results in EPV3. Without recalibration, selection methods with the 1-se penalty performed better than the alternatives with the optimal penalty and the BLaF methods. The results for the BBeF methods in this scenario could not even be shown because the method failed to converge in the majority of simulation runs (about 95%). Recalibrating by selected variables showed a detrimental effect on Brier score and AUC.

The dispersion of the predicted performance is presented in the Figure A5.5. Again, the patterns were very similar to those in the first data generating mechanism. In all methods, the widths of the confidence intervals ranged from 0.001 to 0.008 for AUC and from 0.001 to 0.005 for Brier score. Just like in the other data generating mechanism, models with the optimal penalty selected a substantially larger number of noise variables than models with the 1-se penalty in all scenarios (see Figure A5.4).

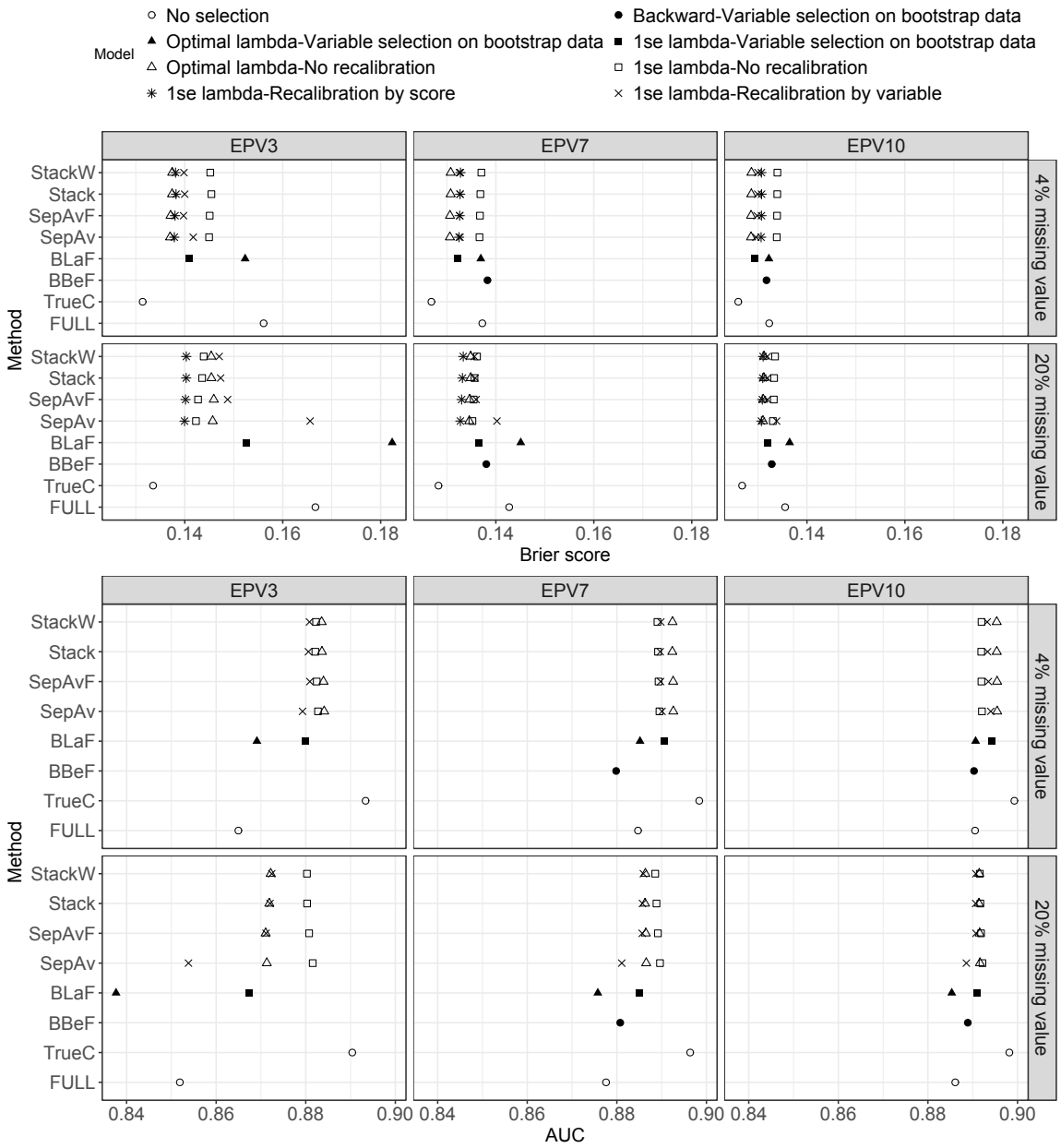


Figure 5.7: *Second data generating mechanism (section 5.3.1.b): Mean Brier score (top figure) and AUC (bottom figure) over 500 generated data sets for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. The performance measures of BBeF in data with EPV of 3 are not available due to convergence problems in many generated data sets.*

Conclusions from the simulation studies

In the lasso models, using the optimal penalty resulted in models with better predictive performance compared to the 1-se rule penalty without correction for over-

shrinkage, unless the amount of information became very low. This gain in predictive performance came with a larger number of selected noise variables. Imposing a stronger penalty, like the 1-se, is preferable if parsimony is important. Moreover, if I corrected for over-shrinkage in the 1-se method, model performance was improved without losing parsimony. In case the EPV was 15, performance approximated that of the optimal penalty if I refitted the model based on the selected variables. On the other side of the spectrum, if the amount of information was very low, refitting based on the score could improve the Brier score to values lower than those from the optimal penalty (Note that recalibrating the score has no effect on the AUC). Method BLaF, and the recalibration-by-selected-variables version of SepAvF, Stack and StackW showed similar predictive performance in most of the data scenarios. The recalibration-by-selected-variables version of SepAv achieved similar predictive performance only when the level of missingness was low to moderate.
BBeF was the most parsimonious model with comparable predictive performance only in data with high EPV and low percentage of missingness. Note however that I only considered the AIC as the stopping rule for variable selection. Performing backward elimination based on larger p-values will create a less parsimonious model with better predictive performance.

5.4 Real data application

5.4.1 Data description

I used the TBM data set for HIV-uninfected patient described in Chapter 3. In total, 951 HIV-uninfected patients were included. The goal was to predict mortality within nine months. Results based on an internally validated Cox proportional hazards regression model were presented in Chapter 3. Specifically, the final model was obtained by performing lasso selection with the 1-se penalty in combination with bootstrapping (BLaF). The chosen variable inclusion frequency in Chapter 3 was 60%. In this analysis, I considered a variable inclusion frequency of 50% to be consistent with the simulated data. I compare these results with the predictive performance (AUC) of the other variable selection methods.

To be consistent with the statistical model used in this simulation study, I additionally used a logistic regression model with nine-month mortality as outcome and a variable inclusion frequency of 50%. For that model, I also included the Brier score.

Within those nine months of follow-up, 219/951 patients died, while 44/951 were lost to follow-up. As the number of patients lost to follow-up is less than 5%, excluding them from the analysis will not have a major impact on the results (White et al., 2011). I applied each of the MI variable selection procedures in Table 5.1 to the same set of candidate predictors as considered in Chapter 3 (see Table 3.1). Amongst them, age, weight, sex, treated with dexamethasone, Medical Research Council (MRC) Grade, and cohort were fully observed. The percentage of missing values in the other predictors ranged from 0.1% (0.1% [1]) in temperature to 12.9% (12.7% [1]) in the ratio of CSF to blood glucose. The total percentage of missing values in the data was 2.7% (1.7% [1]), and the percentage of complete cases was 67.4% (67.8% [1]). The number of events per variable was 11.7. Ten imputed data sets were generated using the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011). I implemented internal validation using the enhanced bootstrap resampling procedure described in Chapter 2.

5.4.2 Results

Results of the logistic regression models

Figure 5.8 presents the bias-corrected predictive model performance. The results for the variable inclusion frequency by the model selection on bootstrap data methods are presented in the Table A5.1. In line with the results from the simulations, BBeF yielded the most parsimonious model, with only five variables included: Age, MRC Grade, presence of focal neurological signs, CSF lymphocyte count, and cohort. Models obtained by lasso with the 1-se rule, i.e. BLaF, SepAv, SepAvF, Stack and StackW, selected the same eight predictors (Table 5.3). After recalibration by selected variables, they gave similar bias-corrected predictive performance measures (AUC ranges from 0.775 to 0.776, Brier score = 0.155) and slightly outperformed BBeF (AUC = 0.767, Brier score = 0.157) and the full model (AUC = 0.773, Brier score = 0.156). The additional three predictors, namely the use of dexamethasone treatment, weight and the history of TB, are easy to measure and do no incur any extra costs in practice. Choosing the optimal penalty, on the other hand, gained minor improvement in the predictive performance at the cost of two to seven extra variables. Two of them, ratio of CSF to blood glucose or CSF protein, required significant extra effort to measure (Table 5.3).

[1] Numbers in the parenthesis represent the corresponding values for the TBM data before excluding censored patients

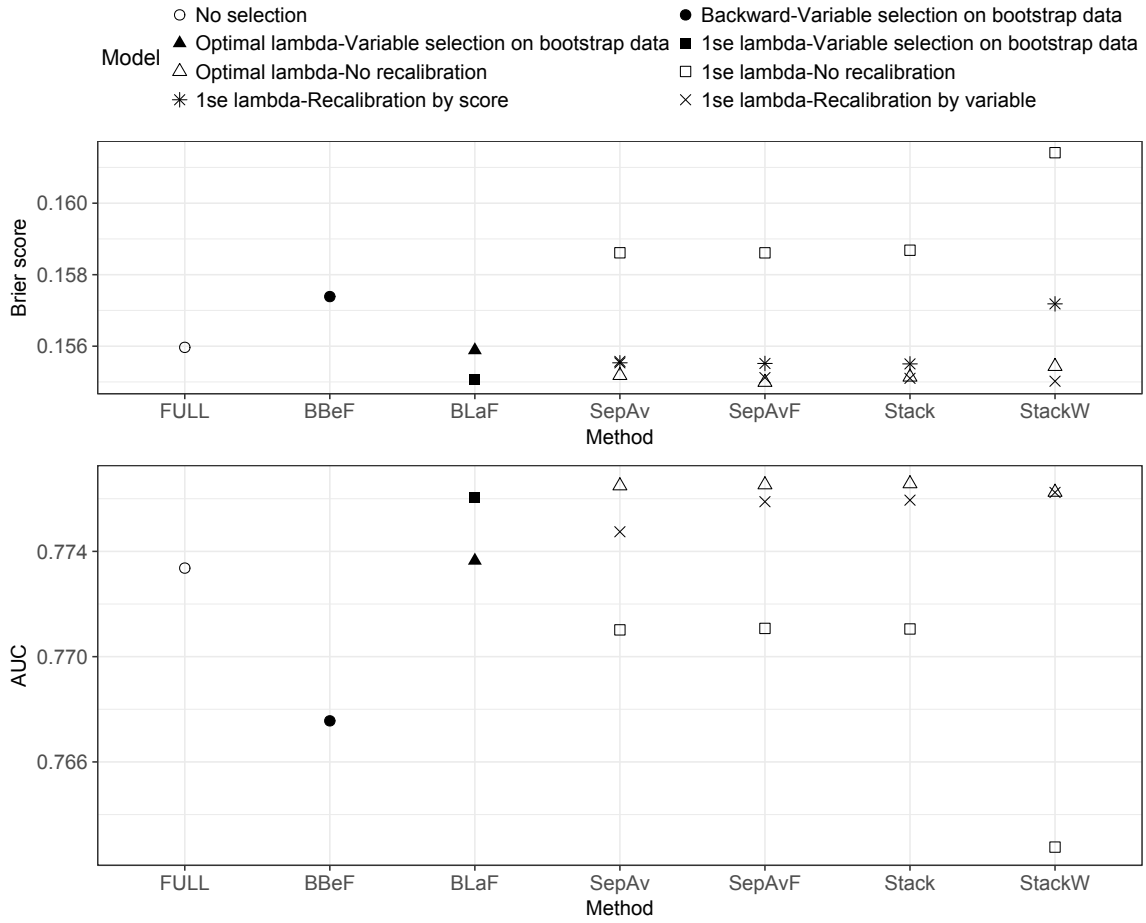


Figure 5.8: *TBM data: Corrected AUC and Brier score via internal validation, with logistic regression models*

Table 5.3: *Variable selection in the TBM data set, all models*

Variables	**Method**											
	BBeF	BLaF	SepAv	SepAvF	Stack	StackW	BLaFo	SepAvo	SepAvFo	Stacko	StackWo	FULL
Age	x	x	x	x	x	x	x	x	x	x	x	x
MRC grade	x	x	x	x	x	x	x	x	x	x	x	x
CSF lymphocyte count	x	x	x	x	x	x	x	x	x	x	x	x
Presence of focal neurological signs	x	x	x	x	x	x	x	x	x	x	x	x
Cohort	L	L	L	L	L	L	L	x	x	x	x	x
Received dexamethasone treatment	C	x	x	x	x	x	x	x	x	x	x	x
Previous TB		x	x	x	x	x	x	x	x	x	x	x
Weight		L	L	L	L	L	x	x	x	x	x	x
Ratio of CSF to blood glucose							L	x	L	L	L	x
Illness duration at entry							x	x	L	L	L	x
CSF protein							x	x	L	L		x
Occurrence of seizures							L	L				x
Plasma sodium							x	L				x
Chest x-ray miliary TB							x					x
Sex							L					x
CSF glucose												x
Body temperature												x

Abbreviations: L: Logistic regression model; C: Cox regression model; x: both models, TB: tuberculosis

Results of the Cox regression models

Figure 5.9 depicts the bias-corrected cumulative/dynamic AUC evaluated at 9 months for the Cox regression models. The results for the selected variables are presented in Table 5.3.

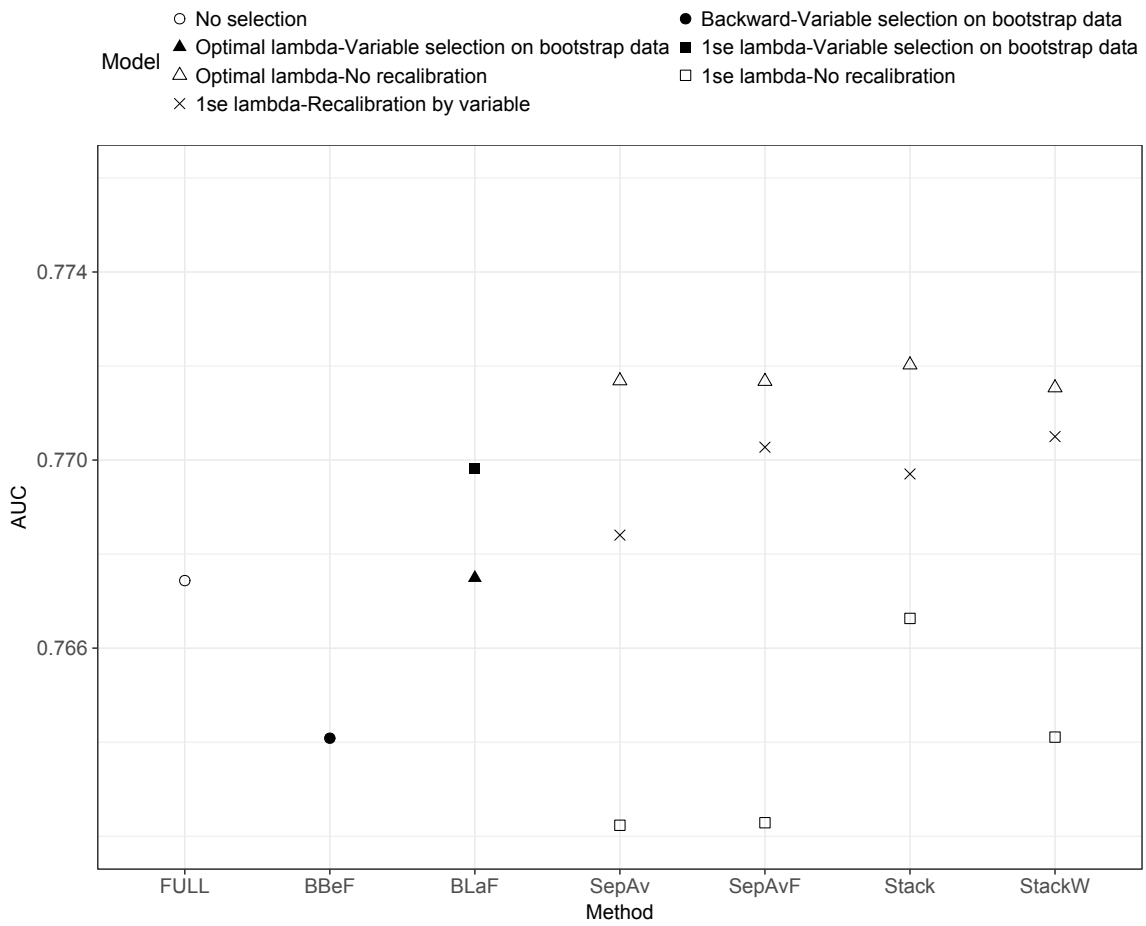


Figure 5.9: *TBM data: Corrected cumulative/dynamic AUC via internal validation, with Cox proportional hazard models*

Regarding predictive performance, conclusions are very similar to those obtained from the simulated data and the real data application with logistic regression chosen as the statistical model. The BLaF, SepAv, SepAvF, Stack and StackW with the 1-se penalty gave analogous AUC values after recalibration by selected variables (AUC ranges from 0.768 to 0.770). They also selected the same six predictors (Table 5.3). This is in concordance to the results presented in Chapter 3 in which BLaF with the 1-se penalty was implemented to obtain the final model. The BBeF method selected one predictor less than the BLaF, and resulted in a model with a smaller AUC value

(AUC = 0.764). Of note, although the SepAvF, Stack and StackW with the optimal penalty selected two more predictors (weight and Cohort), the gains in AUC values are again small.

5.5 Discussion

I described and compared different popular variable selection methods in the presence of multiply imputed data when the aim is to build a prediction model. I assumed the missing data mechanism to be MAR. Results from the simulated data and the real data application were largely similar with respect to the relative predictive performance and number of selected variables. Overall, no method prevailed as uniformly best over all scenarios that I considered in the simulation study. Performance depended on the amount of information in the data. Imposing a stronger penalty than the "optimal" one provided more parsimonious models that might have only slightly inferior and sometimes even better model performance after recalibration. Stacking all MI data was an attractive method. Although using lasso on bootstrap data or on original MI data separately provided just as good performance, it is likely to be sensitive to the chosen variable inclusion threshold. Performing backward elimination on bootstrap data with AIC as stopping rule gave the most parsimonious model. I discussed characteristics of each considered method in more detail.

The first class of methods, variable selection on bootstrap data, has been widely used in model development. This is the method that I implemented in Chapter 3 with the variable inclusion frequency of 60%. In this simulation, I considered an inclusion threshold of 50%. It was observed that the predictive performance of this approach depends on the selection method (backward elimination or lasso) and the amount of information in the data.

Regarding the impact of the selection method, backward elimination based on AIC provided a more parsimonious model but with inferior predictive performance compared to lasso. However, performance of backward elimination was likely to depend on the choice of the stopping rule. In this study, I only considered the popular AIC as stopping criterion, which roughly corresponds to a p-value of 0.157. More liberal choices of the significance level, e.g. 0.5 or 0.1, will certainly result in a less parsimonious model with better predictive performance (Steyerberg et al., 2000). Using lasso with the optimal penalty included almost all considered variables in the final model, leading to poorer predictive performance compared to the 1-se penalty. The

reason is that many variables were selected at the 50% threshold, and we need to refit the model with all these selected variables via standard maximum likelihood. A stricter threshold value for the variable inclusion frequency (e.g. 80% or 90%) may result in a more parsimonious model with potentially better performance. With the 1-se penalty, lasso provided a model with good performance and a reasonable balance between the number of selected true and noise variables.

Regarding the impact of the amount of information in the data, backward elimination may fail to converge in data with a low amount of information (Grogan and Elashoff, 2017). This failure is a consequence of a complete or quasi-complete separation pattern in the data (Albert and Anderson, 1984). Complete or quasi-complete separation happens when there is a variable or a combination of variables that could predict the outcome perfectly or to a certain degree, leading to infinite parameter estimates (Albert and Anderson, 1984). Of note, lasso does not have the same problem due to its regularization properties. On the other hand, in data with a larger number of events per variable and a lower amount of missing values, the performance of backward elimination approached that of the lasso methods. Thus, it is a reasonable choice provided that the amount of information is sufficient.

The approach to build models on bootstrap samples has a couple of drawbacks. First, results are sensitive to the chosen cut-off value of the variable inclusion frequency. In our TBM data, there was a clear separation between strong and weak predictors for two of the three considered methods (Table A5.1), which rendered the 50% threshold value sensible. However, in other data settings, this value will not be optimal. Choosing the cut-off value therefore could be considered as a post hoc decision, i.e. after obtaining the inclusion frequency for all variables. Second, the internal validation for these methods requires double bootstrap to repeat the entire modelling process (i.e. first bootstrap in the model validation step, second bootstrap in the variable selection step). This procedure could be computationally prohibitive (page 71, Harrell (2015)).

"Lasso on original MI data" methods are fairly straightforward to use and require less computation. Lasso selection with the optimal penalty tended to include many noise variables, which made it less suitable as a parsimonious variable selection method. Choosing a larger penalty resulted in a more parsimonious model with fewer noise variables. However, parameters will be shrunk too much which may hamper predictive performance. If we refit the model with the selected variables via unpenalized maximum likelihood, we can improve performance, especially in data with a reasonable number of events per variable (EPV of 10 or more). In data with a low

level of information, this refitting may lead to inflated parameter estimates, and potentially a loss in predictive ability. As an alternative, we can recalibrate the model by score, which takes into account the relative predictive strengths of predictors while updating their coefficients. It was observed that this type of recalibration improves performance in data with a low level of information.
We can perform lasso selection on each MI data set separately and combine estimates afterwards. If we simply take the mean of all coefficient estimates, we select all variables that appear in at least one imputed data set. Hence, the number of selected noise variables tends to be large, and increases with the number of imputations. However, I found that in data with a very low level of information, this method tended to give the best predictive performance in terms of the Brier score if coefficients are recalibrated by score. By taking the mean of variables that are selected in at least half of the MI data sets, I attempted to reduce the number of selected noise variables. However, a disadvantage of this method is that it additionally requires choosing a threshold value for the inclusion frequency.
Stacking all MI data sets before analysis provides an interesting alternative that returns a single set of selected predictors and corresponding coefficient estimates. Therefore, the stacked method avoids making any arbitrary decision about the selection threshold (Vergouwe et al., 2010).Wan et al. (2015) suggested that it is better to weight individuals by the amount of missing values. In their simulation study, it improved mean squared error compared to stacking without using weights. In my simulation, using such weights tended to result in a slightly more parsimonious model in general. Yet it did not appear to lead to a uniformly better model performance. In fact, after recalibrating the model by selected variables, the two methods gave very similar performance in the simulation study as well as in the real data application.

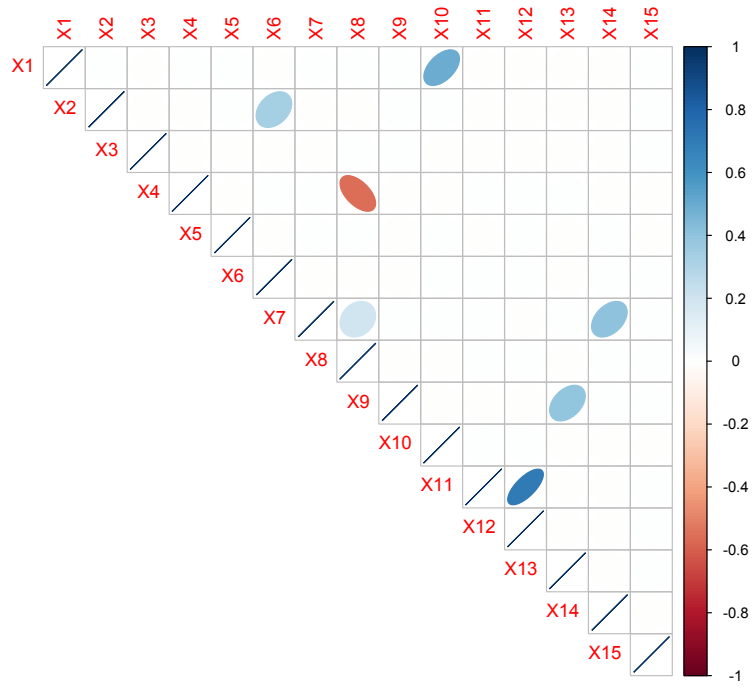
My study has a few limitations. First, only a couple of methods were considered as I aimed at providing a comparison of pragmatic approaches that are frequently used by most data analysts. Second, I only considered two simulation settings for the proportion of noise variables and three values for the number of events per variables for each simulation setting. However, I think that comparable settings are often encountered in medical research. Third, I only considered one scenario for the parameter values. Other possible choices may have an impact on the predictive performance especially in data with the EPV of 3. Finally, the TBM data has a relatively low amount of missingness. Yet, the data still confirms the relative difference in predictive performance that was found in the simulation study.

In conclusion, I gave an overview of a couple of strategies for variable selection with multiply imputed data in the development of a prognostic model. By means of a simulation study, I provided insights into their strengths and weaknesses. None of the methods that I considered is new, but to my knowledge, this paper is one of the few studies on how to combine results from multiply imputed data sets in a penalized likelihood framework. Also, I do not know of any other study that systematically compares different recalibration methods if a stronger penalty than the optimal one is chosen. With the increasing use of MI to handle missing data, and a growing number of prognostic models being developed, further research in this direction is urgently needed.

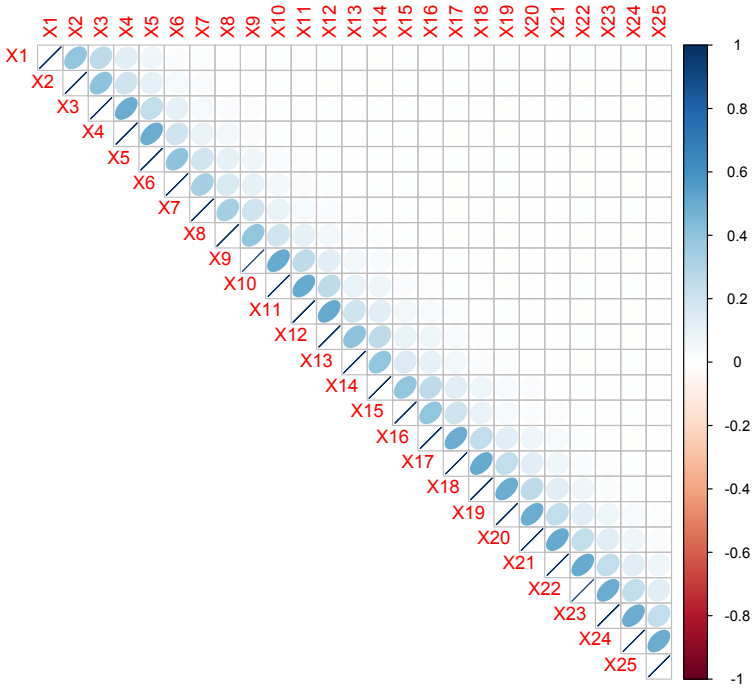
5.6 Appendix

Table A5.1: *Variable inclusion frequency (%) for the model selection on bootstrap data methods in the TBM data using logistic regression*

Variables	**Method**		
	BBeF	**BLaF**	**BLaFo**
Age	100	100	100
MRC grade	98.7	100	100
CSF lymphocyte count	96.3	100	100
Presence of focal neurological signs	59.2	93.0	100
Cohort	70.5	86.5	99.7
Received dexamethasone treatment	21.1	60.9	84.5
Previous TB	25.4	67.4	92.5
Weight	41.6	69.7	97.2
Ratio of CSF to blood glucose	14.9	25.5	76.2
Illness duration at entry	10.7	38.5	83.3
CSF protein	6.5	7.3	60.8
Occurrence of seizures	1.6	4.9	64.1
Plasma sodium	3.6	4.8	61.1
Miliary tuberculosis present on chest radiograph	2.6	3.9	55.8
Sex	2.1	3.4	54.1
CSF glucose	5.3	1.2	42.2
Body temperature	4.0	1.9	49.0



(a) *First data generating mechanism*



(b) *Second data generating mechanism*

Figure A5.1: *Correlation matrix of the covariates*

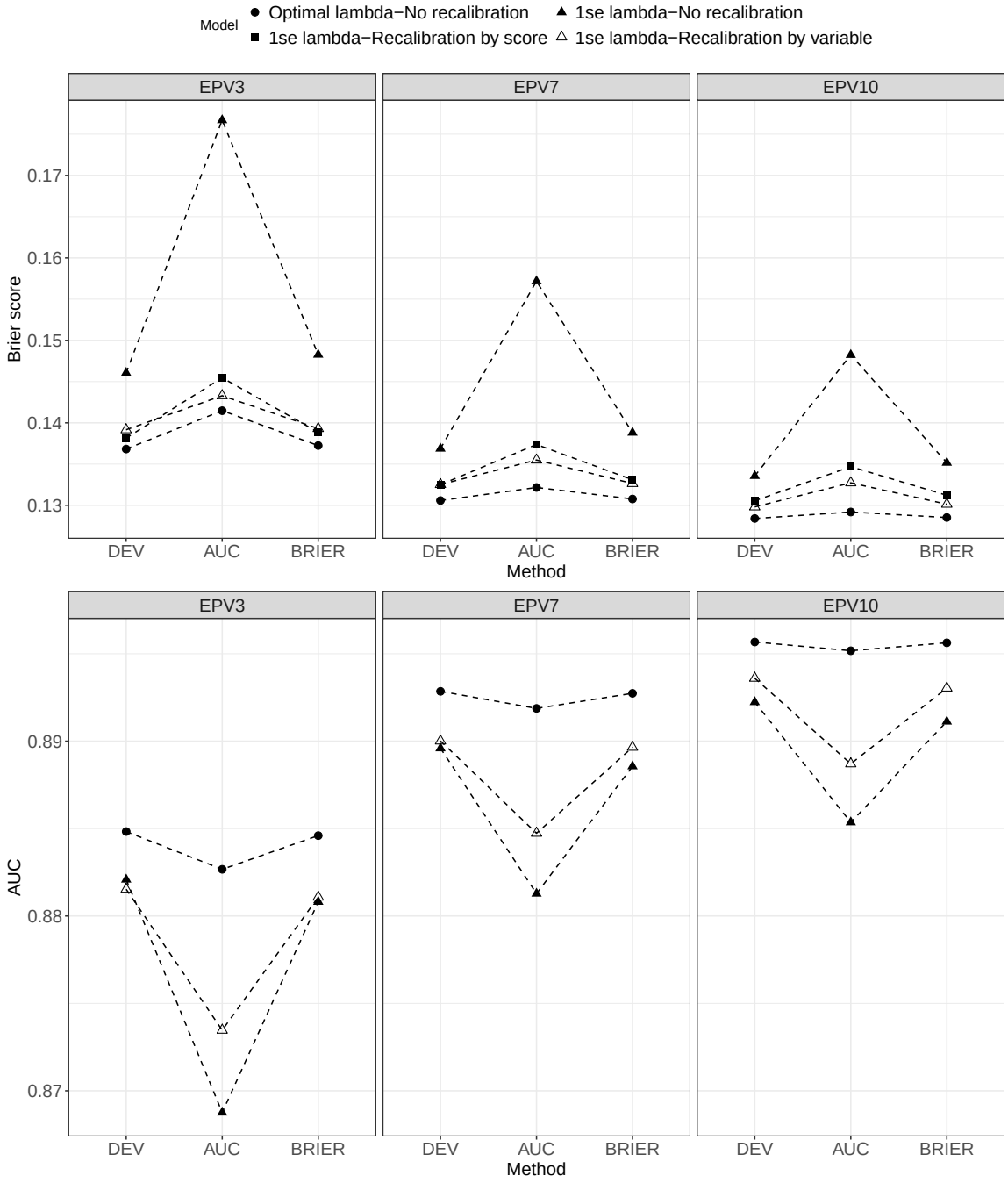
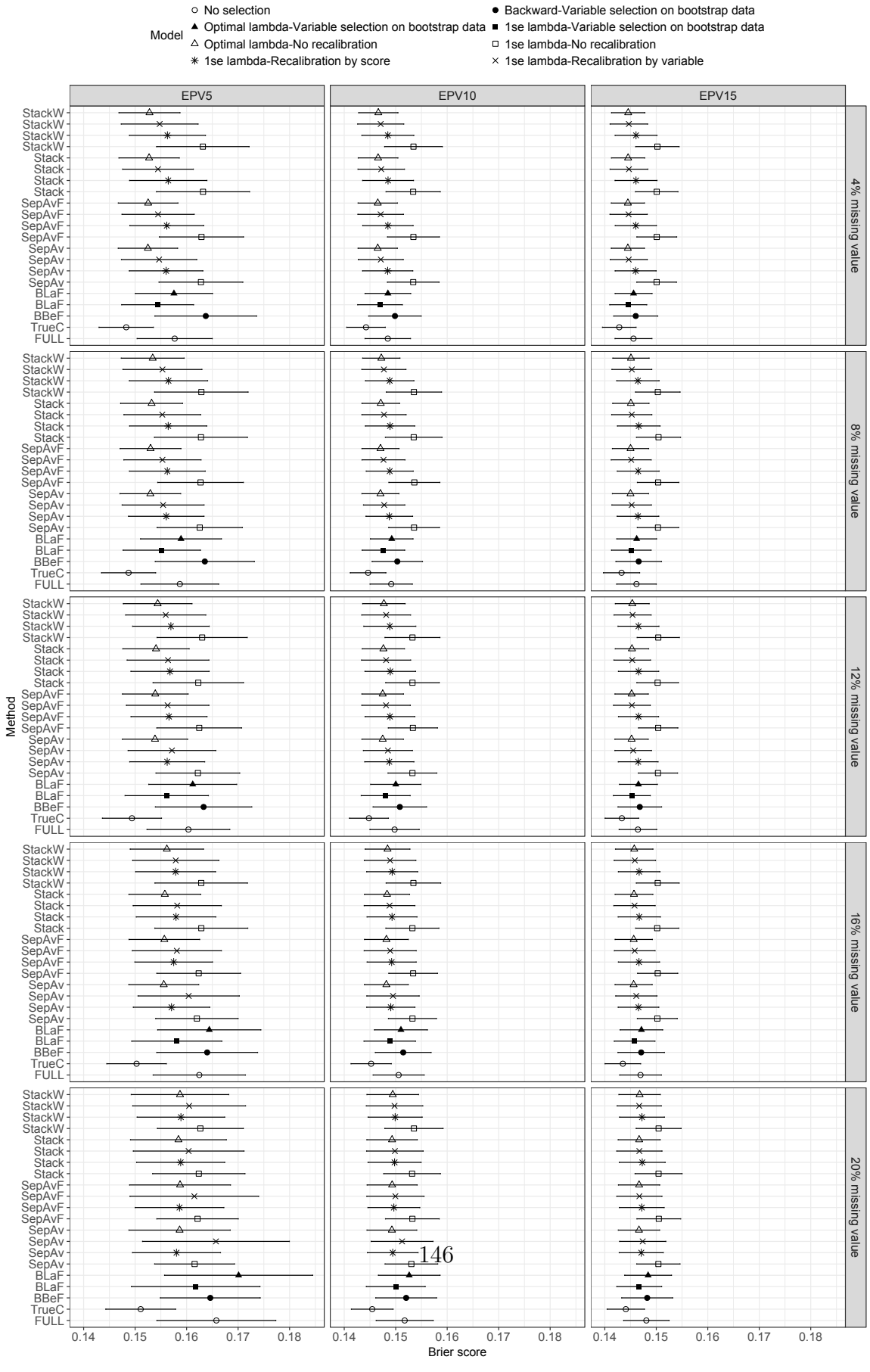
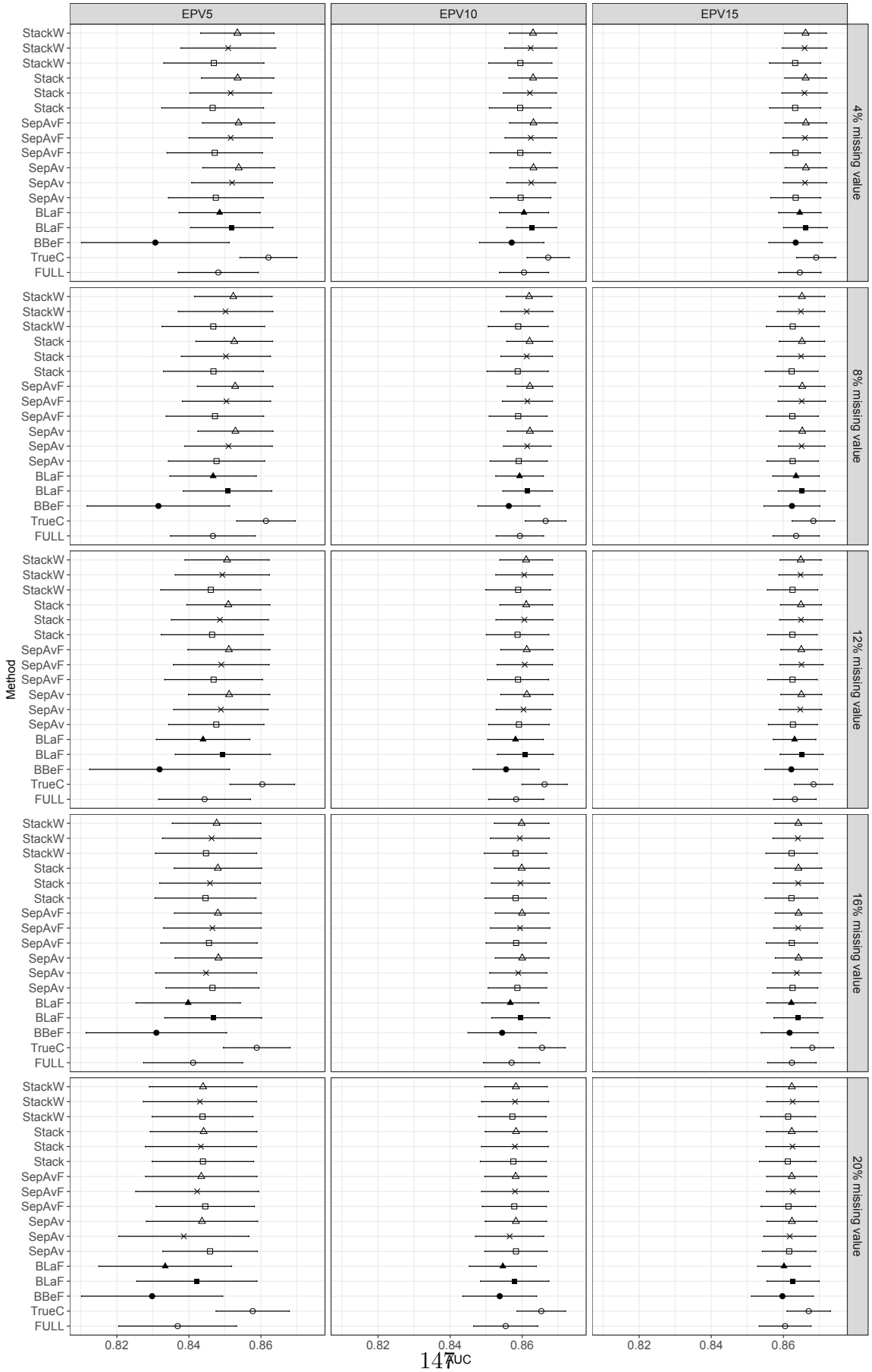


Figure A5.2: *Fully observed data - second data generating mechanism (section 5.3.1.b): Mean AUC and Brier score over 500 fully observed generated data sets for lasso models with three loss function for the selection of the penalty. DEV: deviance; BRIER: brier score; AUC: AUC as loss function.*

Figure A5.3: *First data generating mechanism: Mean Brier score and AUC over 500 generated data sets for all scenarios with 10 imputed data sets. Black bars represent the empirical standard deviation*




EPV5
EPV10
EPV15
4% missing value
8% missing value
12% missing value
16% missing value
20% missing value
Method
StackW
Stack
SepAvF
SepAv
BLaF
BBeF
TrueC
FULL
0.82
0.84
0.86
AUC

(b) AUC

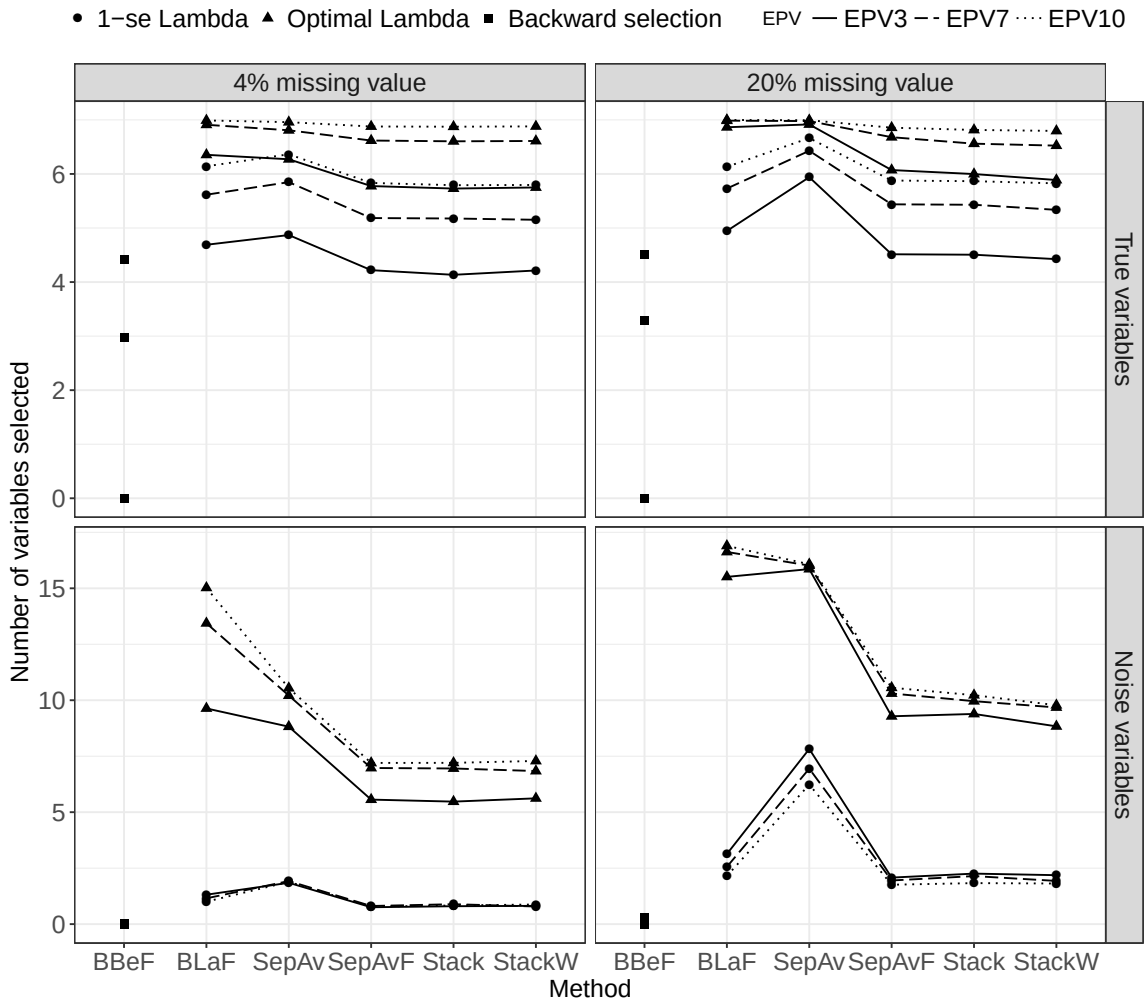
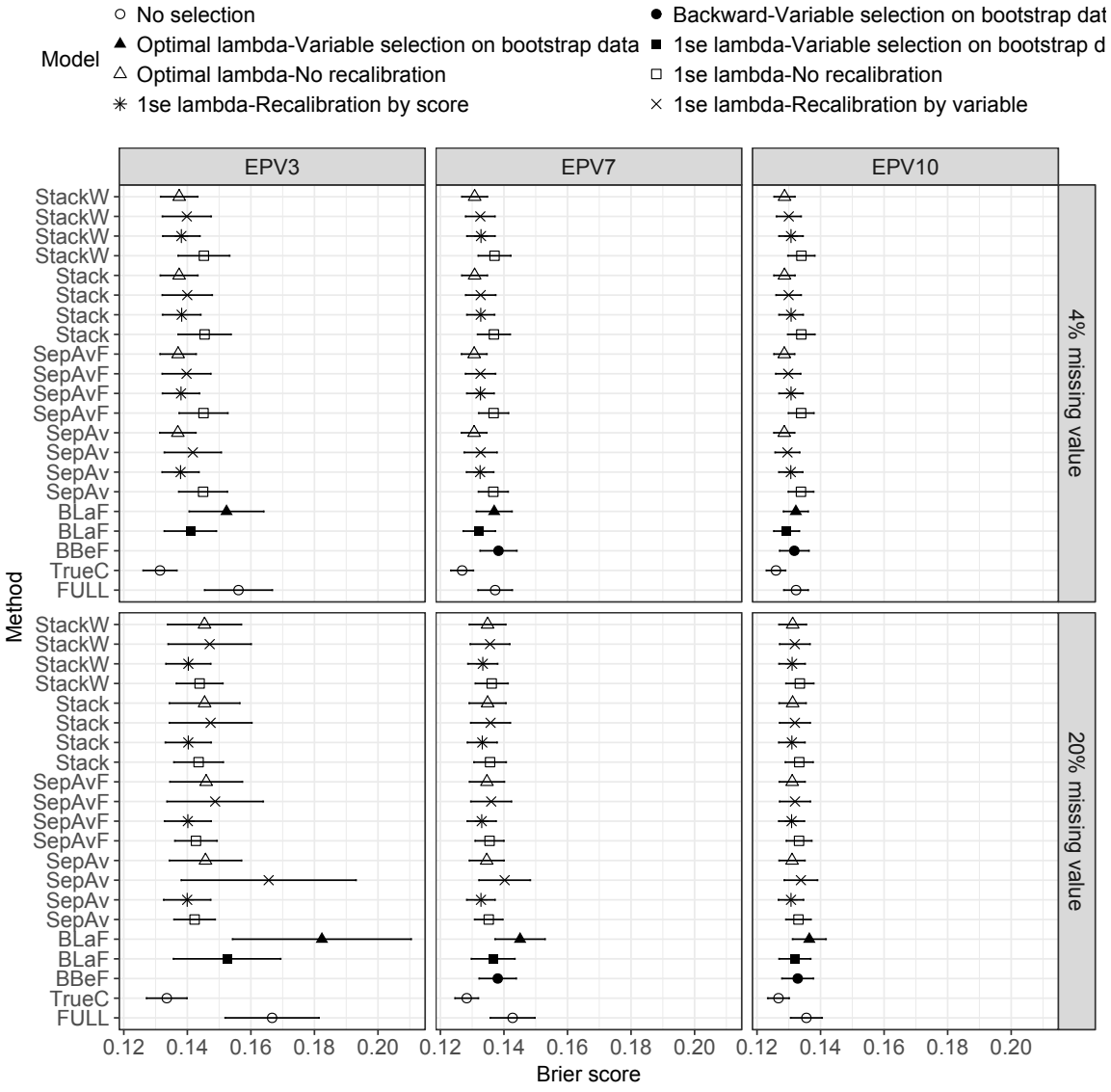
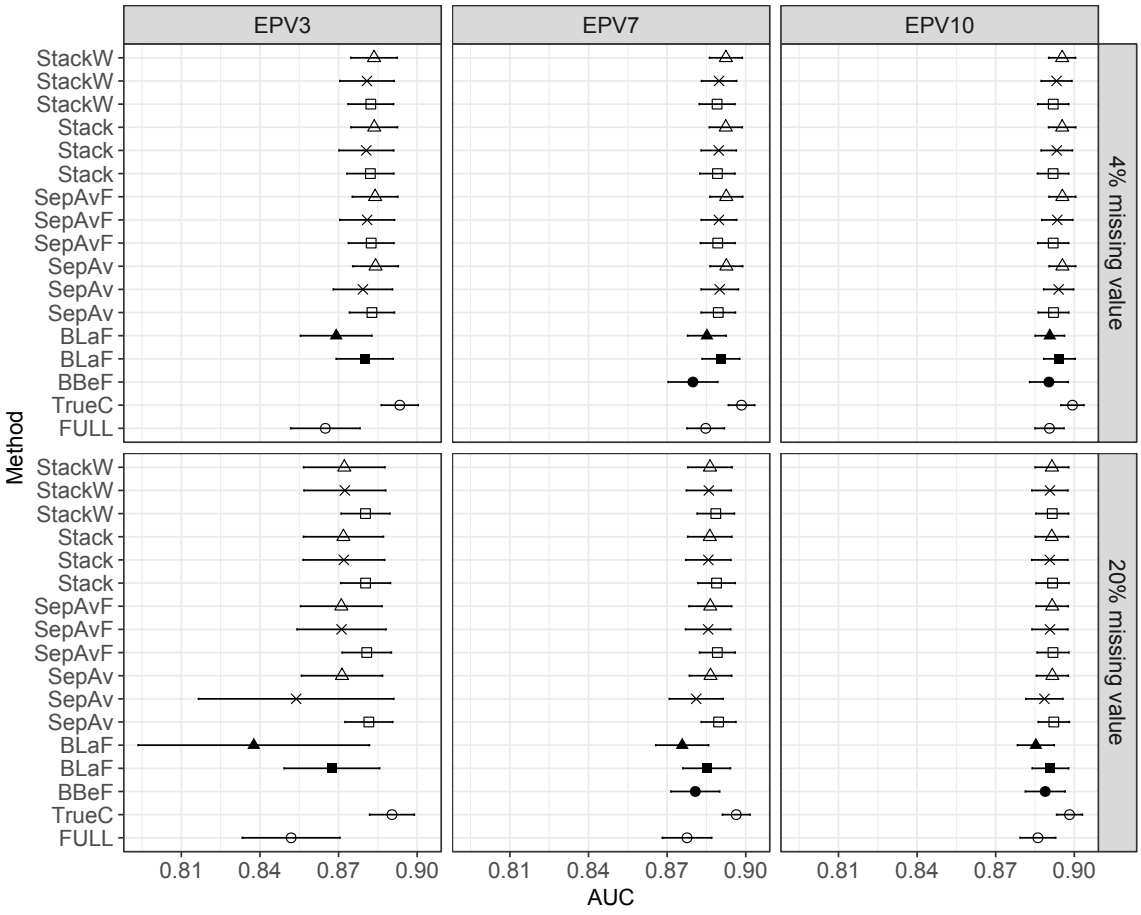


Figure A5.4: *Second data generating mechanism: Number of selected variables for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. The values for BBeF in data with EPV of 3 are not available due to convergence problems in many generated data sets.*



(a) *Brier score*



(b) *AUC*

Figure A5.5: *Second data generating mechanism: Mean Brier score (top figure) and AUC (bottom figure) over 500 generated data sets for the two most extreme cases of missing values (4% and 20%) with 10 imputed data sets. Black bars represent the empirical standard deviation.*

Chapter 6

Conclusion

Tuberculous meningitis (TBM) is one of the most common causes of life-threatening meningitis worldwide, a distinction driven by the continued ability of *Mycobacterium tuberculosis* to evade global control efforts, the strong link between human immunodeficiency virus (HIV) infection and TBM, and the success of vaccination programmes against other bacterial causes of meningitis. Despite proper anti-tuberculosis treatment and innovation in advanced diagnostic techniques, mortality continues to be as high as 30-75% (Bourgi et al., 2017; Heemskerk et al., 2016), and severe neurological sequelae are still common amongst survivors (Wilkinson et al., 2017).

In this DPhil thesis, I developed prognostic models for mortality in TBM based on baseline and dynamic time-updated clinical and laboratory measurements. I also described and compared two approaches of model selection in the presence of multiply imputed data. I first discuss the clinical and statistical contributions of the thesis, then propose several avenues for future research, and end this chapter by some final remarks.

6.1 Contributions of the thesis

6.1.1 Clinical contributions

The larger part of the thesis focused on the development of a prognostic model for mortality in TBM, using one of the largest databases on TBM worldwide. I developed and validated prognostic models based on baseline patient characteristics (obtained

at diagnosis; Chapter 3) and time-updated marker values (measured during the disease course; Chapter 4). These models are amongst the first prognostic models that have been presented in the literature. The main focus of the thesis was to develop prognostic models to enable more accurate predictions. I used fairly standard regression models. Therefore, the models also provide insights into the underlying etiology of the disease. They are in contrast with modern machine learning models, such as random forests. In these black box models, model interpretation is difficult because the relationships between predictors and outcome are hidden in the model structure (Palczewska et al., 2014).

The models presented in Chapter 3 identified and combined baseline risk factors for 9-month mortality in patients with and without HIV co-infection. They were carefully developed, and had good predictive performance in internal validation. The included predictors are easy to measure, and are available at the time of TBM diagnosis in almost all clinical settings. To facilitate the implementation of the models in practice, I developed a web-based calculator which can be accessed from anywhere with Internet connection. The active hour of the application is more than 25 hours per month since it was deployed. The baseline models can be used to identify patients at high risk early. This enables clinicians to put them under closer monitoring and promptly intervene and manage them. The interventions may include temperature regulation, assisted ventilation, complications management and fluid and electrolyte management to monitor and control hyponatremia. One can also use the models to recruit patients in clinical trials where treatment is focused on a certain prognostic risk group.

Prediction models that are based on information at diagnosis are most straightforward to apply in clinical practice. However, baseline information tends to become less relevant as the disease progresses and, therefore, initial predictions become less accurate. The initial prognosis can be updated and improved by using time-updated markers as predictors. Indeed, the dynamic prediction models presented in Chapter 4 illustrated the benefit of the longitudinal Glasgow Coma Score (GCS) and plasma sodium in predicting death within 120 days after the initiation of anti-tuberculous treatment. The dynamic model allows updating of the predicted risk for a patient whenever new assessments of the two markers become available. I also created a web-based calculator to assist the application of the dynamic model in practice. The dynamic model provides an attractive tool for clinicians in the era of personalized medicine. Clinicians can use it as a prognostic tool to assist medical decision-making for each individual patient at a particular time point. They could also use it to gain insights into the underlying disease dynamic. For example, by monitoring the value

of plasma sodium and GCS and their associations with survival, clinicians can learn more about their role in disease progression.

6.1.2 Statistical contributions

Apart from the application in clinical practice, the prediction models presented in Chapter 3 and 4 are valuable case studies that illustrate the development and validation of prognostic models for a specific disease. Although guidelines for building and validating a prediction model are well established (Harrell, 2015; Steyerberg, 2008; Collins et al., 2015), applying these recommendations to a particular disease area and for a concrete data set still requires specific considerations. This includes the selection of candidate predictors, methods for handling missing data, methods for performing variable selection, methods for internal validation especially in the presence of multiply imputed data, and the choice between landmarking, joint modelling or another approach for dynamic prediction. For example, in Chapter 4, the landmarking approach showed better performance on the TBM data set than the more complicated joint modelling approach. This illustrates that a more complicated modelling method is not always preferable over a simpler one, and a correct specification for the longitudinal process is essential in a joint model.

Despite the increasing number of prognostic models and the popularity of multiple imputation techniques, there is still little guidance on how to perform variable selection for multiply imputed data (Zhao and Long, 2017). The simulation presented in Chapter 5 is one of the few studies that provide a systematic comparison for a number of pragmatic strategies in the prediction modelling context. The study showed that neither of the considered approaches outperforms the others in all settings. The choice of the most suitable method depends on the amount of information in the data, and whether parsimony or predictive performance is the most important criterion in building a prediction model. Overall, applying lasso selection with the 1-se penalty shows the best performance. When a stronger penalty than the optimal one is chosen in a lasso model, it is important to recalibrate the obtained model. When the amount of information is sufficient (the number of events per variable $\geq$10 and low amount of missingness), all considered approaches gave analogous performance. Performing lasso selection on the stacked multiply imputed (MI) data is a new and attractive method. The stacked approach enjoys the advantage of analysing only one single data set. Hence, it does not require a decision on how to combine selected variables and estimates from the individual MI data sets.

6.2 Future research

The prognostic models presented in Chapter 3 and 4 showed good predictive performance in internal validation. Their performance could be improved even further by considering predictors that influence survival but were not available for all cohorts in our current pooled database. Specifically, with the baseline models, one could consider inflammatory cytokines (Thuong et al., 2017; Sharma et al., 2017), host genetic factors such as LTA4H (Thuong et al., 2017), or bacterial factors such as drug resistance or bacterial lineage (Tho et al., 2012). Resistance information is currently not available at baseline. However, I hope that novel rapid diagnostic tools for TBM in the future would make it feasible. With the dynamic models, one could investigate the value of adding other longitudinal markers, such as inflammatory cytokines, or clinical markers of inflammation such as CSF lymphocytes, neutrophils, lactate, protein, opening pressure, or the CSF-plasma glucose ratio. These markers have been shown to provide useful information in the monitoring of disease progression and treatment response (Simmons et al., 2005). Of note, the inclusion of multiple longitudinal markers would pose a computational challenge for the joint modeling method because of the increase in dimension of the random effects component (Fieuws and Verbeke, 2004; Hickey et al., 2016). This extension is less of a problem for the landmarking approach since we can model each longitudinal marker separately and use the fitted value at a particular landmark time in the Cox model. Thus, the difference in computational performance between the two methods is expected to be more pronounced in this case.

Validation is an important part in prognostic research. The presented prediction models were developed using data solely from two recruiting hospitals in the South of Vietnam. To prove their usefulness in clinical practice, it is crucial to validate their performance on new patients, preferably from a different population. As mentioned in Chapter 3, the baseline model was validated on an independent Indonesian cohort, and unfortunately, it did not perform well (Dian et al., 2018). The poor performance could be explained by the discrepancy in pathophysiological mechanism across TBM populations. One of the possible driving factors for the difference is the genomic variations in the bacteria and their hosts (Thao et al., 2018a). Thus, more validation research on different populations is desirable to identify potential factors that account for the heterogeneity between populations. When the models perform poorly in a validation study, they could be updated and adjusted by recalibrating or adding newly discovered predictors (Steyerberg et al., 2004; Collins et al., 2015). Of note,

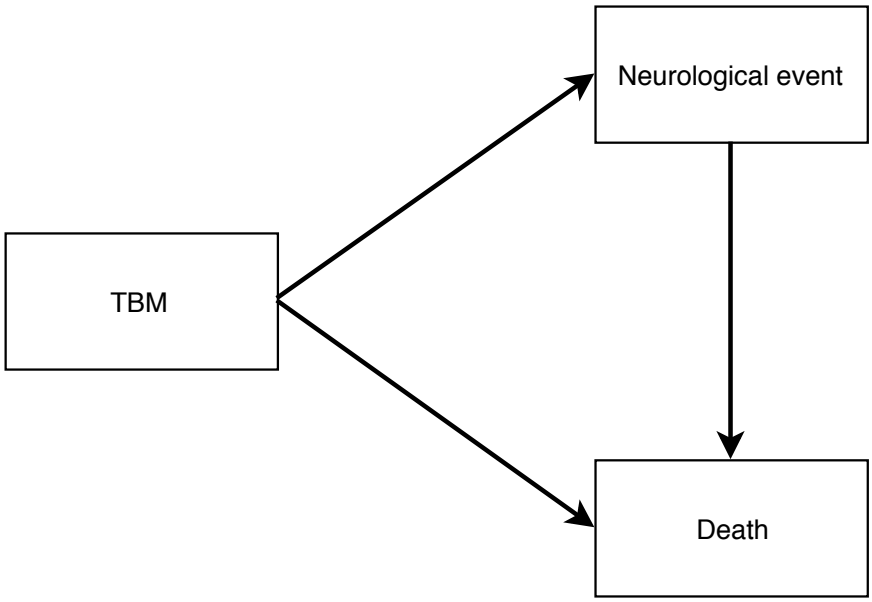


Figure 6.1: *Illness-death model in TBM data.*

a validation study often needs a smaller sample size compared with a model development study because overfitting is no longer a concern. Currently, two randomized clinical trials of adjunctive dexamethasone (Donovan et al., 2018b,a) conducted by OUCRU-VN are collecting data on a large number of TBM patients with (n = 520) and without (n = 720) HIV infection. Once the recruitment is finished, this large dataset could be used to validate and improve the developed models. Of note, most of the trial participants are Vietnamese patients admitted to the two recruiting hospitals: Pham Ngoc Thach hospital and the Hospital for Tropical Diseases. In order to advocate validation studies in different locations, it is important to standardize the data collecting system across institutes and regions. This aligned and standardized data collection will enhance study comparison, generalizability of results and scientific collaboration (Marais et al., 2016). Furthermore, the current web-based calculators can only provide risk predictions. They could be improved to collect data from users to enhance the prediction models.

The prognostic value of the time-updated plasma sodium and GCS were identified in Chapter 4. In the next step, one could investigate the value of incorporating these markers for poor outcomes in TBM trials. Specifically, whether they could act as surrogate outcomes for survival or neurological events. Such outcomes could facilitate trials with shorter period and smaller sample size (Tsiatis et al., 1995; Wilkinson et al., 2017).

Apart from landmarking and joint modelling approaches, a multi-state model is

an alternative approach for prediction. Such model can be useful when the disease process can be described as transition between different states (Meira-Machado et al., 2009). In TBM, a neurological event is also an outcome of interest. Patients might die with or without experiencing a prior neurological event and the occurrence of a neurological event could impact survival in TBM patients. Hence, it would be interesting to investigate whether a prognostic model for mortality could be updated by taking into account the occurrence of a neurological event. The progress of neurological events and survival can be represented via an illness death model (van Houwelingen and Putter, 2011) (i.e. a multistate model with three states (Figure 6.1)). In addition, the model allows estimating the risk of having a neurological event, death prior to and after a neurological event, and exploring potential baseline risk factors for each disease transition.

In conclusion, the developed models allow for a more accurate early identification of patients at highest risk of death, and their implementation in clinical practice could improve patient outcomes. In addition, the thesis contributes to the statistical aspects of prognostic research by providing insights into the strengths and weaknesses of several strategies for variable selection with multiply imputed data.

References

Albert, A. and Anderson, J. A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984. ISSN 0006-3444. doi: 10.1093/biomet/71.1.1. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/71.1.1.

Alboukadel Kassambara and Marcin Kosinski. survminer: Drawing Survival Curves using 'ggplot2', 2018. URL https://cran.r-project.org/package=survminer.

Anderson, N. E., Somaratne, J., Mason, D. F., Holland, D., and Thomas, M. G. A review of tuberculous meningitis at Auckland City Hospital, New Zealand. *Journal of Clinical Neuroscience*, 17(8):1018–1022, 2010. ISSN 09675868. doi: 10.1016/j.jocn.2010.01.007. URL http://dx.doi.org/10.1016/j.jocn.2010.01.007.

Ardito, Fausta, Posteraro, Brunella, Sanguinetti, Maurizio, Zanetti, Stefania, and Fadda, Giovanni. Evaluation of BACTEC Mycobacteria Growth Indicator Tube (MGIT 960) Automated System for Drug Susceptibility Testing of Mycobacterium tuberculosis. *Journal of Clinical Microbiology*, 39(12):4440–4444, dec 2001. ISSN 0095-1137. doi: 10.1128/JCM.39.12.4440-4444.2001. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC88562/.

Austin, Peter C. and Steyerberg, Ewout W. The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, 68(6):627–636, 2015. ISSN 18785921. doi: 10.1016/j.jclinepi.2014.12.014. URL http://dx.doi.org/10.1016/j.jclinepi.2014.12.014.

Austin, Peter C and Steyerberg, Ewout W. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research*, 26(2):796–808, 2017. ISSN 0962-2802. doi: 10.1177/0962280214558972. URL http://journals.sagepub.com/doi/10.1177/0962280214558972.

Austin, Peter C and Tu, Jack V. Bootstrap Methods for Developing Predictive Models. *The American Statistician*, 58(2):131–137, may 2004. ISSN 0003-1305. doi: 10.1198/0003130043277. URL http://dx.doi.org/10.1198/0003130043277.

Blanche, Paul, Latouche, Aurélien, and Viallon, Vivian. Time-dependent AUC with right-censored data: a survey study. oct 2012. URL http://arxiv.org/abs/1210.6805.

Blanche, Paul, Dartigues, Jean-François, and Jacqmin-Gadda, Hélène. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5):687–704, sep 2013a. ISSN 03233847. doi: 10.1002/bimj.201200045. URL http://doi.wiley.com/10.1002/bimj.201200045.

Blanche, Paul, Dartigues, Jean-Francois, and Jacqmin-Gadda, Helene. Estimating and Comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, 2013b. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.5958/full.

Blanche, Paul, Latouche, Aurélien, and Viallon, Vivian. Time-Dependent AUC with Right-Censored Data: A Survey. In Lee, Mei-Ling Ting, Gail, Mitchell, Pfeiffer, Ruth, Satten, Glen, Cai, Tianxi, and Gandy, Axel, editors, *Risk Assessment and Evaluation of Predictions*, pages 239–251. Springer New York, New York, NY, 2013c. ISBN 978-1-4614-8981-8. doi: 10.1007/978-1-4614-8981-8_11. URL http://link.springer.com/10.1007/978-1-4614-8981-8{_}11.

Bourgi, Kassem, Fiske, Christina, and Sterling, Timothy R. Tuberculosis Meningitis. *Current Infectious Disease Reports*, 19(11):4–12, 2017. ISSN 15343146. doi: 10.1007/s11908-017-0595-4.

Bouwmeester, Walter, Zuithoff, Nicolaas P. A., Mallett, Susan, Geerlings, Mirjam I., Vergouwe, Yvonne, Steyerberg, Ewout W., Altman, Douglas G., and Moons, Karel G. M. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Medicine*, 9(5):e1001221, may 2012. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001221. URL https://dx.plos.org/10.1371/journal.pmed.1001221.

Chen, Qixuan and Wang, Sijian. Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, 32(21):3646–3659, 2013. ISSN 02776715. doi: 10.1002/sim.5783.

Choudhury, A R. Non-surgical treatment of tuberculomas of the brain. *British journal of neurosurgery*, 3(6):643–53, 1989. ISSN 0268-8697. URL http://www.ncbi.nlm.nih.gov/pubmed/2627283.

Collins, Gary S., Reitsma, Johannes B., Altman, Douglas G., and Moons, Karel G.M. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Annals of internal medicine*, 162(1):55–63, 2015. ISSN 1539-3704. doi: 10.7326/M14-0697. URL http://www.ncbi.nlm.nih.gov/pubmed/25560714.

Courvoisier, Delphine S., Combescure, Christophe, Agoritsas, Thomas, Gayet-Ageron, Angle, and Perneger, Thomas V. Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*, 64(9):993–1000, 2011. ISSN 08954356. doi: 10.1016/j.jclinepi.2010.11.012.

Croda, Mariana G, Vidal, José E, Hernández, Adrián V, Dal Molin, Tiago, Gualberto, Felipe A, and de Oliveira, Augusto C Penalva. Tuberculous meningitis in HIV-infected patients in Brazil: clinical and laboratory characteristics and factors associated with mortality. *International Journal of Infectious Diseases*, 14(7): e586—-e591, 2010.

Dass, Rashna, Nagaraj, Ravishankar, Murlidharan, Jayashree, and Singhi, Sunit. Hyponatraemia and hypovolemic shock with tuberculous meningitis. *Indian journal of pediatrics*, 70(12):995–7, dec 2003. ISSN 0019-5456. URL http://www.ncbi.nlm.nih.gov/pubmed/14719791.

Davis, Angharad, Meintjes, Graeme, and Wilkinson, Robert J. Treatment of Tuberculous Meningitis and Its Complications in Adults. *Current treatment options in neurology*, 20(3):5, feb 2018. ISSN 1092-8480. doi: 10.1007/s11940-018-0490-9. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5830467.

Dian, Sofiati, Rahmadi, Ridho, van Laarhoven, Arjan, Ganiem, Ahmad Rizal, and van Crevel, Reinout. Predicting Mortality of Tuberculous Meningitis. *Clinical Infectious*

Diseases, may 2018. ISSN 1058-4838. doi: 10.1093/cid/ciy445. URL https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciy445/5025876.

Donovan, Joseph, Phu, Nguyen Hoan, Mai, Nguyen Thi Hoang, Dung, Le Tien, Imran, Darma, Burhan, Erlina, Ngoc, Lam Hong Bao, Bang, Nguyen Duc, Giang, Do Chau, Ha, Dang Thi Minh, Day, Jeremy, Thao, Le Thi Phuong, Thuong, Nguyen TT, Vien, Nguyen Nang, Geskus, Ronald B., Wolbers, Marcel, Hamers, Raph L, van Crevel, Reinout, Nursaya, Mugi, Maharani, Kartika, Hien, Tran Tinh, Baird, Kevin, Lan, Nguyen Huu, Kestelyn, Evelyne, Chau, Nguyen Van Vinh, and Thwaites, Guy E. Adjunctive dexamethasone for the treatment of HIV-infected adults with tuberculous meningitis (ACT HIV): Study protocol for a randomised controlled trial. *Wellcome Open Research*, 3:31, jun 2018a. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.14006.2. URL https://wellcomeopenresearch.org/articles/3-31/v2.

Donovan, Joseph, Phu, Nguyen Hoan, Thao, Le Thi Phuong, Lan, Nguyen Huu, Mai, Nguyen Thi Hoang, Trang, Nguyen Thi Mai, Hiep, Nguyen Thi Thu, Nhu, Tran Bao, Hanh, Bui Thi Bich, Mai, Vu Thi Phuong, Bang, Nguyen Duc, Giang, Do Chau, Ha, Dang Thi Minh, Day, Jeremy, Thuong, Nguyen TT, Vien, Nguyen Nang, Geskus, Ronald B., Hien, Tran Tinh, Kestelyn, Evelyne, Wolbers, Marcel, Chau, Nguyen Van Vinh, and Thwaites, Guy E. Adjunctive dexamethasone for the treatment of HIV-uninfected adults with tuberculous meningitis stratified by Leukotriene A4 hydrolase genotype (LAST ACT): Study protocol for a randomised double blind placebo controlled non-inferiority trial. *Wellcome Open Research*, 3: 32, mar 2018b. ISSN 2398-502X. doi: 10.12688/wellcomeopenres.14007.1. URL https://wellcomeopenresearch.org/articles/3-32/v1.

Efsen, Anne Marie W, Panteleev, Alexander M, Grint, Daniel, Podlekareva, Daria N, Vassilenko, Anna, Rakhmanova, Aza, Zeltina, Indra, Losso, Marcelo H, Miller, Robert F, Girardi, Enrico, Caylá, Joan, Post, Frank A, Miro, Jose M, Bruyand, Mathias, Furrer, Hansjakob, Obel, Niels, Lundgren, Jens D, Mocroft, Amanda, Kirk, Ole, and Study Group, HIV/TB. TB Meningitis in HIV-Positive Patients in Europe and Argentina: Clinical Outcome and Factors Associated with Mortality. *BioMed Research International*, 2013:1–9, 2013. ISSN 2314-6133. doi: 10.1155/2013/373601. URL http://www.ncbi.nlm.nih.gov/pubmed/24699884http://www.hindawi.com/journals/bmri/2013/373601/.

Elashoff, Robert M, Li, Gang, and Li, Ning. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64 (3):762–771, 2008. ISSN 0006-341X.

Erdem, Hakan, Ozturk-Engin, Derya, Tireli, Hulya, Kilicoglu, Gamze, Defres, Sylviane, Gulsun, Serda, Sengoz, Gonul, Crisan, Alexandru, Johansen, Isik Somuncu, Inan, Asuman, Nechifor, Mihai, Al-Mahdawi, Akram, Civljak, Rok, Ozguler, Muge, Savic, Branislava, Ceran, Nurgul, Cacopardo, Bruno, Inal, Ayse Seza, Namiduru, Mustafa, Dayan, Saim, Kayabas, Uner, Parlak, Emine, Khalifa, Ahmad, Kursun, Ebru, Sipahi, Oguz Resat, Yemisen, Mucahit, Akbulut, Ayhan, Bitirgen, Mehmet, Popovic, Natasa, Kandemir, Bahar, Luca, Catalina, Parlak, Mehmet, Stahl, Jean Paul, Pehlivanoglu, Filiz, Simeon, Soline, Ulu-Kilic, Aysegul, Yasar, Kadriye, Yilmaz, Gulden, Yilmaz, Emel, Beovic, Bojana, Catroux, Melanie, Lakatos, Botond, Sunbul, Mustafa, Oncul, Oral, Alabay, Selma, Sahin-Horasan, Elif, Kose, Sukran, Shehata, Ghaydaa, Andre, Katell, Dragovac, Gorana, Gul, Hanefi Cem, Karakas, Ahmet, Chadapaud, Stéphane, Hansmann, Yves, Harxhi, Arjan, Kirova, Valerija, Masse-Chabredier, Isabelle, Oncu, Serkan, Sener, Alper, Tekin, Recep, Elaldi, Nazif, Deveci, Ozcan, Ozkaya, Hacer Deniz, Karabay, Oguz, Senbayrak, Seniha, Agalar, Canan, and Vahaboglu, Haluk. Hamsi scoring in the prediction of unfavorable outcomes from tuberculous meningitis: results of Haydarpasa-II study. *Journal of neurology*, 262(4):890–8, jan 2015. ISSN 1432-1459. doi: 10.1007/s00415-015-7651-5. URL `http://www.ncbi.nlm.nih.gov/pubmed/25634680`.

Ferrer, Loïc, Putter, Hein, and Proust-Lima, Cécile. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. jul 2017. URL `http://arxiv.org/abs/1707.03706`.

Fieuws, Steffen and Verbeke, Geert. Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Statistics in medicine*, 23(20): 3093–104, oct 2004. ISSN 0277-6715. doi: 10.1002/sim.1885. URL `http://www.ncbi.nlm.nih.gov/pubmed/15449333`.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

Garfunkel, Lynn C, Kaczorowski, Jeffrey, and Christy, Cynthia. *Pediatric Clinical Advisor E-Book: Instant Diagnosis and Treatment*. Elsevier Health Sciences, 2007. ISBN 0323070582.

Garg, R K. Classic diseases revisited: Tuberculosis of the central nervous system. *Postgraduate Medical Journal*, 75(881):133–140, mar 1999. doi: 10.1136/pgmj.75.881.133. URL http://dx.doi.org/10.1136/pgmj.75.881.133.

George, Elizabeth Litta, Iype, Thomas, Cherian, Ajith, Chandy, Sinchu, Kumar, Aswini, Balakrishnan, Ajitha, and Vijayakumar, K. Predictors of mortality in patients with meningeal tuberculosis. *Neurology India*, 60(1):18–22, 2012. ISSN 0028-3886. doi: 10.4103/0028-3886.93583. URL http://www.ncbi.nlm.nih.gov/pubmed/22406774.

Gerds, Thomas A. and Schumacher, Martin. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, 48(6):1029–1040, dec 2006. ISSN 03233847. doi: 10.1002/bimj.200610301. URL http://doi.wiley.com/10.1002/bimj.200610301.

Gerds, Thomas A., Cai, Tianxi, and Schumacher, Martin. The Performance of Risk Prediction Models. *Biometrical Journal*, 50(4):457–479, aug 2008. ISSN 03233847. doi: 10.1002/bimj.200810443. URL http://doi.wiley.com/10.1002/bimj.200810443.

Graf, Erika, Schmoor, Claudia, Sauerbrei, Willi, and Schumacher, Martin. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–45, sep 1999. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2529::AID-SIM274⟩3.0.CO;2-5. URL http://www.ncbi.nlm.nih.gov/pubmed/10474158.

Groenwold, Rolf H.H., Moons, Karel G.M., Pajouheshnia, Romin, Altman, Doug G., Collins, Gary S., Debray, Thomas P.A., Reitsma, Johannes B., Riley, Richard D., and Peelen, Linda M. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *Journal of Clinical Epidemiology*, 78:90–100, oct 2016. ISSN 08954356. doi: 10.1016/j.jclinepi.2016.03.017. URL https://linkinghub.elsevier.com/retrieve/pii/S0895435616300300.

Grogan, Tristan R. and Elashoff, David A. A simulation based method for assessing the statistical significance of logistic regression models after common variable

selection procedures. *Communications in Statistics: Simulation and Computation*, 0918(May):1–14, 2017. ISSN 15324141. doi: 10.1080/03610918.2016.1230216. URL https://doi.org/10.1080/03610918.2016.1230216.

Gu, Jin, Xiao, Heping, Wu, Furong, Ge, Yanping, Ma, Jun, and Sun, Wenwen. Prognostic factors of tuberculous meningitis: a single-center study. *International journal of clinical and experimental medicine*, 8(3):4487–93, 2015. ISSN 1940-5901. URL http://www.ncbi.nlm.nih.gov/pubmed/26064373.

Guedj, Jeremie, Thiébaut, Rodolphe, and Commenges, Daniel. Joint modeling of the clinical progression and of the biomarkers' dynamics using a mechanistic model. *Biometrics*, 67(1):59–66, 2011. ISSN 0006-341X.

Harrell, Frank E. *Regression Modeling Strategies*. Springer Series in Statistics. Springer New York, New York, NY, 2015. ISBN 9781475734621. doi: 10.1007/978-1-4757-3462-1. URL http://link.springer.com/10.1007/978-3-319-19425-7.

Harrell Jr, Frank E. rms: Regression Modeling Strategies, 2016. URL https://cran.r-project.org/package=rms.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

Hastie, Trevor, Tibshirani, Robert, and Tibshirani, Ryan J. Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. jul 2017. URL http://arxiv.org/abs/1707.08692.

Heagerty, Patrick J. and Zheng, Yingye. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1):92–105, mar 2005. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2005.030814.x. URL http://doi.wiley.com/10.1111/j.0006-341X.2005.030814.x.

Heemskerk, A. Dorothee, Bang, Nguyen Duc, Mai, Nguyen Thi Hoang, Chau, Tran Thi Hong, Phu, Nguyen Hoan, Loc, Pham Phu, Chau, Nguyen Van Vinh, Hien, Tran Tinh, Dung, Nguyen Duc, Lan, Nguyen Thi Ngoc, Lan, Nguyen Huu, Lan, Nguyen Ngoc, Phong, Le Tan, Vien, Nguyen Nang, Hien, Nguyen Quang, Yen, Nguyen Thi Bich, Ha, Dang Thi Minh, Day, Jeremy, Maxine Caws', Merson, Laura,

Thinh, Tran Thi Van, Wolbers, Marcel, Thwaites, Guy E., and Farrar, Jeremy J. Intensified treatment with high dose rifampicin and levofloxacin compared to standard treatment for adult patients with tuberculous meningitis (TBM-IT): protocol for a randomized controlled trial. *Trials*, 12(1):25, 2011. ISSN 1745-6215. doi: 10.1186/1745-6215-12-25. URL http://www.trialsjournal.com/content/12/1/25.

Heemskerk, A Dorothee, Bang, Nguyen D, Mai, Nguyen T.H., Chau, Tran T.H., Phu, Nguyen H, Loc, Pham P, Chau, Nguyen V.V., Hien, Tran T., Dung, Nguyen H, Lan, Nguyen T.N., Lan, Nguyen H., Lan, Nguyen N, Phong, Le T, Vien, Nguyen N, Hien, Nguyen Q, Yen, Nguyen T.B., Ha, Dang T.M., Day, Jeremy N, Caws, Maxine, Merson, Laura, Thinh, Tran T.V., Wolbers, Marcel, Thwaites, Guy E, and Farrar, Jeremy J. Intensified Antituberculosis Therapy in Adults with Tuberculous Meningitis. *New England Journal of Medicine*, 374 (2):124–134, jan 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1507062. URL http://www.nejm.org/doi/10.1056/NEJMoa1507062.

Heemskerk, A. Dorothee, Nguyen, Mai Thi Hoang, Dang, Ha Thi Minh, Vinh Nguyen, Chau Van, Nguyen, Lan Huu, Do, Thu Dang Anh, Nguyen, Thuong Thuy Thuong, Wolbers, Marcel, Day, Jeremy, Le, Thao Thi Phuong, Nguyen, Bang Duc, Caws, Maxine, and Thwaites, Guy E. Clinical Outcomes of Patients with Drug-Resistant Tuberculous Meningitis Treated with an Intensified Antituberculosis Regimen. *Clinical Infectious Diseases*, 65(1):20–28, 2017. ISSN 15376591. doi: 10.1093/cid/cix230.

Hemingway, Harry, Croft, Peter, Perel, Pablo, Hayden, Jill A., Abrams, Keith, Timmis, Adam, Briggs, Andrew, Udumyan, Ruzan, Moons, Karel G. M., Steyerberg, Ewout W., Roberts, Ian, Schroter, Sara, Altman, Douglas G., Riley, Richard D., and PROGRESS Group. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ (Clinical research ed.)*, 346(feb05 1):e5595, feb 2013. ISSN 1756-1833. doi: 10.1136/bmj.e5595. URL http://www.ncbi.nlm.nih.gov/pubmed/23386360.

Henderson, Robin, Diggle, Peter, and Dobson, Angela. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000. ISSN 1468-4357.

Hernández, Adrián V, Steyerberg, Ewout W, and Habbema, J.Dik F. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases sta-

tistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*, 57(5):454–460, may 2004. ISSN 08954356. doi: 10.1016/j.jclinepi.2003.09.014. URL http://linkinghub.elsevier.com/retrieve/pii/S0895435603003792.

Heymans, Martijn W, van Buuren, Stef, Knol, Dirk L, van Mechelen, Willem, and de Vet, Henrica CW. Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, 7(1):33, dec 2007. ISSN 1471-2288. doi: 10.1186/1471-2288-7-33. URL http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-7-33.

Hickey, Graeme L., Philipson, Pete, Jorgensen, Andrea, and Kolamunnage-Dona, Ruwanthi. Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Medical Research Methodology*, 16(1):1–15, 2016. ISSN 14712288. doi: 10.1186/s12874-016-0212-5. URL http://dx.doi.org/10.1186/s12874-016-0212-5.

Hingorani, Aroon D., van der Windt, Daniëlle A., Riley, Richard D., Abrams, Keith, Moons, Karel G. M., Steyerberg, Ewout W., Schroter, Sara, Sauerbrei, Willi, Altman, Douglas G., Hemingway, Harry, and PROGRESS Group. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ (Clinical research ed.)*, 346(feb05 1):e5793, feb 2013. ISSN 1756-1833. doi: 10.1136/bmj.e5793. URL http://www.bmj.com/cgi/doi/10.1136/bmj.e5793http://www.ncbi.nlm.nih.gov/pubmed/23386361http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3565686.

Hosoglu, S, Geyik, M, and Balik, I. Predictors of outcome in patients with tuberculous meningitis. *International Journal of Tuberculosis and Lung Disease*, 6(1):64–70, 2002. ISSN 1027-3719.

Hosolu, S., Ayaz, C., Geyik, M. F., Kökolu, Ö F., and Ceviz, A. Tuberculous meningitis in adults: An eleven-year review. *International Journal of Tuberculosis and Lung Disease*, 2(7):553–557, 1998. ISSN 10273719.

Hsu, Po-Chang, Yang, Chien-Chang, Ye, Jung-Jr, Huang, Po-Yen, Chiang, Ping-Cherng, and Lee, Ming-Hsun. Prognostic factors of tuberculous meningitis in adults: a 6-year retrospective study at a tertiary hospital in northern Taiwan. *Journal of microbiology, immunology, and infection*, 43(2):111–118, 2010. ISSN 1995-9133. doi: 10.1016/S1684-1182(10)60018-7. URL http://dx.doi.org/10.1016/S1684-1182(10)60018-7.

Huang, Xuelin, Yan, Fangrong, Ning, Jing, Feng, Ziding, Choi, Sangbum, and Cortes, Jorge. A two-stage approach for dynamic prediction of time-to-event distributions. *Statistics in Medicine*, 35(13):2167–2182, jun 2016. ISSN 02776715. doi: 10.1002/sim.6860. URL http://doi.wiley.com/10.1002/sim.6860.

Hung, Hung and Chiang, ChinTsang. Estimation methods for timedependent AUC models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010. ISSN 0319-5724.

Jewell, Nicholas P. and Nielsen, Jens P. A Framework for Consistent Prediction Rules Based on Markers. *Biometrika*, 80(1):153, mar 1993. ISSN 00063444. doi: 10.2307/2336765. URL https://www.jstor.org/stable/2336765?origin=crossref.

Jha, Sneh Kumar, Garg, Ravindra Kumar, Jain, Amita, Malhotra, Hardeep Singh, Verma, Rajesh, and Sharma, Praveen Kumar. Definite (microbiologically confirmed) tuberculous meningitis: predictors and prognostic impact. *Infection*, 43 (6):639–45, dec 2015. ISSN 1439-0973. doi: 10.1007/s15010-015-0756-z. URL http://www.ncbi.nlm.nih.gov/pubmed/25724799.

Kalbfleisch, John D. and Prentice, Ross L. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, aug 2002. ISBN 9781118032985. doi: 10.1002/9781118032985. URL http://doi.wiley.com/10.1002/9781118032985.

Kalita, J and Misra, U K. Outcome of tuberculous meningitis at 6 and 12 months: a multiple regression analysis. *The International Journal of Tuberculosis and Lung Disease*, 3(3):261–265, 1999. ISSN 1027-3719.

Kamarudin, Adina Najwa, Cox, Trevor, and Kolamunnage-Dona, Ruwanthi. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, 17(1):53, dec 2017. ISSN 1471-2288. doi: 10.1186/s12874-017-0332-6. URL http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0332-6.

Kuhn, Max and Johnson, Kjell. *Applied Predictive Modeling*. Springer New York, 2013. ISBN 978-1-4614-6848-6. doi: 10.1007/978-1-4614-6849-3. URL http://link.springer.com/10.1007/978-1-4614-6849-3.

Lachenbruch, Peter A. Variable selection when missing values are present: a case study. *Statistical Methods in Medical Research*, 20(4):429–444, may 2010. ISSN 0962-2802. doi: 10.1177/0962280209358003. URL https://doi.org/10.1177/0962280209358003.

Lambert, Jérôme and Chevret, Sylvie. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical Methods in Medical Research*, 25(5):2088–2102, oct 2016. ISSN 0962-2802. doi: 10.1177/0962280213515571. URL http://journals.sagepub.com/doi/10.1177/0962280213515571.

Lau, Kwok Kwong, Yu, I. T S, Chan, A. C K, Wong, L. K S, Tam, C. M., Sheng, B., Li, H. L., Chan, H. S., Yam, K. L., Yeung, H. M., Chow, F. L., Ip, S., Tse, Philomena, Tam, C. M., Chan, C. K., Yew, W. W., Chan, K. S., Chen, W., Mok, Y. W., Ko, K. F., Kwan, M. C., Tsoi, T. H., Cheung, C. M., Tai, L. B., Wong, T. Y., Lau, K. K., Sheng, B., Leung, C. W., Li, H. L., Wong, K. S., Ip, Margaret, Yeung, W. L., Yu, I. T S, Ng, K. K., Wong, C. K., Tsang, Dominic, Wu, S. P., Mok, K. Y., Lam, C. W., Tsang, K. L., Chan, K. H., Lau, Y. L., Li, T. T., Tsang, K. L., Szeto, M. L., Mok, C. K., Tam, C. Y., Ng, P. W., Wong, T. P., and Kong, F. Y. A registry of tuberculous meningitis in Hong Kong. *International Journal of Tuberculosis and Lung Disease*, 9(12):1391–1397, 2005. ISSN 10273719.

Lesaffre, Emmanuel and Lawson, Andrew B. *Bayesian Biostatistics*. John Wiley & Sons, Ltd, Chichester, UK, jul 2012. ISBN 9781119942412. doi: 10.1002/9781119942412. URL http://doi.wiley.com/10.1002/9781119942412.

Lin, Haiqun, McCulloch, Charles E, and Mayne, Susan T. Maximum likelihood estimation in the joint analysis of timetoevent and multiple longitudinal variables. *Statistics in Medicine*, 21(16):2369–2382, 2002. ISSN 0277-6715.

Little, Roderick J A and Rubin, Donald B. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken, NJ, USA, aug 2002. ISBN 9781119013563. doi: 10.1002/9781119013563. URL http://doi.wiley.com/10.1002/9781119013563.

Long, Qi and Johnson, Brent A. Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics*, 16(3):596–610, 2014. ISSN 14684357. doi: 10.1093/biostatistics/kxv003.

Lu, C. H., Chang, W. N., and Chang, H. W. The prognostic factors of adult tuberculous meningitis. *Infection*, 29(6):299–304, 2001. ISSN 03008126. doi: 10.1007/s15010-001-1100-3.

Mai, Nguyen TH, Dobbs, Nicholas, Phu, Nguyen Hoan, Colas, Romain A, Thao, Le TP, Thuong, Nguyen TT, Nghia, Ho DT, Hanh, Nguyen HH, Hang, Nguyen T, Heemskerk, A Dorothee, Day, Jeremy N, Ly, Lucy, Thu, Do DA, Merson, Laura, Kestelyn, Evelyne, Wolbers, Marcel, Geskus, Ronald, Summers, David, Chau, Nguyen VV, Dalli, Jesmond, and Thwaites, Guy E. A randomised double blind placebo controlled phase 2 trial of adjunctive aspirin for tuberculous meningitis in HIV-uninfected adults. *eLife*, 7:1–20, feb 2018. ISSN 2050-084X. doi: 10.7554/eLife.33478. URL `https://elifesciences.org/articles/33478`.

Marais, Ben J, Heemskerk, Anna D, Marais, Suzaan S, van Crevel, Reinout, Rohlwink, Ursula, Caws, Maxine, Meintjes, Graeme, Misra, Usha K, Mai, Nguyen T H, and Ruslami, Rovina. Standardized methods for enhanced quality and comparability of tuberculous meningitis studies. *Clinical Infectious Diseases*, 64(4):501–509, 2016. ISSN 1058-4838.

Marais, Suzaan, Thwaites, Guy, Schoeman, Johan F, Török, M Estée, Misra, Usha K, Prasad, Kameshwar, Donald, Peter R, Wilkinson, Robert J, and Marais, Ben J. Tuberculous meningitis: a uniform case definition for use in clinical research. *The Lancet Infectious Diseases*, 10(11):803–812, nov 2010. ISSN 14733099. doi: 10.1016/S1473-3099(10)70138-9. URL `https://linkinghub.elsevier.com/retrieve/pii/S1473309910701389`.

Marais, Suzaan, Pepper, Dominique J, Schutz, Charlotte, Wilkinson, Robert J, and Meintjes, Graeme. Presentation and Outcome of Tuberculous Meningitis in a High HIV Prevalence Setting. *PLOS ONE*, 6(5):e20077, may 2011. doi: 10.1371/journal.pone.0020077. URL `https://doi.org/10.1371/journal.pone.0020077`.

Marais, Suzaan, Meintjes, Graeme, Pepper, Dominique J., Dodd, Lori E., Schutz, Charlotte, Ismail, Zahiera, Wilkinson, Katalin A., and Wilkinson, Robert J. Frequency, Severity, and Prediction of Tuberculous Meningitis Immune Reconstitution Inflammatory Syndrome. *Clinical Infectious Diseases*, 56(3):450–460, feb 2013. ISSN 1537-6591. doi: 10.1093/cid/cis899. URL `https://academic.oup.com/cid/article-lookup/doi/10.1093/cid/cis899`.

Marais, Suzaan, Wilkinson, Katalin A., Lesosky, Maia, Coussens, Anna K., Deffur, Armin, Pepper, Dominique J., Schutz, Charlotte, Ismail, Zahiera, Meintjes, Graeme, and Wilkinson, Robert J. Neutrophil-Associated Central Nervous System Inflammation in Tuberculous Meningitis Immune Reconstitution Inflammatory Syndrome. *Clinical Infectious Diseases*, 59(11):1638–1647, dec 2014. ISSN 1537-6591. doi: 10.1093/cid/ciu641. URL `https://academic.oup.com/cid/article-lookup/doi/10.1093/cid/ciu641`.

Mauff, Katya, Steyerberg, Ewout W, Nijpels, Giel, van der Heijden, Amber A W A, and Rizopoulos, Dimitris. Extension of the association structure in joint models to include weighted cumulative effects. *Statistics in medicine*, 36(23):3746–3759, 2017. ISSN 0277-6715.

Maziarz, Marlena, Heagerty, Patrick, Cai, Tianxi, and Zheng, Yingye. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics*, 73(1):83–93, 2017. ISSN 15410420. doi: 10.1111/biom.12562.

McPherson, Winston Chang, Cheng, Joe, Allaire, JJ, Xie, Yihui, and Jonathan. shiny: Web Application Framework for R. 2018. URL `https://cran.r-project.org/package=shiny`.

Meira-Machado, Luís, de Uña-Álvarez, Jacobo, Cadarso-Suárez, Carmen, and Andersen, Per K. Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, 18(2):195–222, 2009. ISSN 0962-2802.

Misra, U K, Kalita, J., Srivastava, M., and Mandal, S K. Prognosis of tuberculous meningitis: a multivariate analysis. *Journal of the neurological sciences*, 137(1): 57–61, apr 1996. ISSN 0022-510X. doi: 10.1016/0022-510X(95)00334-X. URL `http://www.ncbi.nlm.nih.gov/pubmed/9120488`.

Misra, U K, Kalita, J, Roy, A K, Mandal, S K, and Srivastava, M. Role of clinical, radiological, and neurophysiological changes in predicting the outcome of tuberculous meningitis: a multivariable analysis. *Journal of neurology, neurosurgery, and psychiatry*, 68(3):300–3, mar 2000. ISSN 0022-3050. doi: 10.1136/jnnp.68.3.300. URL `http://www.ncbi.nlm.nih.gov/pubmed/10675210`.

Misra, Usha K, Kalita, Jayantee, Bhoi, Sanjeev K, and Singh, Rajesh K. A study of hyponatremia in tuberculous meningitis. *Journal of the neurological sciences*, 367:152–7, aug 2016. ISSN 1878-5883. doi: 10.1016/j.jns.2016.06.004. URL `http://www.ncbi.nlm.nih.gov/pubmed/27423581`.

Mogensen, Ulla B, Ishwaran, Hemant, and Gerds, Thomas A. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, 50(11):1–23, 2012. URL http://www.jstatsoft.org/v50/i11/.

Moons, K. G M, Royston, P., Vergouwe, Y., Grobbee, D. E, and Altman, D. G. Prognosis and prognostic research: what, why, and how? *BMJ*, 338(feb23 1):b375–b375, feb 2009. ISSN 0959-8138. doi: 10.1136/bmj.b375. URL http://www.bmj.com/cgi/doi/10.1136/bmj.b375.

Musoro, Jammbe Z, Zwinderman, Aeilko H, Puhan, Milo a, ter Riet, Gerben, and Geskus, Ronald B. Validation of prediction models based on lasso regression with multiply imputed data. *BMC Medical Research Methodology*, 14(1):116, dec 2014. ISSN 1471-2288. doi: 10.1186/1471-2288-14-116. URL http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-116.

Nguyen, Duc Bui, Do, Nhan Thi, Shiraishi, Ray W, Le, Yen Ngoc, Tran, Quang Hong, Huu Nguyen, Hai, Medland, Nicholas, Nguyen, Long Thanh, and Struminger, Bruce Baird. Outcomes of Antiretroviral Therapy in Vietnam: Results from a National Evaluation. *PLOS ONE*, 8(2):e55750, feb 2013. URL https://doi.org/10.1371/journal.pone.0055750.

Nhu, Nguyen Thi Quynh, Heemskerk, Dorothee, Thu, Do Dang Anh, Chau, Tran Thi Hong, Mai, Nguyen Thi Hoang, Nghia, Ho Dang Trung, Loc, Pham Phu, Ha, Dang Thi Minh, Merson, Laura, Thinh, Tran Thi Van, Day, Jeremy, Chau, Nguyen van Vinh, Wolbers, Marcel, Farrar, Jeremy, and Caws, Maxine. Evaluation of GeneXpert MTB/RIF for diagnosis of tuberculous meningitis. *Journal of clinical microbiology*, 52(1):226–33, jan 2014. ISSN 1098-660X. doi: 10.1128/JCM.01834-13. URL http://www.ncbi.nlm.nih.gov/pubmed/24197880.

Ogundimu, Emmanuel O, Altman, Douglas G, and Collins, Gary S. Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of clinical epidemiology*, 76:175–82, 2016. ISSN 1878-5921. doi: 10.1016/j.jclinepi.2016.02.031. URL http://www.ncbi.nlm.nih.gov/pubmed/26964707.

Pajouheshnia, Romin, Damen, Johanna A. A. G., Groenwold, Rolf H. H., Moons, Karel G. M., and Peelen, Linda M. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. *Diagnostic and Prognostic Research*, 1(1):15, dec 2017. ISSN 2397-7523. doi:

10.1186/s41512-017-0015-0. URL http://diagnprognres.biomedcentral.com/articles/10.1186/s41512-017-0015-0.

Palczewska, Anna, Palczewski, Jan, Robinson, Richard Marchese, and Neagu, Daniel. Interpreting random forest classification models using a feature contribution method. In *Integration of reusable systems*, pages 193–218. Springer, 2014.

Papageorgiou, Grigorios, Mauff, Katya, Tomer, Anirudh, and Rizopoulos, Dimitris. An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes. *Annual Review of Statistics and Its Application*, aug 2018. ISSN 2326-8298. doi: 10.1146/annurev-statistics-030718-105048. URL https://doi.org/10.1146/annurev-statistics-030718-105048.

Patel, Vinod B, Connolly, Cathy, Singh, Ravesh, Lenders, Laura, Matinyenya, Brian, Theron, Grant, Ndung'u, Thumbi, and Dheda, Keertan. Comparison of amplicor and GeneXpert MTB/RIF tests for diagnosis of tuberculous meningitis. *Journal of clinical microbiology*, 52(10):3777–80, oct 2014. ISSN 1098-660X. doi: 10.1128/JCM.01235-14. URL http://www.ncbi.nlm.nih.gov/pubmed/25056328.

Peduzzi, P, Concato, J, Kemper, E, Holford, T R, and Feinstein, A R. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–9, dec 1996. ISSN 0895-4356. URL http://www.ncbi.nlm.nih.gov/pubmed/8970487.

Prasad, Kameshwar and Singh, M. B. Corticosteroids for managing tuberculous meningitis. *Cochrane Database of Systematic Reviews*, (1), 2008. ISSN 1469493X. doi: 10.1002/14651858.CD002244.pub3.

R. Brown, Elizabeth and G. Ibrahim, Joseph. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59(2):221–228, 2003. ISSN 0006-341X.

R Core Team. R: A Language and Environment for Statistical Computing. 2018. URL https://www.r-project.org/.

Rich, Arnold R. The pathogenesis of tuberculous meningitis. *Bull. Johns Hopkins Hosp*, 52:5, 1933.

Rizopoulos, Dimitris. Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics*, 67(3):819–829, sep 2011.

ISSN 0006341X. doi: 10.1111/j.1541-0420.2010.01546.x. URL http://doi.wiley.com/10.1111/j.1541-0420.2010.01546.x.

Rizopoulos, Dimitris. *Joint models for longitudinal and time-to-event data*. Chapman and Hall/CRC, 2012. ISBN 1439872872.

Rizopoulos, Dimitris. The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *Journal of Statistical Software*, 72(7):1–45, 2016. ISSN 1548-7660. doi: 10.18637/jss.v072.i07. URL http://www.jstatsoft.org/v72/i07/.

Rizopoulos, Dimitris and Ghosh, Pulak. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30(12):1366–1380, 2011. ISSN 02776715. doi: 10.1002/sim.4205.

Rizopoulos, Dimitris, Molenberghs, Geert, and Lesaffre, Emmanuel M.E.H. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276, 2017. ISSN 15214036. doi: 10.1002/bimj.201600238.

Rock, R B, Olin, M, Baker, C A, Molitor, T W, and Peterson, P K. Central Nervous System Tuberculosis: Pathogenesis and Clinical Aspects. *Clinical Microbiology Reviews*, 21(2):243–261, apr 2008. doi: 10.1128/cmr.00042-07. URL http://dx.doi.org/10.1128/CMR.00042-07.

Roozenbeek, Bob, Maas, Andrew I. R., Lingsma, Hester F., Butcher, Isabella, Lu, Juan, Marmarou, Anthony, McHugh, Gillian S., Weir, James, Murray, Gordon D., and Steyerberg, Ewout W. Baseline characteristics and statistical power in randomized controlled trials: Selection, prognostic targeting, or covariate adjustment? *Critical Care Medicine*, 37(10):2683–2690, oct 2009. ISSN 0090-3493. doi: 10.1097/CCM.0b013e3181ab85ec. URL https://insights.ovid.com/crossref?an=00003246-200910000-00001.

Royston, P., Moons, K. G M, Altman, D. G, and Vergouwe, Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ*, 338(mar31 1):b604–b604, mar 2009. ISSN 0959-8138. doi: 10.1136/bmj.b604. URL http://www.bmj.com/cgi/doi/10.1136/bmj.b604.

Royston, Patrick and Altman, Douglas G. External validation of a Cox prognostic model: principles and methods. *BMC medical research methodology*, 13:33, mar 2013. ISSN 1471-2288. doi: 10.1186/1471-2288-13-33. URL http://www.ncbi.nlm.nih.gov/pubmed/23496923.

Royston, Patrick and Sauerbrei, Willi. *Multivariable Model-Building*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, apr 2008. ISBN 9780470770771. doi: 10.1002/9780470770771. URL http://doi.wiley.com/10.1002/9780470770771.

Sakula, A. Robert koch: centenary of the discovery of the tubercle bacillus, 1882. *The Canadian veterinary journal*, 24(4):127–31, apr 1983. ISSN 0008-5286. URL http://www.ncbi.nlm.nih.gov/pubmed/17422248http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1790283.

Salekeen, Sirajus, Mahmood, Khalid, Naqvi, Iftikhar Haider, Baig, Mirza Yousuf, Akhter, Syed Tehssen, and Abbasi, Amanullah. Clinical course, complications and predictors of mortality in patients with tuberculous meningitis–an experience of fifty two cases at Civil Hospital Karachi, Pakistan. *JPMA. The Journal of the Pakistan Medical Association*, 63(5):563–7, may 2013. ISSN 0030-9982. URL http://www.ncbi.nlm.nih.gov/pubmed/23757980.

Seddon, James A, Visser, Douwe H, Bartens, Margaux, Jordaan, Annemie M, Victor, Thomas C, van Furth, A Marceline, Schoeman, Johan F, and Schaaf, H Simon. Impact of drug resistance on clinical outcome in children with tuberculous meningitis. *The Pediatric infectious disease journal*, 31(7):711–6, jul 2012. ISSN 1532-0987. doi: 10.1097/INF.0b013e318253acf8. URL http://www.ncbi.nlm.nih.gov/pubmed/22411053.

Sharma, S, Goyal, M K, Sharma, K, Modi, M, Sharma, M, Khandelwal, N, Prabhakar, S, Sharma, N, Shree, R, and Gairolla, J. Cytokines do play a role in pathogenesis of tuberculous meningitis: A prospective study from a tertiary care center in India. *Journal of the neurological sciences*, 379:131–136, 2017. ISSN 0022-510X.

Shaw, Joanna E T, Pasipanodya, Jotam G, and Gumbo, Tawanda. Meningeal tuberculosis: high long-term mortality despite standard therapy. *Medicine*, 89(3):189–95, may 2010. ISSN 1536-5964. doi: 10.1097/MD.0b013e3181df9070. URL http://www.ncbi.nlm.nih.gov/pubmed/20453606.

Shmueli, Galit. To Explain or to Predict? *Statistical Science*, 25(3):289–310, aug 2010. ISSN 0883-4237. doi: 10.1214/10-STS330. URL http://projecteuclid.org/euclid.ss/1294167961.

Simmons, Cameron P, Thwaites, Guy E, Quyen, Nguyen Than Ha, Chau, Tran Thi Hong, Mai, Pham Phuong, Dung, Nguyen Thi, Stepniewska, Kasia, White, Nicholas J, Hien, Tran Tinh, and Farrar, Jeremy. The clinical benefit of adjunctive dexamethasone in tuberculous meningitis is not associated with measurable attenuation of peripheral or local immune responses. *Journal of immunology (Baltimore, Md. : 1950)*, 175(1):579–590, 2005. ISSN 0022-1767. doi: 10.4049/jimmunol.175.1.579.

Steyerberg, E W, Eijkemans, M J, and Habbema, J D. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of clinical epidemiology*, 52(10):935–42, oct 1999. ISSN 0895-4356. URL http://www.ncbi.nlm.nih.gov/pubmed/10513756.

Steyerberg, E.W. *Clinical Prediction Models*. Statistics for Biology and Health. Springer New York, New York, NY, 2009. ISBN 978-0-387-77243-1. doi: 10.1007/978-0-387-77244-8.

Steyerberg, Ewout W. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2008. ISBN 0387772448.

Steyerberg, Ewout W, Eijkemans, Marinus J.C., Harrell, Frank E, and Habbema, J Dik F. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, 19(8):1059–1079, apr 2000. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(20000430)19:8⟨1059::AID-SIM412⟩3.0.CO;2-0.

Steyerberg, Ewout W., Borsboom, Gerard J. J. M., van Houwelingen, Hans C., Eijkemans, Marinus J. C., and Habbema, J. Dik F. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine*, 23(16):2567–2586, aug 2004. ISSN 0277-6715. doi: 10.1002/sim.1844. URL http://doi.wiley.com/10.1002/sim.1844.

Steyerberg, Ewout W, Vickers, Andrew J, Cook, Nancy R, Gerds, Thomas, Gonen, Mithat, Obuchowski, Nancy, Pencina, Michael J, and Kattan, Michael W. Assessing the performance of prediction models: a framework for traditional and novel

measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–38, jan 2010. ISSN 1531-5487. doi: 10.1097/EDE.0b013e3181c30fb2. URL http://www.ncbi.nlm.nih.gov/pubmed/20010215.

Steyerberg, Ewout W., Moons, Karel G. M., van der Windt, Danielle A., Hayden, Jill A., Perel, Pablo, Schroter, Sara, Riley, Richard D., Hemingway, Harry, and Altman, Douglas G. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Medicine*, 10(2):e1001381, feb 2013. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001381. URL http://dx.plos.org/10.1371/journal.pmed.1001381.

Streptomycin in Tuberculosis Trials Commitee, Medical Research Council. STREPTOMYCIN treatment of tuberculous meningitis. *Lancet (London, England)*, 1 (6503):582–96, apr 1948. ISSN 0140-6736. doi: 10.1016/S0140-6736(48)92003-0. URL http://www.ncbi.nlm.nih.gov/pubmed/18911226.

Suresh, Krithika, Taylor, Jeremy M.G., Spratt, Daniel E., Daignault, Stephanie, and Tsodikov, Alexander. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal*, 59(6):1277–1300, 2017. ISSN 15214036. doi: 10.1002/bimj.201600235.

Sütla, P N, Unal, A, Forta, H, Senol, S, and Kirba, D. Tuberculous meningitis in adults: review of 61 cases. *Infection*, 31(6):387–91, dec 2003. ISSN 0300-8126. doi: 10.1007/s15010-003-3179-1. URL http://www.ncbi.nlm.nih.gov/pubmed/14735380.

Sweeting, Michael J, Barrett, Jessica K, Thompson, Simon G, and Wood, Angela M. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. *Statistics in medicine*, 36 (28):4514–4528, dec 2017. ISSN 1097-0258. doi: 10.1002/sim.7144. URL http://www.ncbi.nlm.nih.gov/pubmed/27730661.

Teasdale, Graham and Jennett, Bryan. Assessment of coma and impaired consciousness. *The Lancet*, 304(7872):81–84, jul 1974. ISSN 01406736. doi: 10.1016/S0140-6736(74)91639-0. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673674916390.

Thao, Le Thi Phuong, Geskus, Ronald, and Thwaites, Guy. Reply to Dian et al. *Clinical Infectious Diseases*, 67(12):1955, nov 2018a. ISSN 1058-4838. URL http://dx.doi.org/10.1093/cid/ciy447.

Thao, Le Thi Phuong, Heemskerk, A Dorothee, Geskus, Ronald B, Mai, Nguyen Thi Hoang, Ha, Dang Thi Minh, Chau, Tran Thi Hong, Phu, Nguyen Hoan, Chau, Nguyen Van Vinh, Caws, Maxine, Lan, Nguyen Huu, Thu, Do Dang Anh, Thuong, Nguyen Thuy Thuong, Day, Jeremy, Farrar, Jeremy J, Torok, M Estee, Bang, Nguyen Duc, Thwaites, Guy E, and Wolbers, Marcel. Prognostic Models for 9-Month Mortality in Tuberculous Meningitis. *Clinical Infectious Diseases*, 66(4):523–532, feb 2018b. ISSN 1058-4838. doi: 10.1093/cid/cix849. URL `http://academic.oup.com/cid/article/doi/10.1093/cid/cix849/4237651/Prognostic-models-for-9-month-mortality-in`.

Therneau, Terry M. A Package for Survival Analysis in S, 2015. URL `http://cran.r-project.org/package=survival`.

Therneau, Terry M and Grambsch, Patricia M. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013. ISBN 1475732945.

Thiébaut, Rodolphe, JacqminGadda, Hélène, Babiker, Abdel, Commenges, Daniel, and Collaboration, Cascade. Joint modelling of bivariate longitudinal data with informative dropout and leftcensoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Statistics in medicine*, 24(1):65–82, 2005. ISSN 0277-6715.

Tho, Dau Quang, Estée Török, M., Yen, Nguyen Thi Bich, Bang, Nguyen Duc, Lan, Nguyen Thi Ngoc, Kiet, Vo Sy, Chau, Nguyen Van Vinh, Dung, Nguyen Huy, Day, Jeremy, Farrar, Jeremy, Wolbers, Marcel, Caws, Maxine, Török, M Estée, Yen, Nguyen Thi Bich, Bang, Nguyen Duc, Lan, Nguyen Thi Ngoc, Kiet, Vo Sy, van Vinh Chau, Nguyen, Dung, Nguyen Huy, Day, Jeremy, Farrar, Jeremy, Wolbers, Marcel, and Caws, Maxine. Influence of Antituberculosis Drug Resistance and Mycobacterium tuberculosis Lineage on Outcome in HIV-Associated Tuberculous Meningitis. *Antimicrobial Agents and Chemotherapy*, 56(6):3074–3079, jun 2012. ISSN 0066-4804. doi: 10.1128/AAC.00319-12. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3370761/`.

Thuong, Nguyen T T, Heemskerk, Dorothee, Tram, Trinh T B, Thao, Le T P, Ramakrishnan, Lalita, Ha, Vu T N, Bang, Nguyen D, Chau, Tran T H, Lan, Nguyen H, Caws, Maxine, Dunstan, Sarah J, Chau, Nguyen V V, Wolbers, Marcel, Mai, Nguyen T H, and Thwaites, Guy E. Leukotriene A4 Hydrolase Genotype and

HIV Infection Influence Intracerebral Inflammation and Survival From Tuberculous Meningitis. *The Journal of Infectious Diseases*, 215(7):1020–1028, apr 2017. ISSN 0022-1899. URL http://dx.doi.org/10.1093/infdis/jix050.

Thwaites, G, Chau, TTH, Mai, NTH, Drobniewski, F, McAdam, K, and Farrar, J. Tuberculous meningitis. *J Neurol Neurosurg Psychiatry*, 68:289–299, 2000. ISSN 13507540. doi: 10.1016/j.rmed.2006.08.006.

Thwaites, Guy. Tuberculous meningitis. *Medicine (United Kingdom)*, 45(11):670–673, 2017. ISSN 13654357. doi: 10.1016/j.mpmed.2017.08.010. URL https://doi.org/10.1016/j.mpmed.2017.08.010.

Thwaites, Guy E and Tran, Tinh Hien. Tuberculous meningitis: many questions, too few answers. *The Lancet. Neurology*, 4(3):160–70, mar 2005. ISSN 1474-4422. doi: 10.1016/S1474-4422(05)01013-6. URL http://www.ncbi.nlm.nih.gov/pubmed/15721826.

Thwaites, Guy E, Simmons, Cameron P, Than Ha Quyen, Nguyen, Thi Hong Chau, Tran, Phuong Mai, Pham, Thi Dung, Nguyen, Hoan Phu, Nguyen, White, Nicholas P, Tinh Hien, Tran, and Farrar, Jeremy J. Pathophysiology and prognosis in vietnamese adults with tuberculous meningitis. *The Journal of infectious diseases*, 188(8):1105–15, oct 2003. ISSN 0022-1899. doi: 10.1086/378642. URL http://www.ncbi.nlm.nih.gov/pubmed/14551879.

Thwaites, Guy E, Bang, Nguyen Duc, Dung, Nguyen Huy Thi, Quy, Hoang Thi, Oanh, Do Thi Tuong, Thoa, Nguyen Thi Cam, Hien, Nguyen Quang, Thuc, Nguyen Tri, Hai, Nguyen Ngoc, Lan, Nguyen Ngoc Thi Ngoc, Lan, Nguyen Ngoc Thi Ngoc, Duc, Nguyen Hong, Tuan, Vu Ngoc, Hiep, Cao Huu, Chau, Tran Thi Hong, Mai, Pham Phuong, Dung, Nguyen Huy Thi, Stepniewska, Kasia, White, Nicholas J, Hien, Tran Tinh, Farrar, Jeremy J, Nguyen, Duc Bang, Nguyen, Hong Duc Huy Dung, Hoang, Thi Quy, Do, Thi Tuong Oanh, Nguyen, Thi Cam Thoa Dung Ngoc Lan Tri Thuc, Nguyen, Quang Hien, Nguyen, Thi Cam Thoa Dung Ngoc Lan Tri Thuc, Nguyen, Ngoc Hai Lan, Nguyen, Thi Cam Thoa Dung Ngoc Lan Tri Thuc, Nguyen, Ngoc Hai Lan, Nguyen, Hong Duc Huy Dung, Vu, Ngoc Tuan, Cao, Huu Hiep, Tran, Tinh Hien Thi Hong Chau, Pham, Phuong Mai, Nguyen, Thi Cam Thoa Dung Ngoc Lan Tri Thuc, Stepniewska, Kasia, White, Nicholas J, Tran, Tinh Hien Thi Hong Chau, and Farrar, Jeremy J. Dexamethasone for the Treatment of Tuberculous Meningitis in Adolescents and Adults. *The*

New England journal of medicine, 351(17):1741–1751, 2004a. ISSN 1533-4406. doi: 10.1056/NEJMoa040573. URL http://dx.doi.org/10.1056/NEJMoa040573.

Thwaites, Guy E, Chau, Tran Thi Hong, and Farrar, Jeremy J. Improving the bacteriological diagnosis of tuberculous meningitis. *Journal of clinical microbiology*, 42(1):378–9, jan 2004b. ISSN 0095-1137. URL http://www.ncbi.nlm.nih.gov/pubmed/14715783.

Thwaites, Guy E, Thi, Tran, Chau, Hong, and Farrar, Jeremy J. Improving the Bacteriological Diagnosis of Tuberculous Meningitis Improving the Bacteriological Diagnosis of Tuberculous Meningitis. 42(1):378–380, 2004c. doi: 10.1128/JCM.42.1.378.

Thwaites, Guy E., Lan, Nguyen Thi Ngoc, Dung, Nguyen Huy, Quy, Hoang Thi, Oanh, Do Thi Tuong, Thoa, Nguyen Thi Cam, Hien, Nguyen Quang, Thuc, Nguyen Tri, Hai, Nguyen Ngoc, Bang, Nguyen Duc, Lan, Nguyen Ngoc, Duc, Nguyen Hong, Tuan, Vu Ngoc, Hiep, Cao Huu, Chau, Tran Thi Hong, Mai, Pham Phuong, Dung, Nguyen Thi, Stepniewska, Kasia, White, Nicholas J., Hien, Tran Tinh, and Farrar, Jeremy J. Effect of antituberculosis drug resistance on response to treatment and outcome in adults with tuberculous meningitis. *The Journal of infectious diseases*, 192(1):79–88, jul 2005. ISSN 0022-1899. doi: 10.1086/430616. URL http://www.ncbi.nlm.nih.gov/pubmed/15942897.

Thwaites, Guy E., Macmullen-Price, Jeremy, Chau, Tran Thi Hong, Phuong Mai, Pham, Dung, Nguyen Thi, Simmons, Cameron P., White, Nicholas J., Hien, Tran Tinh, Summers, David, and Farrar, Jeremy J. Serial MRI to determine the effect of dexamethasone on the cerebral pathology of tuberculous meningitis: an observational study. *Lancet Neurology*, 6(3):230–236, 2007. ISSN 14744422. doi: 10.1016/S1474-4422(07)70034-0.

Thwaites, Guy E., Bhavnani, Sujata M., Chau, Tran Thi Hong, Hammel, Jeffrey P., Török, M Estée, Van Wart, Scott a., Mai, Pham Phuong, Reynolds, Daniel K., Caws, Maxine, Dung, Nguyen Thi, Hien, Tran Tinh, Kulawy, Robert, Farrar, Jeremy, and Ambrose, Paul G. Randomized pharmacokinetic and pharmacodynamic comparison of fluoroquinolones for tuberculous meningitis. *Antimicrobial agents and chemotherapy*, 55(7):3244–53, jul 2011. ISSN 1098-6596. doi: 10.1128/AAC.00064-11. URL http://www.ncbi.nlm.nih.gov/pubmed/21502621.

Thwaites, Guy E, van Toorn, Ronald, and Schoeman, Johan. Tuberculous meningitis: more questions, still too few answers. *The Lancet. Neurology*, 12(10):999–1010, oct 2013. ISSN 1474-4465. doi: 10.1016/S1474-4422(13)70168-6. URL `http://www.ncbi.nlm.nih.gov/pubmed/23972913`.

Tibshirani, Robert. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346178`.

Torok, M. Estee, Chau, Tran Thin Hong, Mai, Pham Phuong, Phong, Nguyen Duy, Dung, Nguyen Thi, Van Chuong, Ly, Lee, Sue J., Caws, M., de Jong, Menno, Hien, Tran Tinh, and Farrar, Jeremy J. Clinical and microbiological features of HIV-associated tuberculous meningitis in Vietnamese adults. *PLoS ONE*, 3(3):1–7, 2008. ISSN 19326203. doi: 10.1371/journal.pone.0001772.

Török, M. Estee, Yen, Nguyen Thi Bich, Chau, Tran Thi Hong, Mai, Nguyen Thi Hoang, Phu, Nguyen Hoan, Mai, Pham Phuong, Dung, Nguyen Thi, Chau, Nguyen Van Vinh, Bang, Nguyen Duc, Tien, Nguyen Anh, Minh, N. H., Hien, Nguyen Quang, Thai, Phan Vuong Khac, Dong, Doan The, Anh, Do Thi Tuong, Thoa, Nguyen Thi Cam, Hai, Nguyen Ngoc, Lan, Nguyen Ngoc, Lan, Nguyen Thi Ngoc, Quy, Hoang Thi, Dung, Nguyen Huy, Hien, Tran Tinh, Chinh, Nguyen Tran, Simmons, Cameron Paul, de Jong, Menno, Wolbers, Marcel, and Farrar, Jeremy James. Timing of initiation of antiretroviral therapy in human immunodeficiency virus (HIV)–associated tuberculous meningitis. *Clinical infectious diseases*, 52(11):1374–83, jun 2011a. ISSN 1537-6591. doi: 10.1093/cid/cir230. URL `http://www.ncbi.nlm.nih.gov/pubmed/21596680`.

Török, M. Estee, Yen, Nguyen Thi Bich, Chau, Tran Thi Hong, Mai, Nguyen Thi Hoang, Phu, Nguyen Hoan, Mai, Pham Phuong, Dung, Nguyen Thi, Chau, Nguyen Van Vinh, Bang, Nguyen Duc, Tien, Nguyen Anh, Minh, N. H., Hien, Nguyen Quang, Thai, Phan Vuong Khac, Dong, Doan The, Anh, Do Thi Tuong, Thoa, Nguyen Thi Cam, Hai, Nguyen Ngoc, Lan, Nguyen Ngoc, Lan, Nguyen Thi Ngoc, Quy, Hoang Thi, Dung, Nguyen Huy, Hien, Tran Tinh, Chinh, Nguyen Tran, Simmons, Cameron Paul, De Jong, Menno, Wolbers, Marcel, and Farrar, Jeremy James. Timing of initiation of antiretroviral therapy in human immunodeficiency virus (HIV)-associated tuberculous meningitis. *Clinical Infectious Diseases*, 52(11):1374–1383, 2011b. ISSN 10584838. doi: 10.1093/cid/cir230.

Tsiatis, A A, Degruttola, Victor, and Wulfsohn, M S. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90 (429):27–37, 1995. ISSN 0162-1459.

Tsiatis, Anastasios A and Davidian, Marie. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004. ISSN 1017-0405.

Uno, Hajime, Cai, Tianxi, Tian, Lu, and Wei, L. J. Evaluating Prediction Rules for t -Year Survivors With Censored Regression Models. *Journal of the American Statistical Association*, 102(478):527–537, jun 2007. ISSN 0162-1459. doi: 10.1198/016214507000000149. URL `http://www.tandfonline.com/doi/abs/10.1198/016214507000000149`.

Van, Thinh Tran Thi and Farrar, Jeremy. Tuberculous meningitis. *Journal of Epidemiology and Community Health*, 68(3):195–196, mar 2014. ISSN 0143-005X. doi: 10.1136/jech-2013-202525. URL `http://jech.bmj.com/lookup/doi/10.1136/jech-2013-202525`.

van Buuren, Stef. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, jun 2007. ISSN 0962-2802. doi: 10.1177/0962280206074463. URL `http://journals.sagepub.com/doi/10.1177/0962280206074463`.

van Buuren, Stef. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton, FL, 2012. ISBN 9781439868249.

van Buuren, Stef and Groothuis-Oudshoorn, Karin. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL `http://www.jstatsoft.org/v45/i03/`.

van Houwelingen, Hans and Putter, Hein. *Dynamic Prediction in Clinical Survival Analysis*, volume 9. CRC press, 2011. ISBN 9781439835333.

Van Houwelingen, Hans C. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007. ISSN 03036898. doi: 10.1111/j.1467-9469.2006.00529.x.

van Houwelingen, Hans C. and Sauerbrei, Willi. Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited. *Open Journal of Statistics*, 03

(02):79–102, 2013. ISSN 2161-718X. doi: 10.4236/ojs.2013.32011. URL http://www.scirp.org/journal/doi.aspx?DOI=10.4236/ojs.2013.32011.

van Laarhoven, Arjan, Dian, Sofiati, Ruesen, Carolien, Hayati, Ela, Damen, Michelle S M A, Annisa, Jessi, Chaidir, Lidya, Ruslami, Rovina, Achmad, Tri Hanggono, Netea, Mihai G, Alisjahbana, Bachti, Rizal Ganiem, Ahmad, and van Crevel, Reinout. Clinical Parameters, Routine Inflammatory Markers, and LTA4H Genotype as Predictors of Mortality Among 608 Patients With Tuberculous Meningitis in Indonesia. *The Journal of Infectious Diseases*, 215(7):1029–1039, apr 2017. ISSN 0022-1899. URL http://dx.doi.org/10.1093/infdis/jix051.

van Smeden, Maarten, de Groot, Joris A. H., Moons, Karel G. M., Collins, Gary S., Altman, Douglas G., Eijkemans, Marinus J. C., and Reitsma, Johannes B. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1):163, dec 2016. ISSN 1471-2288. doi: 10.1186/s12874-016-0267-3. URL http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0267-3.

van Smeden, Maarten, Moons, Karel GM, de Groot, Joris AH, Collins, Gary S, Altman, Douglas G, Eijkemans, Marinus JC, and Reitsma, Johannes B. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, page 096228021878472, jul 2018. ISSN 0962-2802. doi: 10.1177/0962280218784726. URL http://journals.sagepub.com/doi/10.1177/0962280218784726.

van Well, Gijs T J, Paes, Berbe F, Terwee, Caroline B, Springer, Priscilla, Roord, John J, Donald, Peter R, van Furth, A Marceline, and Schoeman, Johan F. Twenty years of pediatric tuberculous meningitis: a retrospective cohort study in the western cape of South Africa. *Pediatrics*, 123(1):e1–8, 2009. ISSN 1098-4275. URL http://www.ncbi.nlm.nih.gov/pubmed/19367678.

Vergouwe, Yvonne, Royston, Patrick, Moons, Karel G.M., and Altman, Douglas G. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology*, 63(2):205–214, 2010. ISSN 08954356. doi: 10.1016/j.jclinepi.2009.03.017. URL http://linkinghub.elsevier.com/retrieve/pii/S0895435609001188.

Vinnard, Christopher, Winston, Carla A, Wileyto, E Paul, Macgregor, Rob Roy, and Bisson, Gregory P. Isoniazid resistance and death in patients with tuberculous

meningitis: retrospective cohort study. *BMJ*, 341:c4451, sep 2010. ISSN 1756-1833. doi: 10.1136/bmj.c4451. URL http://www.ncbi.nlm.nih.gov/pubmed/20819874.

Vittinghoff, E. and McCulloch, C. E. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, 165(6):710–718, jan 2007. ISSN 0002-9262. doi: 10.1093/aje/kwk052. URL https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwk052.

Wahl, Simone, Boulesteix, Anne-Laure, Zierer, Astrid, Thorand, Barbara, and van de Wiel, Mark A. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. *BMC medical research methodology*, 16(1):144, 2016. ISSN 1471-2288. doi: 10.1186/s12874-016-0239-7. URL http://www.ncbi.nlm.nih.gov/pubmed/27782817.

Wan, Y., Datta, S., Conklin, D.J., and Kong, M. Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation*, 85(9): 1902–1916, jun 2015. ISSN 0094-9655. doi: 10.1080/00949655.2014.907801. URL http://www.tandfonline.com/doi/abs/10.1080/00949655.2014.907801.

Wang, Jann-Tay, Hung, Chien-Ching, Sheng, Wang-Huei, Wang, Jann-Yuan, Chang, Shan-Chwen, and Luh, Kwen-Tay. Prognosis of tuberculous meningitis in adults in the era of modern antituberculous chemotherapy. *Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi*, 35(4):215–22, dec 2002. ISSN 1684-1182. URL http://www.ncbi.nlm.nih.gov/pubmed/12542246.

Wang, Yan and Taylor, Jeremy M G. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96(455):895–905, 2001. ISSN 0162-1459.

White, Ian R and Royston, Patrick. Imputing missing covariate values for the Cox model. *Statistics in medicine*, 28(15):1982–98, jul 2009. ISSN 1097-0258. doi: 10.1002/sim.3618. URL http://www.ncbi.nlm.nih.gov/pubmed/19452569.

White, Ian R, Royston, Patrick, and Wood, Angela M. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4): 377–99, feb 2011. ISSN 1097-0258. doi: 10.1002/sim.4067. URL http://www.ncbi.nlm.nih.gov/pubmed/21225900.

Wickham, Hadley. tidyverse: Easily Install and Load the 'Tidyverse', 2017. URL https://cran.r-project.org/package=tidyverse.

Wilkinson, Robert J., Rohlwink, Ursula, Misra, Usha Kant, van Crevel, Reinout, Mai, Nguyen Thi Hoang, Dooley, Kelly E., Caws, Maxine, Figaji, Anthony, Savic, Rada, Solomons, Regan, Thwaites, Guy E., and Tuberculous Meningitis International Research Consortium. Tuberculous meningitis. *Nature reviews. Neurology*, 13(10): 581–598, oct 2017. ISSN 1759-4766. doi: 10.1038/nrneurol.2017.120. URL http://www.ncbi.nlm.nih.gov/pubmed/28884751.

Wood, Angela M, White, Ian R, and Royston, Patrick. How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27(17):3227–3246, jul 2008. ISSN 02776715. doi: 10.1002/sim.3177. URL http://doi.wiley.com/10.1002/sim.3177.

Wood, S N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

World Health Organization. Global Tuberculosis Report. WHO, 2016.

World Health Organization. Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2016. Technical report, 2018.

Wulfsohn, Michael S and Tsiatis, Anastasios A. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997. ISSN 0006-341X.

Xu, Jane and Zeger, Scott L. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):375–387, 2001. ISSN 0035-9254.

Yasar, K K, Pehlivanoglu, F, and Sengoz, G. Predictors of mortality in tuberculous meningitis: a multivariate analysis of 160 cases. *The International Journal of Tuberculosis and Lung Disease*, 14(10):1330–1335, 2010.

Zhao, Yize and Long, Qi. Variable selection in the presence of missing data: imputation-based methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5):e1402, sep 2017. ISSN 19395108. doi: 10.1002/wics.1402. URL http://doi.wiley.com/10.1002/wics.1402.